خوشه‌بندی خودکار داده‌ها با بهره‌گیری از الگوریتم رقابت استعماری بهبودیافته

آرش چاقری* **و محمدرضا فیضی درخشی**

دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

چکیده

الگوریتم رقابت استعماری (ICA)، یکی از کاراترین الگوریتم‌های فراابتکاری برای پیداکردن جواب بهینه سراسری در مسائل بهینه‌سازی است. در این مقاله از الگوریتم رقابت استعماری برای خوشه‌بندی خودکار مجموعه داده‌های بزرگ و واقعی بدون برچسب استفاده شده است. با بهره‌گیری از ساختار مناسب برای هر یک از کروموزم‌ها و استفاده از الگوریتم رقابت استعماری، در زمان اجرا تعداد بهینه خوشه‌ها هم‌زمان با خوشه‌بندی بهینه داده‌ها به‌دست می‌آید. همچنین برای افزایش دقت و افزایش سرعت هم‌گرایی، ساختار الگوریتم رقابت استعماری با تغییراتی همراه است. روش پیشنهادی (ACICA) نیاز به هیچ‌گونه دانش قبلی برای خوشه‌بندی داده‌ها ندارد؛ علاوه‌بر آن روش پیشنهادی در مقایسه با سایر روش‌های خوشه‌بندی مبتنی بر الگوریتم‌های تکاملی، دقت بیشتری را دارد. از معیارهای ارزیابی خوشه‌بندی DB و CS به‌عنوان تابع هدف استفاده شده است. برای نشان‌دادن برتری روش پیشنهادی، میانگین مقدار بهینه تابع هدف و تعداد خوشه‌های تعیین‌شده توسط روش پیشنهادی با سه الگوریتم خوشه‌بندی خودکار مبتنی بر الگوریتم‌های تکاملی مقایسه می‌شود.

واژگان کلیدی: خوشه‌بندی تفکیکی، خوشه‌بندی خودکار، الگوریتم رقابت استعماری

Automatic Clustering Using Improved Imperialist Competitive Algorithm

Arash Chaghari* & Mohammad-Reza Feizi-Derakhshi

Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

Abstract

Imperialist Competitive Algorithm (ICA) is considered as a prime meta-heuristic algorithm to find the general optimal solution in optimization problems. This paper presents a use of ICA for automatic clustering of huge unlabeled data sets. By using proper structure for each of the chromosomes and the ICA, at run time, the suggested method (ACICA) finds the optimum number of clusters while optimal clustering of the data simultaneously.To increase the accuracy and speed of convergence, the structure of ICA changes. As in different applications, there is a need for data clustering which the number of clusters is not known before it is necessary to have methods that can cluster data without knowing the correct prediction of the number of clusters. In the other words, the proposed algorithm requires no background knowledge to classify the data. In addition, the proposed method is more accurate in comparison with other clustering methods based on evolutionary algorithms. In Imperialist Competitive Algorithm, firstly steps should be taken to increase search rates and explore possible solution while approaching to the global optimal response the steps should be reduced to ensure that the algorithm is not lost and it is not in the local optimal manner. For this purpose and improvement of imperialist competitive algorithm, mutation rate and revolution operator's operation

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

rate are determined dynamically. DB and CS are cluster validity Indexes. In this paper, DB and CS cluster validity measurements are used as the objective function. To demonstrate the superiority of the proposed method, the average of fitness function and the number of clusters determined by the proposed method is compared with three automatic clustering algorithms based on evolutionary algorithms. The partitional clustering algorithms are based on three powerful well-known optimization algorithms, namely the genetic algorithm, the particle swarm optimization and differential evolutionary algorithm.

Keywords: Partitional Clustering, Automatic Clustering, Imperialist Competitive Algorithm (ICA)

۱- مقدمه

خوشه‌بندی به فرآیند تبدیل حجم عظیمی از داده‌ها به گروه‌های داده‌ای مشابه گفته می‌شود. هر گروه یک خوشه نامیده می‌شود. در هر خوشه داده‌هایی که بیشترین شباهت را به هم و کمترین میزان شباهت را با سایر خوشه‌ها دارند، قرار می‌گیرند. در دهه‌های اخیر، تحلیل خوشه نقش مهمی را در حوزه‌های مختلف مهندسی (به‌عنوان مثال یادگیری ماشین، هوش مصنوعی، شناسایی آماری الگو، مهندسی مکانیکی و مهندسی الکتریکی)، علوم کامپیوتر (به‌عنوان مثال داده‌کاوی وب، تحلیل پایگاه داده سه‌بعدی، مجموعه اسناد متنی و بخش‌بندی متن)، علوم پزشکی (به‌عنوان مثال ژنتیک، زیست‌شناسی، میکرب‌شناسی و آسیب‌شناسی)، علوم زمین (به‌عنوان مثال جغرافیا، زمین‌شناسی)، علوم اجتماعی (به‌عنوان مثال جامعه‌شناسی، روان‌شناسی، باستان‌شناسی و آموزش) و اقتصاد (به‌عنوان مثال بازاریابی و تجارت) دارد [1]- [3].

الگوریتم‌های خوشه‌بندی داده به دو دستهٔ سلسله‌مراتبی و تفکیکی [4]،[5] تقسیم می‌شوند. در هر دسته الگوریتم‌های زیادی برای پیداکردن خوشه‌ها وجود دارند. در خوشه‌بندی سلسله‌مراتبی، خروجی در قالب یک درخت ارائه می‌شود که هر سطح مراحل خوشه‌بندی را نشان می‌دهد. الگوریتم‌های سلسله‌مراتبی تجمعی (پائین- بالا) و یا تقسیم‌کننده می‌توانند باشند. در الگوریتم‌های تجمعی، در ابتدا هر داده به‌عنوان یک خوشه در نظر گرفته می‌شود. در هر مرحله خوشه‌های مشابه با هم ادغام می‌شوند و یک خوشه بزرگ‌تر را تشکیل می‌دهند. در الگوریتم‌های تقسیم‌کننده، مجموعه داده به‌عنوان یک خوشه در نظر گرفته و در مراحل بعدی خوشه‌بندی، به خوشه‌های کوچک‌تر تقسیم می‌شود. الگوریتم‌های خوشه‌بندی سلسله‌مراتبی دو مزیت عمده دارند. نخستین مزیت آن‌ها عدم نیاز به دانش قبلی برای تعداد خوشه‌ها است. دومین مزیت، مستقل‌بودن این دسته از الگوریتم‌ها به شرایط اولیه است. با این وجود مهم‌ترین ایراد این نوع از الگوریتم‌ها، ایستابودن خوشه‌ها در طی مراحل خوشه‌بندی است. به عبارت دیگر، اعضای خوشه‌ها در طی مراحل خوشه‌بندی نمی‌توانند به خوشه‌های دیگر منتقل شوند. به‌علاوه به‌دلیل عدم وجود اطلاعات درباره شکل کلی داده‌ها و یا تعداد خوشه‌ها، این الگوریتم‌ها ممکن است نتوانند خوشه‌های دارای هم‌پوشانی را از هم تفکیک کنند [6]. از سوی دیگر الگوریتم‌های تفکیکی سعی می‌کنند با استفاده از بهینه‌سازی یک معیار مشخص (به‌عنوان مثال، معیار مربع خطا) مجموعه داده را به‌صورت مستقیم به خوشه‌های مجزا از هم تقسیم کنند. در این نوع معیارهای ارزیابی خوشه‌بندی با استفاده از اندازه‌گیری شباهت بین داده‌ها، نمونه‌ها طوری خوشه‌بندی می‌شوند که نمونه‌های موجود در هر خوشه بیشترین شباهت را نسبت به هم و کمترین شباهت را نسبت به سایر خوشه‌ها داشته باشند. مزیت الگوریتم‌های سلسله‌مراتبی ایراد الگوریتم‌های تفکیکی محسوب می‌شود و برعکس.

خوشه‌بندی به دو صورت می‌تواند انجام شود: ۱.خوشه‌بندی غیر فازی[1] ۲.خوشه‌بندی فازی. درجه عضویت هر یک از نمونه‌ها در روش‌های غیر فازی صفر و یا یک است. درحقیقت می‌توان روش‌های غیر فازی را حالت خاصی از الگوریتم‌های فازی در نظر گرفت؛ به عبارت دیگر، داده‌ای که به یک خوشه تعلق دارد، درجه عضویت آن به خوشه مربوطه یک و درجه عضویت آن برای مابقی خوشه‌ها، صفر است. مزیت روش‌های غیر فازی، پیاده‌سازی آسان و کارابودن آن‌هاست. یکی از روش‌های معروف در این حوزه، الگوریتم K-means است. در روش‌های فازی، درجه عضویت یک نمونه به یک خوشه عددی بین صفر و یک است. به عبارت بهتر، هر چقدر یک نمونه داده به یک خوشه شباهت بیش‌تری داشته باشد، درجه عضویت آن به عدد یک نزدیک‌تر است و برعکس. در این مقاله فقط الگوریتم خوشه‌بندی غیر فازی بررسی می‌شوند.

در این مقاله با استفاده از محاسبات تکاملی سعی در ارائه الگوریتمی کارا برای خوشه‌بندی می‌شود. در روش‌های تکاملی، خوشه‌بندی مجموعه داده از منظر یک مسئله

[1] Crisp

بهینه‌سازی دیده می‌شود و با استفاده از الگوریتم‌های فراابتکاری، مانند الگوریتم ژنتیک [7]، این مسئله بهینه‌سازی حل می‌شود. ایدهٔ اصلی در الگوریتم‌های فرا ابتکاری، ایجاد جمعیتی از جواب‌های منتخب برای مسئله بهینه‌سازی است؛ به‌طوری‌که در هر تکرار الگوریتم، مجموعه جواب‌ها با استفاده از عملگرهای تغییر، بهبود می‌یابد. انتخاب جواب‌ها در جمعیت بر اساس معیار ارزیابی مربوط به مسئله بهینه‌سازی انجام می‌شود. در الگوریتم ژنتیک الگوریتم‌های تقاطع و جهش برای تغییر جواب‌ها استفاده می‌شوند. از عملگر جهش به‌منظور بررسی نقاط همسایگی یک جواب و از عملگر تقاطع به‌منظور ترکیب دو جواب و به دست آوردن یک جواب جدید استفاده می‌شود. این دسته از الگوریتم‌های بهینه‌سازی، با استفاده از حفظ، ترکیب و مقایسه بین چندین جواب به‌طور هم‌زمان از نقاط بهینه محلی می‌توانند فرار کنند که یکی از مزیت‌های مهم این نوع از الگوریتم‌ها محسوب می‌شود. الگوریتم‌های فراابتکاری محلی مانند شبیه‌سازی تبرید [8]، در یک زمان فقط یک جواب منتخب را بهینه می‌کنند و نمی‌توانند از نقاط بهینه محلی فرار کنند.

تلاش‌های بسیاری برای خوشه‌بندی داده‌های پیچیده و حجیم با استفاده از الگوریتم‌های تکاملی در سال‌های اخیر انجام شده است؛ ولی بیشتر این الگوریتم‌ها، نیاز به دانش قبلی برای تعداد خوشه‌ها دارند و نمی‌توانند تعداد بهینه خوشه‌ها را در حین اجرای برنامه به‌دست آورند. به‌علاوه تعداد خوشه‌ها برای تعدادی از مجموعه داده‌ها به‌طورکامل نامشخص است و حتی تقریبی از تعداد خوشه‌ها را نیز نمی‌توان مشخص کرد. به‌عنوان مثال برای خوشه‌بندی مجموعه‌ای از اسناد با توجه به سؤال کاربر برای موتور جستجو، تعداد خوشه‌ها با تغییر سؤال کاربر تغییر می‌کند و بنابراین نمی‌توان یک تعداد مشخص برای خوشه‌ها در نظر گرفت. همچنین برای مجموعه داده‌های با ابعاد زیاد، امکان رؤیت تصویری مجموعه داده برای به‌دست‌آوردن تعداد خوشه‌های آن غیرممکن است.

در مسائل مربوط به خوشه‌بندی داده‌ها، با توجه به اینکه دانشی در مورد مسأله در دسترس نیست، تولید جمعیت اولیه در بیشترمسائل به‌طورکامل تصادفی است و بنابراین رفتار الگوریتم تکاملی قابل پیش‌بینی نیست. به‌علاوه در بسیاری از روش های خوشه بندی معروف مانند الگوریتم K-means به تعدادی ورودی کاربر از جمله تعداد خوشه ها نیاز دارند. تعیین تعداد خوشه‌ها ازقبل برای کاربر بسیار مشکل می‌تواند باشد. به‌علاوه روش‌هایی مانند K-means این گرایش را دارند که در بهینه محلی گرفتار شوند. درنتیجه روش‌های خوشه‌بندی مبتنی بر الگوریتم های تکاملی ارائه شده‌اند. به‌طورمعمول جمعیت اولیه در الگوریتم‌های تکاملی به‌صورت تصادفی مقداردهی می‌شود. درحالی‌که با کنترل‌کردن تولید جمعیت اولیه نتایج به‌دست‌آمده برای خوشه‌بندی را می‌توان بهبود داد. بنابراین بعضی از روش‌های موجود مانند الگوریتم GenClust [9] اعضای اولیه جمعیت را با دقت بالا از لحاظ کیفیت انتخاب می‌کنند. پیچیدگی این کار $O(n^2)$ است که مقدار زیادی می‌باشد. در مقاله [10] یک الگوریتم خوشه‌بندی ارائه شده است که علاوه‌بر انتخاب بهینه اعضای اولیه جمعیت با پیچیدگی $O(n)$ از چندین اجزای جدید برای بهبود دقت خوشه‌بندی استفاده می‌کند. همچنین در مقاله [11] راه‌کاری برای تولید جمعیت اولیه ارائه می‌شود. با استفاده از این راه‌کار، تلاش می‌شود جواب‌هایی تولید شود که به نواحی امیدبخش نزدیک‌تر باشند. همچنین در مقاله [12] روش ارائه‌شده برای خوشه‌بندی داده‌ها، تعداد خوشه‌ها را نیز به‌صورت خودکار مشخص می‌کند.

در مقاله [13] از الگوریتم بهینه‌سازی ازدحام ذرات برای خوشه‌بندی به‌صورت سلسله‌مراتبی استفاده شده است. در این مقاله دو رویکرد خوشه‌بندی با استفاده از الگوریتم تکاملی PSO و خوشه‌بندی سلسله‌مراتبی با استفاده از الگوریتم PSO بررسی شده است. با استفاده از ویژگی‌های خوشه‌بندی سلسله‌مراتبی و همچنین اضافه‌کردن بهینه‌سازی هوش گروهی، خوشه‌بندی داده‌ها انجام می‌شود. در مقاله [14] خوشه‌بندی گراف‌های احتمالاتی بزرگ با استفاده از الگوریتم تکاملی چندجمعیته انجام می‌شود. الگوریتم تکاملی چندین جعیت را مقداردهی می‌کند. هر کدام از جمعیت‌ها به‌عنوان یک نسخه قطعی از گراف احتمالاتی ورودی هستند. ایجاد چندین نسخه قطعی از گراف ورودی با بهره‌گیری از اعمال حد آستانه بر روی یال‌ها امکان‌پذیر می‌شود. هر کروموزم یک راه حل کامل برای خوشه‌بندی را در بردارد.

ازجمله کارهای انجام‌شده در زمینه خوشه‌بندی خودکار داده‌ها با استفاده از الگوریتم‌های تکاملی می‌توان به کارهای [18]-[15] اشاره کرد. در[16] با استفاده از الگوریتم ژنتیک و در [15] با استفاده از الگوریتم بهینه‌سازی ازدحام ذرات و با درنظرگرفتن توابع ارزیابی خوشه‌بندی، تعداد بهینه خوشه‌ها مشخص می‌شود و خوشه‌بندی داده‌ها بر اساس تعداد خوشه بهینه مشخص‌شده، انجام می‌گیرد. همچنین در [17] با بهره‌گیری از الگوریتم تکامل تفاضلی[1] بهبودیافته و در

[1] Differential evolutionary

[18] با درنظرگرفتن سه تابع ارزیابی به‌صورت هم‌زمان و بهره‌گیری از الگوریتم بهینه‌سازی چندهدفه، عمل خوشه‌بندی خودکار داده‌ها انجام می‌شود. الگوریتم بهینه‌سازی چندهدفه به‌کارگرفته‌شده، مبتنی بر الگوریتم شبیه‌سازی تبرید[1] است.

هدف این مقاله، برآورده‌کردن دو مورد، است. یکی مشخص‌کردن تعداد بهینه خوشه‌ها و دیگری خوشه‌بندی بهینه مجموعه داده بدون برچسب است. روش پیشنهادی با استفاده از الگوریتم بهینه‌سازی رقابت استعماری و بهبود آن، این هدف را برآورده می‌کند. با بهره‌گیری از ساختار مناسب برای هر یک از کروموزم‌ها که در بخش ۳-۲ توضیح داده می‌شود و استفاده از الگوریتم رقابت استعماری، در زمان اجرا تعداد بهینه خوشه‌ها هم‌زمان با خوشه‌بندی بهینه داده‌ها به دست می‌آید. همچنین برای افزایش دقت و افزایش سرعت همگرایی، ساختار الگوریتم رقابت استعماری با تغییراتی همراه است. شرح این تغییرات در بخش ۳-۱ آورده شده است. تابع هدف مورد استفاده در الگوریتم بهینه‌سازی، مانند سایر الگوریتم‌های خوشه‌بندی مبتنی بر الگوریتم‌های تکاملی، استفاده از معیارهای ارزیابی خوشه‌بندی است. ایراد این دسته از الگوریتم‌های خوشه‌بندی، وابستگی کارایی آن‌ها به معیار ارزیابی استفاده شده است. یک معیار ارزیابی ناکارا منجر به ایجاد خوشه‌های کاذب می‌تواند شود.

در این مقاله، کارایی الگوریتم به‌وسیله سه معیار ارزیابی می‌شود: ۱. دقت نتایج خوشه‌بندی نهایی ۲. سرعت هم‌گرایی ۳. میزان پایداری جواب‌ها (توانایی تولید جواب‌های یکسان در تکرارهای مختلف الگوریتم). برای به‌دست‌آوردن نتایج از پنج مجموعه داده واقعی استفاده می‌شود.

بخش‌های بعدی این مقاله بدین‌صورت است. در بخش ۲ مسئله خوشه‌بندی به‌صورت رسمی با زبان ریاضی بیان می‌شود. در بخش ۳، الگوریتم رقابت استعماری، روش بهبود عملکرد آن، روش نمایش جواب‌ها و تابع هدف مورد استفاده برای حل مسئله خوشه‌بندی خودکار، شرح داده می‌شود. در بخش ۴، مجموعه داده‌های استفاده‌شده، الگوریتم‌های استفاده‌شده برای مقایسه و پارامترهای آن ذکر می‌شود. در بخش ۵، نتایج حاصل از خوشه‌بندی خودکار با استفاده از ۵ مجموعه داده واقعی نشان داده می‌شود و در بخش ۶ نتیجه‌گیری آورده می‌شود.

[1] Simulated annealing

۲- مفاهیم پایه

۲-۱- تعریف مسئله

یک الگو، یک ساختار فیزیکی یا انتزاعی از اشیا است و تمایزدهنده آن از سایر الگوها مجموعه‌ای از ویژگی‌ها هستند [19]. فرض می‌شود $P=\{P_1,P_2,\ldots,P_n\}$ مجموعه‌ای از n الگو یا نمونه داده‌ای باشند. همچنین این مجموعه الگوها در قالب یک ماتریس $X_{n\times d}$ می‌توانند نمایش داده شوند که n تعداد نمونه‌ها یا الگوها و d تعداد ویژگی‌های آن‌هاست. با فرض داشتن چنین ماتریسی، الگوریتم خوشه‌بندی تفکیکی سعی می‌کند تا مجموعه $C=\{C_1,C_2,\ldots,C_K\}$ با K خوشه را ایجاد کند به‌طوری‌که نمونه‌ها در هر خوشه بیشترین شباهت را نسبت به هم و کمترین شباهت را نسبت به سایر خوشه‌ها داشته باشند. مجموعه C باید ۳ ویژگی زیر را ارضا نماید: ۱. هر خوشه باید حداقل یک نمونه داده داشته باشد. ۲. دو خوشه نباید دارای نمونه داده مشترک باشند. ۳. هر نمونه داده باید به یک خوشه نسبت داده شده باشد. به عبارت بهتر هر نمونه داده باید خوشه آن مشخص باشد.

برای برآورده‌کردن سه ویژگی یادشده نیاز به یک تابع هدف است. بنابراین در مسئله خوشه‌بندی، باید یک مجموعه خوشه بهینه C^* از بین جواب‌های ممکن $C=\{C^1,C^2,\ldots,C^{N(n,K)}\}$ پیدا شود، به‌طوری‌که تعداد جواب‌های ممکن طبق رابطه (۱) به دست می‌آید.

$$N(n,K)=\frac{1}{K!}\sum_{i=1}^{K}(-1)^i\binom{K}{i}(K-i)^i \qquad (۱)$$

در رابطه (۱)، n تعداد نمونه‌های داده‌ای و K تعداد طبقه‌های موجود است.

ارزیابی هر یک از جواب‌ها توسط تابع ارزیابی انجام می‌گیرد. تابع ارزیابی کیفیت خوشه‌بندی را بر مبنای فاصله بین نمونه‌های داده‌ای نشان می‌دهد. طبق [20] نشان داده شده است که مسئله خوشه‌بندی برای تعداد خوشه‌های بیشتر از ۳، یک مسئله NP-hard است.

۲-۲- معیارهای شباهت

همان‌طور که در بخش قبل ذکر شد، خوشه‌بندی، فرایندی است که در نتیجهٔ آن گروهی از خوشه‌ها بر اساس معیار شباهت ایجاد می‌شوند. بنابراین معیارهای شباهت نقش مهمی را در خوشه‌بندی ایفا می‌کنند. یکی از روش‌های رایج برای ارزیابی شباهت بین الگوها، استفاده از معیار فاصله است. بیشترین معیار فاصله مورد استفاده در الگوریتم‌های مختلف

استفاده از فاصله اقلیدسی است که مقدار آن برای دو الگوی d بعدی $\vec{X}_i$ و $\vec{X}_j$ به‌صورت رابطه (۲) به دست می‌آید.

$$D\left(\vec{X}_i,\vec{X}_j\right)=\sqrt{\sum_{p=1}^{d}(X_{i,p}-X_{j,p})^2}=\vec{X}_i-\vec{X}_j \qquad (۲)$$

در رابطه (۲)، D فاصله اقلیدسی بین الگوی X_i و X_j است. فاصله اقلیدسی حالت خاصی از فاصله مینوسکی[1] (به‌ازای α=۲) است. فاصله مینوسکی به‌صورت رابطه (۳) تعریف می‌شود:

$$D^{\alpha}\left(\vec{X}_i,\vec{X}_j\right)=\left(\sum_{p=1}^{d}\left(X_{i,p}-X_{j,p}\right)^{\alpha}\right)^{1/\alpha}=\vec{X}_i-\vec{X}_j{}^{\alpha} \qquad (۳)$$

در رابطه (۳)، D فاصله مینوسکی بین الگوی X_i و X_j است. وقتی که α=1 باشد، به‌عنوان فاصله منهتنی[2] شناخته می‌شود.

فاصله مینوسکی به‌طورمعمول برای خوشه‌بندی مجموعه داده با ابعاد زیاد کارآمد نیست؛ بنابراین بر طبق برای فاصله مینوسکی، ویژگی‌های با مقیاس بالا، ویژگی‌های دیگر را مغلوب می‌کنند. این مشکل می‌تواند با نرمال‌کردن مقادیر ویژگی‌ها برطرف شود. یکی از روش‌ها برای انجام این کار، استفاده از فاصله کسینوسی است که به‌صورت رابطه (۴) تعریف می‌شود.

$$\vec{X}_i,\vec{X}_j=\frac{\sum_{p=1}^{d}X_{i,p}\cdot X_{j,p}}{\vec{X}_i\,\vec{X}_j} \qquad (۴)$$

فاصله کسینوسی، اختلاف زاویه‌ای دو بردار داده‌ای را محاسبه می‌کند و به مقدار آن‌ها توجهی ندارد. معیار فاصله دیگر، فاصله ماهالانبیس[3] است که طبق رابطه (۵) محاسبه می‌شود:

$$D_M\left(\vec{X}_i,\vec{X}_i\right)=\left(\vec{X}_i-\vec{X}_i\right)\Sigma^{-1}\left(\vec{X}_i-\vec{X}_i\right) \qquad (۵)$$

در رابطه (۵)، $\sum$ ماتریس کوواریانس الگوهاست. فاصله ماهالانبیس بر اساس واریانس ویژگی‌ها و وابستگی خطی مابین هر دو جفت ویژگی، وزن‌های مختلفی به ویژگی‌ها نسبت می‌دهد.

۲-۳- معیارهای ارزیابی خوشه‌بندی

معیارهای ارزیابی خوشه‌بندی مطابق با توابع ریاضی- آماری هستند که میزان خوب‌بودن نتیجه یک خوشه‌بندی را نشان می‌دهند. یک معیار ارزیابی خوشه‌بندی دو هدف را دنبال می‌کند: ۱. مشخص‌کردن تعداد خوشه‌ها ۲. به‌دست‌آوردن بهترین حالت خوشه‌بندی با توجه به تعداد خوشه‌ها. هر معیار ارزیابی خوشه‌بندی باید دو وجه خوشه‌بندی زیر را مدنظر قرار دهد: ۱. پیوستگی یا فشردگی: الگوهای موجود در یک خوشه باید تا حد امکان به یکدیگر شبیه باشند. واریانس یا پراکندگی الگوهای موجود در یک خوشه نمایان‌گر پیوستگی یا فشردگی الگوهای درون یک خوشه هستند. ۲. تفکیک: خوشه‌ها باید تا حد امکان از هم فاصله داشته باشند. فاصله بین مراکز خوشه‌ها (به‌عنوان مثال فاصله اقلیدسی) می‌تواند نمایان‌گر تفکیک خوشه‌ها باشد.

معیارهای ارزیابی مختلفی برای ارزیابی خوشه‌بندی غیر فازی وجود دارند که ازجمله آن‌ها می‌توان به معیار DI [20]، DB [21]، CS [22] و PBM [23] اشاره کرد. برای تمام این معیارها، مقدار بیشینه و یا کمینهٔ آن‌ها نشان‌دهنده خوشه‌بندی بهینه مجموعه الگوها و یا داده‌هاست. ازاین‌رو آن‌ها را با الگوریتم‌های بهینه‌سازی فرا ابتکاری از جمله الگوریتم ژنتیک می‌توان استفاده کرد. از بین معیارهای ارزیابی خوشه‌بندی، در این مقاله فقط معیارهای ارزیابی DB و CS مدنظر قرار می‌گیرند.

۱. معیار ارزیابی DB: این معیار تابعی از نسبت مجموع پراکندگی درون‌خوشه‌ای به پراکندگی بین خوشه‌هاست. ابتدا پراکندگی خوشه iام و سپس فاصله بین خوشه i و خوشه j طبق روابط (۶) و (۷) محاسبه می‌شوند.

$$S_{i,q}=\left[\frac{1}{N_i}\sum_{\vec{X}\in C_i}\|\vec{X}-\vec{m}_i\|_2^{q}\right]^{1/q} \qquad (۶)$$

$$D_{ij,t}=\left\{\sum_{p=1}^{d}\left|m_{i,p}-m_{j,p}\right|^{t}\right\}^{1/t}=\|\vec{m}_i-\vec{m}_j\|_t \qquad (۷)$$

که $\vec{m}_i$، مرکز خوشه iام، $q,t\geq 1$، q و t به‌طور مستقل مقداردهی می‌توانند شوند. N_i تعداد الگوهای متعلق به خوشه C_i است. $D_{ij,t}$ نرم[4] tام برای مراکز خوشه C_i (m_i) و خوشه C_j (m_j) را محاسبه می‌کند. مقدار معیار ارزیابی DB به صورت رابطه (۹) تعریف می‌شود.

$$R_{i,qt}=\max_{j\in K,\,j\neq i}\left\{\frac{S_{i,q}+S_{j,q}}{D_{ij,t}}\right\} \qquad (۸)$$

$$DB(K)=\frac{1}{K}\sum_{i=1}^{K}R_{i,qt} \qquad (۹)$$

[1] Minkowski distance

[2] Manhattan distance

[3] Mahalanobis distance

[4] Norm

در رابطه (۸)، K نشان‌دهندهٔ خوشه Kام است. کمترین مقدار به‌دست‌آمده برای معیار DB، نشان‌دهنده خوشه‌بندی بهینه است.

۲. معیار ارزیابی CS: قبل از محاسبه معیار ارزیابی CS، مرکز هر خوشه به‌وسیله میانگین الگوهای آن خوشه طبق رابطه (۱۰) مشخص می‌شود.

$$\vec{m}_i = \frac{1}{N_i}\sum_{x_j \in C_i} \vec{x}_j \qquad (۱۰)$$

در رابطه (۱۰)، N_i تعداد الگوهای متعلق به خوشهٔ C_i است. معیار فاصله بین دو نمونه $\vec{X}_i$ و $\vec{X}_j$ به‌صورت $d\left(\vec{X}_i, \vec{X}_j\right)$ و معیار CS به‌صورت رابطه (۱۱) تعریف می‌شود:

$$CS(K) = \frac{\sum_{i=1}^{K}\left[\frac{1}{N_i}\sum_{\vec{X}_i \in C_i}\left[\max_{\vec{X}_q \in C_i}\left\{d\left(\vec{X}_i, \vec{X}_q\right)\right\}\right]\right]}{\sum_{i=1}^{K}\left[\min_{j \in K, j \neq i}\left\{d\left(\vec{m}_i, \vec{m}_j\right)\right\}\right]} \qquad (۱۱)$$

همانند معیار ارزیابی DB، معیار CS برابر با نسبت فاصله درون‌خوشه‌ای به فاصله بین‌خوشه‌ای است و بنابراین باید صورت کسر کمینه و مخرج بیشینه و در کل مقدار CS کمینه شود.

۳- الگوریتم بهینه‌سازی رقابت استعماری (ICA)

الگوریتم رقابت استعماری [24] از مفاهیم امپراتوری، مستعمره و استعمارگر ایده گرفته است. در هر الگوریتم بهینه‌سازی فراابتکاری یک جمعیت از جواب‌های ممکن ایجاد می‌شود. در الگوریتم رقابت استعماری، جمعیت را امپراتوری‌ها تشکیل می‌دهند. به‌عبارت بهتر هر امپراتوری یک جواب ممکن برای مسئله بهینه‌سازی است. در هر امپراتوری تعدادی کشور وجود دارد. یک کشور که از بقیه کشورها بهتر است، به‌عنوان استعمارگر و سایر کشورها به‌عنوان مستعمره لحاظ می‌شوند.

در الگوریتم بهینه‌سازی رقابت استعماری، دو عمل‌گر جذب و انقلاب وجود دارد. در یک امپراتوری ازآنجاکه کشور استعمارگر از جهات مختلف اقتصادی، سیاسی، فرهنگی و غیره بهتر از کشورهای تحت سلطه و مستعمره‌اش است؛ بنابراین کشورهای مستعمره تمایل به رشد و ترقی در جهت کشور استعمارگرشان دارند. برای شبیه‌سازی حرکت کشورهای مستعمره به سمت کشور استعمارگر از عمل‌گر جذب استفاده می‌شود. عمل‌گر جذب طبق رابطه (۱۲) محاسبه می‌شود.

$$x' = x + \beta\left(t - x\right) \qquad (۱۲)$$

که t موقعیت هدف، x جواب فعلی و x' جواب جدید ایجاد شده است و β ضریب جذب نامیده می‌شود. اگر مقدار β برابر با صفر باشد، جواب جدید با جواب فعلی مساوی و اگر مقدار β برابر با یک باشد، جواب جدید با مقدار t برابر است. به‌طورمعمول مقدار β عددی بین صفر تا ۲ انتخاب می‌شود. ازآنجاکه ممکن است، نقاط واقع در همسایگی موقعیت هدف بهتر از آن باشند، به همین دلیل مقدار ضریب جذب بیشتر از یک انتخاب می‌شود. همچنین برای افزایش جستجو درخصوص موقعیت هدف t، یک انحرافی به اندازه زاویه θ نیز ایجاد می‌شود.

مقدار تابع هدف برای یک امپراتوری، برابر با مجموع مقدار تابع هدف به‌ازای جواب استعمارگر و ضریبی (ξ) از میانگین مقادیر تابع هدف برای جواب‌های تحت سلطه آن است. به عبارت بهتر:

میانگین مقادیر تابع هدف به‌ازای جواب‌های مستعمره × ξ + مقدار تابع هدف به‌ازای جواب استعمارگر = مقدار شاخص برای هر امپراتوری مقدار ضریب ξ معمولاً برابر با ۰٫۱ در نظر گرفته می‌شود.

در این الگوریتم بهینه‌سازی، دو نوع رقابت وجود دارد: ۱. رقابت درون‌گروهی: پس از انجام عمل‌گرهای جذب و انقلاب، بهترین جواب موجود در یک امپراتوری جایگزین استعمارگر می‌شود. ۲. رقابت میان گروهی: پس از یک تکرار الگوریتم، یکی از اعضای ضعیف‌ترین امپراتوری به تصادف انتخاب و به اعضای قوی‌ترین امپراتوری اضافه می‌شود.

مراحل الگوریتم ICA:

۱. ایجاد کشورهای اولیه
۲. انتخاب بهترین کشورها به‌عنوان استعمارگر
۳. تخصیص سایر کشورها به‌عنوان مستعمره به استعمارگرها
۴. اعمال عمل‌گر جذب
۵. اعمال عمل‌گر انقلاب (ابتدا با یک احتمال ازپیش‌تعریف‌شده، بر روی ژن‌های یک جواب تغییرات تصادفی (جهش) روی می‌دهد و با یک احتمال

ازپیش‌تعریف‌شده دیگر، جواب‌های جدید تولیدشده حاصل از جهش بر روی جواب‌ها اعمال می‌شود)

۶. مقایسه مستعمرات با استعمارگرها و جایگزینی استعمارگر با مستعمره مربوطه (درصورتی‌که مستعمره بهتر از استعمارگر باشد)
۷. ارزیابی امپراتوری‌ها (محاسبه شاخص مربوط به هر امپراتوری)
۸. یک مستعمره از ضعیف‌ترین امپراتوری حذف و به تصادف به یک امپراتوری دیگر منتقل می‌شود.
۹. اگر ضعیف‌ترین امپراتوری دارای هیچ مستعمره‌ای نباشد، استعمارگر مربوطه به‌عنوان مستعمره به یک امپراتوری دیگر منتقل می‌شود.
۱۰. گزارش بهترین پاسخ یافته‌شده
۱۱. بازگشت به مرحله (۴) در صورتی که شرایط خاتمه محقق نشده باشد.
۱۲. پایان

۳-۱- بهبود الگوریتم رقابت استعماری

در این مقاله، ساختار الگوریتم رقابت استعماری بررسی و از آن برای خوشه‌بندی خودکار داده‌ها استفاده می‌شود. برای افزایش دقت نتایج حاصل از خوشه‌بندی و همچنین افزایش سرعت هم‌گرایی آن، تغییراتی در ساختار الگوریتم رقابت استعماری داده می شود.

در ابتدا برای افزایش نرخ جستجو و کشف راه‌حل‌های ممکن، باید گام‌های بلندتری برداشته شود. در ادامه با نزدیک‌شدن به جواب بهینه سراسری، این گام‌ها باید کاهش پیدا کند تا هم جواب بهینه یافته‌شده از دست نرود و هم الگوریتم، در بهینه محلی گرفتار نشود. برای این منظور نرخ جهش و نرخ اعمال عمل‌گر انقلاب طبق روابط زیر به‌صورت پویا تعیین می‌شوند.

$$prevolution = (pmax - pmin) \times \frac{maxit - it}{maxit} \qquad (۱۳)$$

$$mu(t) = mu(t) \times \left(\frac{muf}{mu(t)} \right)^{\frac{it}{maxit}} \qquad (۱۴)$$

در روابط بالا، prevolution احتمال انقلاب در کشورهای مستعمره، mu(t) احتمال جهش در لحظه t، muf مقدار نهایی احتمال جهش (مقدار نهایی مورد نظر در تکرار آخر الگوریتم)، maxit تعداد کل تکرارهای الگوریتم، it تعداد تکرار فعلی الگوریتم، pmax و pmin به‌ترتیب مقدار بیشینه و کمینه برای احتمال انقلاب هستند.

۳-۲- نحوه نمایش جواب‌ها

ازآنجاکه در ابتدا تعداد خوشه‌ها توسط کاربر مشخص نشده است، نحوه نمایش جواب‌ها باید طوری باشد که بتوان تعداد خوشه‌ها را هم‌زمان با انجام خوشه‌بندی داده‌ها در زمان اجرا به‌دست آورد. بنابراین با فرض اینکه تعداد بیشینه خوشه‌ها برابر با K و تعداد ابعاد یا ویژگی‌های مجموعه داده برابر با d باشد، آنگاه هر یک از جواب‌های ممکن یک ماتریس با ابعاد k×(d+1) خواهد بود؛ زیرا برای به‌دست‌آوردن تعداد بهینه خوشه‌ها، یک متغیر تصمیم به متغیرهای تصمیم موجود برای حل مسئله خوشه‌بندی اضافه می‌شود. به‌عبارتی هر یک از مرکز خوشه‌ها که دارای d بعد است یک متغیر تصمیم به نام متغیر تصمیم فعال‌سازی دارد که مشخص می‌کند خوشه مربوطه فعال باشد و یا نباشد. مقادیر مجاز برای متغیر تصمیم فعال‌سازی اعداد حقیقی در بازه (۰،۱) است. مرکز یک خوشه درصورتی فعال می‌شود که مقدار متغیر تصمیم مربوط به آن بزرگ‌تر از ۰،۵ باشد. در غیر این صورت مرکز خوشه غیر فعال است. درواقع عدد ۰،۵ آستانه فعال‌سازی خوشه محسوب می‌شود. همان‌طور که در شکل (۱) نشان داده شده است، مرکز خوشه iام دارای d بعد است که یک بعد برای متغیر تصمیم فعال‌سازی نیز به ساختار جواب اضافه می‌شود.

(شکل-۱): نمایش مرکز خوشه iام با درنظر گرفتن متغیر تصمیم فعال‌سازی

(Figure-1): Representation of ith cluster center considering activation decision variable

به‌عنوان مثال با فرض اینکه مجموعه داده دارای سه بعد یا ویژگی و بیشینه تعداد خوشه‌ها برابر با چهار باشد، یکی از جواب‌های ممکن به‌صورت شکل (۲) است. ازآنجاکه مقدار متغیر تصمیم فعال‌سازی برای مرکز خوشه دوم کمتر از مقدار حد آستانه فعال‌سازی خوشه یعنی ۰،۵ و مابقی بیشتر از ۰،۵ است، مرکز خوشه دوم در نظر گرفته نمی‌شود و تعداد خوشه‌ها سه در نظر گرفته می‌شود.

	مرکز خوشه اول			
خوشه اول	2.5	5	6.1	0.58
خوشه دوم	5.1	2.3	9.4	0.48
خوشه سوم	3.8	6	2	0.8
خوشه چهارم	7.4	9.4	1	0.64

(شکل-۲): نمایش یک راه حل ممکن برای حل مسئله خوشه‌بندی

(Figure-2): Representation of a possible solution to solve the clustering problem

۳-۳- تابع هدف

برای ارزیابی جواب‌ها در الگوریتم رقابت استعماری، هر یک از معیارهای ارزیابی خوشه‌بندی می‌توانند مورد استفاده قرار گیرند. در این مقاله دو معیار ارزیابی خوشه‌بندی DB و CS به‌عنوان تابع هدف طبق روابط (۱۵) و (۱۶) در نظر گرفته می‌شوند که هدف کمینه‌سازی مقدار آن‌ها برای مسئله خوشه‌بندی است. نتایج به‌ازای درنظرگرفتن هر کدام از آن‌ها به‌عنوان تابع هدف در قسمت نتایج آورده می‌شود (K تعداد خوشه‌هاست).

$$f_1 = CS\left(K\right) \qquad (۱۵)$$

$$f_2 = DB\left(K\right) \qquad (۱۶)$$

۳-۴- شبه‌کد روش پیشنهادی (ACICA)

۱. مقداردهی اولیه جواب‌ها
۲. پیداکردن خوشه‌های فعال با استفاده از مقدار متغیر تصمیم فعال‌سازی
۳. به‌ازای هر تکرار الگوریتم:

a. فاصله هر یک از نمونه‌ها تا هر یک از مراکز خوشه‌ها محاسبه می‌شود.
b. هر کدام از نمونه‌ها به خوشه‌ای نسبت داده می‌شود که کمترین فاصله را تا مرکز آن خوشه داشته باشد.
c. هر کدام از خوشه‌ها بررسی می‌شوند. اگر خوشه‌ای کمتر از دو نمونه داده داشته باشد، مجموعه داده به n/k تقسیم می‌شود. میانگین هر کدام از قسمت‌های داده‌ای ایجادشده به‌عنوان یک مرکز خوشه لحاظ می‌شود و جایگزین مراکز خوشه‌های قبلی می‌شود.
d. الگوریتم رقابت استعماری بهبودیافته بر روی مراکز خوشه‌ها اعمال می‌شود (استفاده از عمل‌گرهای جذب و انقلاب و ارزیابی جواب‌های جدید).

۴. با استفاده از بهترین جواب به‌دست‌آمده حاصل از گام سوم، مراکز و تعداد بهینه خوشه‌ها گزارش می‌شود.

۴- نتایج

در این بخش، نتایج حاصل از روش پیشنهادی بر اساس معیارهای ارزیابی خوشه‌بندی و همچنین تعداد فراخوانی‌های تابع هدف برای رسیدن به یک مقدار ازپیش تعریف‌شده برای معیار DB، ذکر می‌شود و با سه الگوریتم خوشه‌بندی به نام‌های GUCK [16]، DCPSO [15] و ACDE [17] مقایسه می‌شود. الگوریتم GUCK با استفاده از الگوریتم بهینه‌سازی ژنتیک، الگوریتم DCPSO با استفاده از الگوریتم بهینه‌سازی ازدحام ذرات و الگوریتم ACDE با استفاده از الگوریتم تکامل تفاضلی بهبودیافته، به خوشه‌بندی خودکار الگوها می‌پردازند.

۴-۱- مجموعه داده‌ها

۱. مجموعه‌داده Iris: این مجموعه‌داده دارای ۱۵۰ الگو است. هر یک از این الگوها دارای چهار ویژگی و تعداد خوشه‌ها برابر با سه است (n=150، d=4، K=3).
۲. مجموعه‌داده Wine: این مجموعه‌داده دارای ۱۷۸ الگو است. هر یک از این الگوها دارای سیزده ویژگی و تعداد خوشه‌ها برابر با سه است (n=178، d=13، K=3).
۳. مجموعه‌داده Glass: این مجموعه داده دارای ۲۱۴ الگو می‌باشد. هر یک از این الگوها دارای نه ویژگی و تعداد خوشه‌ها برابر با شش است (n=214، d=9، K=6).
۴. مجموعه‌داده Vowel: این مجموعه‌داده دارای ۸۷۱ الگو است. هر یک از این الگوها دارای سه ویژگی و تعداد خوشه‌ها برابر با شش است (n=871، d=3، K=6).
۵. مجموعه‌داده Wisconsin breast cancer: این مجموعه‌داده دارای ۶۸۳ الگو است. هر یک از این الگوها دارای نه ویژگی و تعداد خوشه‌ها برابر با دو است (n=683، d=9، K=2).

۴-۲- مقادیر پارامترها

مقادیر مورد استفاده برای پارامترها برای الگوریتم‌های DCPSO، GCUK و ACDE به‌ترتیب طبق [15]، [16] و [17] در جدول (۱) و همچنین مقادیر مورد استفاده برای

پارامترها برای الگوریتم رقابت استعماری در جدول (۲) آمده است. مقادیر بهینه مورد استفاده در جدول (۲)، با بررسی مقادیر مختلف برای الگوریتم رقابت استعماری به‌دست‌آمده است.

مقادیر مربوط به تعداد خوشه‌های تعیین‌شده و مقدار تابع هدف مربوط به معیارهای CS و DB در جدول (۳) آورده شده‌اند. منظور از Classic ICA، نسخه اصلی الگوریتم رقابت استعماری به کار گرفته شده برای خوشه‌بندی خودکار است و ACICA نسخه بهبودیافته الگوریتم رقابت استعماری به کار گرفته شده برای خوشه‌بندی خودکار الگوها است. همچنین تعداد فراخوانی‌های تابع هدف برای رسیدن به مقدار ازپیش‌تعریف‌شده برای معیار ارزیابی DB، برای مجموعه‌های داده‌های مختلف در جدول (۴) آمده است.

(جدول-۱): مقادیر پارامترها برای GCUK و DCPSO

(Table-1): Parameters values for GCUK and DCPSO

GCUK		DCPSO	
تعداد اعضای جمعیت	50	تعداد اعضای جمعیت	100
احتمال ترکیب	0.8	وزن لختی (ω)	0.72
احتمال جهش	0.001	C_1 و C_2	1.494
k_{min} k_{max}	2 20	k_{min} k_{max}	2 20

(جدول-۲): مقادیر پارامترها برای ACDE و ACICA

(Table-2): Parameters values for ACDE and ACICA

ACDE		ACICA	
تعداداعضای جمعیت	۱۰×تعداد ابعاد مسئله	تعداد اعضای جمعیت	۱۰×تعداد ابعاد مسئله
کمینه احتمال ترکیب	0.5	تعداد امپراتوری	10
بیشینه احتمال ترکیب	1.0	کمینه احتمال انقلاب بیشینه احتمال انقلاب	0.5 1.0
k_{min} k_{max}	2 20	k_{min} k_{max}	2 20

(جدول-۳): مقادیر CS و DB

(Table-3): The values of DB and CS

مجموعه داده	الگوریتم	میانگین تعداد خوشه‌های تعیین شده	معیار CS	معیار DB
Iris	DCPSO	2.23 ± 0.0443	0.7361 ± 0.671	0.6899 ± 0.008
	GCUK	2.35 ± 0.0985	0.7282 ± 2.003	0.7377 ± 0.065
	ACDE	3.25 ±0.0382	0.6643 ±0.097	0.4645 ±0.022
	Classic ICA	3.5 ±0.0505	0.2463 ±0.0603	0.4011 ±0.011
	ACICA	3.15 ±0.008	0.2132 ±0.541	0.356 ±0.025
Wine	DCPSO	3.05 ± 0.0352	1.8721 ± 0.037	4.3432 ± 0.232
	GCUK	2.95 ± 0.0112	1.5842 ± 0.328	5.3424 ± 0.343
	ACDE	3.25 ±0.0391	0.9249 ±0.032	3.0432 ±0.021
	Classic ICA	3.5833 ±0.752	0.3924 ±0.0794	2.855 ±0.115
	ACICA	2.9167 ±0.7299	0.3402 ±0.757	2.565 ±0.001
Breast-Cancer	DCPSO	2.25 ± 0.0632	0.4854 ± 0.009	0.5754 ± 0.073
	GCUK	2.00 ± 0.0083	0.6089 ± 0.016	0.6328 ± 0.002
	ACDE	2.00 ±0.00	0.4532 ±0.034	0.5203 ±0.006
	Classic ICA	2.110 ±0.002	0.4122 ±0.012	0.5011 ±0.002
	ACICA	2.001 ±0.003	0.3907 ±0.002	0.4725 ±0.001
Vowel	DCPSO	7.025 ± 0.0183	1.1827 ± 0.431	1.2821 ± 0.009
	GCUK	5.05 ± 0.0075	1.9978 ± 0.966	2.9482 ± 0.028
	ACDE	5.75 ±0.0751	0.9089 ±0.051	0.9224 ±0.334
	Classic ICA	5.30 ±0.022	0.802 ±0.120	0.8525 ±0.544
	ACICA	5.89 ±0.0435	0.791 ±0.021	0.8001 ±0.021
Glass	DCPSO	5.95 ± 0.0346	0.7642 ± 0.073	1.5152 ± 0.073
	GCUK	5.85 ± 0.0093	1.4743 ± 0.236	1.8371 ± 0.073
	ACDE	6.05 ±0.0148	0.3324 ±0.487	1.0092 ±0.083
	Classic ICA	6.0833 ±0.225	0.3082 ±0.0248	1.0011 ± 0.002
	ACICA	5.9167 ±0.005	0.2811 ±0.0305	0.9581 ±0.012

(جدول-۴): تعداد فراخوانی‌های تابع هدف

(Table-4): The number of fitness function evaluations

مجموعه داده	الگوریتم	میانگین تعداد فراخوانی‌های تابع ارزیابی	مقدار DB موردنظر
Iris	DCPSO	679084.75 ± 16.57	0.8
	GCUK	790865.90 ± 10.21	
	ACDE	504783.45 ± 12.65	
	Classic ICA	512001.21 ± 10.25	
	ACICA	496584.52 ± 2.61	
Wine	DCPSO	486885.85 ± 2.85	6.0
	GCUK	598743.35 ± 8.09	
	ACDE	464653.35 ± 5.50	
	Classic ICA	486536.85 ± 2.32	
	ACICA	456589.25 ± 5.62	
Breast-Cancer	DCPSO	467854.60 ± 10.12	0.9
	GCUK	678874.90 ± 7.82	
	ACDE	424732.30 ± 8.93	
	Classic ICA	452698.36 ± 4.70	
	ACICA	448965.54 ± 6.60	
Vowel	DCPSO	556865.00 ± 4.26	3.0
	GCUK	575854.65 ± 1.29	
	ACDE	435743.05 ± 2.65	
	Classic ICA	432689.21 ± 7.81	
	ACICA	412255.36 ± 9.14	
Glass	DCPSO	569787.95 ± 10.83	2.0
	GCUK	687678.75 ± 10.97	
	ACDE	506754.00 ± 12.27	
	Classic ICA	516647.23 ± 2.55	
	ACICA	502265.47 ± 3.01	

۵- نتیجه‌گیری

از آنجا که در کاربردهای مختلفی، نیاز به خوشه‌بندی داده‌هایی وجود دارد که تعداد خوشه‌ها از پیش مشخص نیست، نیاز به روش‌هایی وجود دارد که بتوانند بدون داشتن دانش درباره طبقه داده‌ها، آنها را خوشه‌بندی کرد و در عین حال پیش بینی صحیحی از تعداد خوشه‌ها انجام داد. در این مقاله سعی شد با استفاده از یکی از روش‌های فراابتکاری به نام الگوریتم رقابت استعماری، خوشه‌بندی داده‌ها بدون داشتن دانش در مورد تعداد خوشه‌ها، انجام شود.

از مجموعه داده‌هایی که تعداد خوشه‌ها از پیش مشخص نیست، می‌توان به داده‌های متنی اشاره کرد. برای کارهای آینده می‌توان از روش پیشنهادی مطرح‌شده در این مقاله برای خوشه‌بندی متون بهره برد. خوشه‌بندی متون کاربردهای زیادی در پردازش زبان طبیعی و بازیابی اطلاعات مانند خلاصه‌سازی متن دارد.

۶- مراجع 6- Refrences

[1] I. Evangelou, "DG Hadjimitsis, AA Lazakidou, Clayton,"" in *Data Mining and Knowledge Discovery in Complex Image Data using Artific-ial Neural Networks", Workshop on Complex Reasoning an Geographical Data, Cyp-rus*, 2001.

[2] T. Lillesand, R. W. Kiefer, and J. Chipman, *Remote sensing and image interpretation*. John Wiley & Sons, 2014.

[3] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 2013.

[4] H. Frigui and R. Krishnapuram, "A robust competitive clustering algorithm with applica-tions in computer vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 21, no. 5, pp. 450-465, 1999.

[5] Y. Leung, J.-S. Zhang, and Z.-B. Xu, "Clustering by scale-space filtering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 22, no. 12, pp. 1396-1410, 2000.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR),* vol. 31, no. 3, pp. 264-323, 1999.

[7] J. Holland, "Adaption in natural and artificial systems," *Ann Arbor MI: The University of Michigan Press,* 1975.

[8] S. Z. Selim and K. Alsultan, "A simulated annealing algorithm for the clustering problem," *Pattern recognition,* vol. 24, no. 10, pp. 1003-1008, 1991.

[9] V. Di Gesú, R. Giancarlo, G. L. Bosco, A. Raimondi, and D. Scaturro, "GenClust: A genetic algorithm for clustering gene expression data,"

BMC bioinformatics, vol. 6, no. 1, p. 289, 2005.

[10] A. Beg, M. Z. Islam, and V. Estivill-Castro, "Genetic algorithm with healthy population and multiple streams sharing information for clustering," *Knowledge-Based Systems,* vol. 114, pp. 61-78, 2016.

[11] D. Bajer, G. Martinović, and J. Brest, "A population initialization method for evolutionary algorithms based on clustering and Cauchy deviates," *Expert Systems with Applications,* vol. 60, pp. 294-310, 2016.

[12] J. de Andrade Silva, E. R. Hruschka, and J. Gama, "An evolutionary algorithm for clustering data streams with a variable number of clusters," *Expert Systems with Applications,* vol. 67, pp. 228-238, 2017.

[13] S. Alam, G. Dobbie, and S. U. Rehman, "Analysis of particle swarm optimization based hierarchical data clustering approach-es," *Swarm and Evolutionary Computation,* vol. 25, pp. 36-51, 2015.

[14] Z. Halim, M. Waqas, and S. F. Hussain, "Clustering large probabilistic graphs using multi-population evolutionary algorithm," *Information Sciences,* vol. 317, pp. 78-95, 2015.

[15] M. G. Omran, A. Salman, and A. P. Engelbrecht, "Dynamic clustering using particle swarm optimization with application in image segmentation," *Pattern Analysis and Applications,* vol. 8, no. 4, pp. 332-344, 2006.

[16] S. Bandyopadhyay and U. Maulik, "Genetic clustering for automatic evolution of clusters and application to image classification," *Pattern Recognition,* vol. 35, no. 6, pp. 1197-1208, 2002.

[17] S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on,* vol. 38, no. 1, pp. 218-237, 2008.

[18] S. Saha and S. Bandyopadhyay, "A generalized automatic clustering algorithm in a multiobjective framework," *Applied Soft Computing,* vol. 13, no. 1, pp. 89-108, 2013.

[19] A. Konar, *Computational intelligence: principles, techniques and applications*. Springer Science & Business Media, 2006.

[20] P. Brucker, "On the complexity of clustering problems," in *Optimization and operations research*: Springer, 1978, pp. 45-54.

[21] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* no. 2, pp. 224-227, 1979.

[22] C.-H. Chou, M.-C. Su, and E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Analysis and Applications,* vol. 7, no. 2, pp. 205-220, 2004.

[23] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern recognition,* vol. 37, no. 3, pp. 487-501, 2004.

[24] E. Atashpaz-Gargari and C. Lucas, "Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition," in *Evolutionary computation, 2007. CEC 2007. IEEE Congress on*, 2007, pp. 4661-4667: IEEE.

آرش چاقری مدرک کارشناسی خود را در رشته مهندسی رایانه گرایش نرم‌افزار از دانشگاه خوارزمی و مدرک کارشناسی ارشد خود را در رشته مهندسی رایانه گرایش هوش مصنوعی از دانشگاه شهید بهشتی تهران اخذ کرده است. ایشان هم‌اکنون دانشجوی دکترای مهندسی رایانه گرایش هوش مصنوعی در دانشگاه تبریز است. زمینه‌های پژوهشی مورد علاقه ایشان، پردازش زبان‌های طبیعی، داده‌کاوی و الگوریتم‌های بهینه‌سازی است.

نشانی رایانامه ایشان عبارت است از:

a.chaghari@tabrizu.ac.ir

محمدرضا فیضی درخشی دارای دکترا و کارشناسی ارشد مهندسی رایانه در گرایش هوش مصنوعی از دانشگاه علم و صنعت ایران بوده و مدرک کارشناسی خود را نیز در رشته مهندسی رایانه، گرایش نرم‌افزار از دانشگاه اصفهان اخذ کرده است. ایشان هم‌اکنون عضو هیئت علمی گروه مهندسی رایانه دانشگاه تبریز است. زمینه‌های پژوهشی مورد علاقه ایشان پردازش زبان‌های طبیعی، پردازش معنایی وب و اسناد، الگوریتم‌های بهینه‌سازی و پایگاه داده است.

نشانی رایانامه ایشان عبارت است از:

mfeizi@tabrizu.ac.ir