New Researches in the Smart City
Vol 1 (Issue 4), 6-25
Nrsc.apadana.ac.ir

Apadana Institute of Higher Education

Improving the efficiency of data clustering with chaotic evolutionary algorithms

Sajad Manteghi *
Head of Technology and Information Department, General Directorate of Education, Kohgiluyeh and Boyer-Ahmad Province, Iran

Sara Khosravani Pour
Instructor, Computer Group, General Directorate of Education and Training, Kohgiluyeh and Boyer-Ahmad Province, Iran.

Received: 27/January/2023 Accepted: 15/August/2023 ISSN: 2980-8936

Abstract

Nowadays, clustering plays an important role in most research fields such as engineering, medicine, biology, data mining, etc. In fact, clustering means unsupervised division. By using it, the data are divided into categories that are more similar to each other in terms of the parameters of interest. One of the famous methods in this field is k-means. In this method, despite the dependence on initial conditions and convergence to local optimal points, N numbers of data are grouped into k clusters with high speed. In this article, to solve the existing problems, the combined method is used based on evolutionary algorithms, chaos theory and k-means; that is in addition to solving the mentioned problems, it will also be independent of the number of variables. In this article, for the purpose of validation, the proposed methods are implemented on 13 different famous collections, and the results are compared with genetic algorithm, particle community, bee colony, simulated refrigeration, differential evolution, harmony search, and k-means methods. The high ability and robustness of these methods will be evident based on the results.

Keywords: clustering, k-means algorithm, evolutionary algorithms, chaos, chaotic evolutionary algorithm

* Corresponding Author: S.manteghi@eng.ui.ac.ir

فصلنامه پژوهش‌های نوین در شهر هوشمند
دوره ۱، شماره ٤، تابستان ۱٤۰۲، ٦-۲۵
Nrsc.apadana.ac.ir

بهبود کارآیی خوشه‌بندی داده‌ها با الگوریتم‌های تکاملی آشوب‌گونه

سجاد منطقی * ریاست فناوری و اطلاعات، اداره کل آموزش و پرورش، کهگیلویه و بویراحمد، ایران.

سارا خسروانی پور مربی گروه کامپیوتر، اداره کل آموزش و پرورش، کهگیلویه و بویراحمد، ایران.

تاریخ دریافت: ۱۴۰۱/۱۱/۰۷ تاریخ پذیرش: ۱۴۰۲/۰۵/۲۴ ISSN: 2980-8936

چکیده

امروزه، خوشه‌بندی نقش مهمی را در اغلب زمینه‌های تحقیقاتی مانند مهندسی، پزشکی، زیست‌شناسی، داده‌کاوی و ... ایفا می‌نماید. در واقع خوشه‌بندی به معنای تقسیم‌بندی بدون نظارت می‌باشد. داده‌ها با استفاده از آن به دسته‌هایی که از نظر پارامترهای موردعلاقه، شباهت بیشتری به یکدیگر دارند، تقسیم می‌گردند. یکی از روش‌های معروف در این زمینه k-means می‌باشد. در این روش علی‌رغم وابستگی به شرایط اولیه و همگرایی به نقاط بهینهٔ محلی، تعداد N داده به k خوشه با سرعت بالا، دسته‌بندی می‌شوند. در این مقاله جهت رفع مشکلات موجود از روش ترکیبی مبتنی بر الگوریتم‌های تکاملی و تئوری آشوب و k-means بهره گرفته خواهد شد؛ که علاوه بر رفع مشکلات ذکرشده، مستقل از تعداد متغیرها نیز خواهد بود. در این مقاله به‌منظور اعتبارسنجی، روش‌های پیشنهادی بر روی ۱۳ مجموعهٔ متفاوت مشهور پیاده‌سازی می‌گردد و نتایج با روش‌های الگوریتم ژنتیک، اجتماع ذرات، کلونی زنبور عسل، تبرید شبیه‌سازی‌شده، تکاملی تفاضلی، جستجوی هارمونی و k-means مقایسه خواهند گردید. توانایی بالا و مقاوم‌بودن این روش‌ها بر اساس نتایج مشهود خواهد بود.

کلیدواژه‌ها: خوشه‌بندی، الگوریتم k-means، الگوریتم‌های تکاملی، آشوب، الگوریتم تکاملی آشوب‌گونه

* نویسندهٔ مسئول: S.manteghi@eng.ui.ac.ir

مقدمه

پردازش داده، یکی از شاخص‌های بسیار مهم در دنیای اطلاعات است. خوشه‌بندی یکی از بهترین روش‌هایی است که برای کار با داده‌ها ارائه شده است. خوشه‌بندی قابلیت ورود به فضای داده و تشخیص ساختار داده‌ها را امکان‌پذیر می‌نماید. لذا به‌عنوان یکی از ایده‌آل‌ترین روش‌ها، برای کار با دنیای عظیم داده‌ها محسوب می‌شود. خوشه‌بندی، یافتن ساختاری در مجموعه‌ای از داده‌ها است که طبقه‌بندی نشده‌اند. به بیان دیگر می‌توان گفت که خوشه‌بندی قراردادن داده‌ها در گروه‌هایی است که اعضای هر گروه از جهات خاصی شبیه یکدیگرند. معیار شباهت در این‌جا، فاصله می‌باشد؛ یعنی نمونه‌هایی که به یکدیگر نزدیکند در یک خوشه قرار می‌گیرند. با محاسبهٔ فاصلهٔ بین دو داده می‌توان فهمید که چقدر این دو داده به هم نزدیک هستند و بر این اساس در یک خوشه قرار داده می‌شوند. توابع ریاضی مختلفی برای محاسبهٔ فاصله وجود دارند؛ فاصلهٔ اقلیدسی، فاصلهٔ همینگ و ... .

خوشه‌بندی یک فرآیند بدون ناظر است که یک مجموعه از اشیاء را به گروه‌های متجانس تقسیم می‌کند. خوشه بندی یکی از مهم‌ترین روش‌های داده‌کاوی می‌باشد ولی درزمینه‌های دیگری مثل شناخت الگو، فن‌آوری اطلاعات، پردازش تصاویر، زیست‌شناسی، روانشناسی و بازاریابی نیز به کار می‌رود. هدف اصلی خوشه‌بندی پیداکردن یک ساختار معنادار برای داده‌ها می‌باشد. الگوریتم k-means یکی از معروف‌ترین الگوریتم‌های خوشه‌بندی است که در سال ۱۹۷۹ انتشار یافت و هنوز هم به طور گسترده‌ای از آن استفاده می‌شود. متأسفانه این الگوریتم وابسته به مقادیر اولیهٔ مراکز خوشه‌هاست و به همین دلیل همیشه خوشه‌بندی به طور کاملاً صحیح انجام نمی‌شود. با توجه به قرارگیری خوشه‌بندی در دستهٔ مسائل غیرقطعی - سخت، لزوم استفاده از الگوریتم‌های فرامکاشفه‌ای و تکاملی جهت دست‌یابی به پاسخ مناسب در زمان منطقی مشخص می‌باشد.

بهینه‌سازی شاخه‌ای از علوم ریاضیات است که در آن سعی می‌شود نقطهٔ بهینهٔ توابع را با توجه به تعدادی محدودیت به دست آورد. امروزه بسیاری از مسائل بهینه‌سازی اغلب از نوع مسائل غیرچندجمله‌ای سخت هستند. از جمله راه-حل‌های موجود در برخورد با این‌گونه مسائل، استفاده از الگوریتم‌های تقریبی یا مکاشفه‌ای است. این‌گونه الگوریتم‌ها تضمینی نمی‌دهند که جواب به‌دست‌آمده بهینه باشد. امروزه الگوریتم‌های تکاملی نظیر الگوریتم ژنتیک، بهینه‌سازی انبوه ذرات، کلونی مورچگان و غیره در حل بسیاری از مسائل پیچیده که روش‌های عددی قادر به حل آن‌ها نیستند و یا در حل آن دچار مشکلاتی می‌شوند، کاربرد دارند. مسئلهٔ خوشه‌بندی نیز یکی از این چنین مسائل است و با توجه به کاربرد بالای این موضوع، محققین همواره در پی یافتن روشی کامل و دقیق برای حل این مسئله بوده‌اند؛ اما با توجه به ماهیت این موضوع، روش‌های عددی موجود قابلیت حل این مسئله را ندارند و معضل پاسخ‌های محلی در این روش‌ها مشکل‌ساز بوده است.

الگوریتم‌های تکاملی الهام‌گرفته از پدیده‌های طبیعی هستند. این الگوریتم‌ها از مولدهای تصادفی برای تولید جمعیت اولیه استفاده می‌کنند؛ ولی در طبیعت پدیده‌ها تصادفی نیستند و از نظم خاصی پیروی می‌کنند. در این مقاله در الگوریتم‌های تکاملی از پدیدهٔ آشوب به‌منظور تولید اعداد شبه‌تصادفی برای خوشه‌بندی استفاده شده است. آشوب نوعی بی‌نظمی منظم است. بی‌نظمی از آن جهت که نتایج آن غیرقابل‌پیش‌بینی است و منظم بدان جهت که از نوعی قطعیت برخوردار است؛ بنابراین می‌توان راه‌کاری جامع‌تر و کامل‌تر ارائه نمود. لذا هدف از این مقاله ارائهٔ الگوریتم-های خوشه‌بندی بر مبنای الگوریتم‌های تکاملی آشوب‌گونه می‌باشد که بتواند نسبت به الگوریتم‌های مورد مقایسه بهبودی را حاصل نماید. در این مقاله چندین الگوریتم تکاملی آشوب‌گونه همانند الگوریتم ژنتیک آشوب‌گونه، الگوریتم بهینه‌سازی انبوه ذرات آشوب‌گونه و الگوریتم کلونی زنبور آشوب‌گونه برای مسئلهٔ خوشه‌بندی پیشنهاد شده است.

الگوریتم‌های تکاملی در مقایسه با سایر الگوریتم‌های بهینه‌سازی برتری‌هایی دارند که موجب شده است به طور گسترده مورد استفاده قرار بگیرند. هم‌چنین محدودیتی در مورد تابع شایستگی ندارند و لزومی ندارد که این تابع مثلاً مشتق‌پذیر باشد. علاوه بر این موارد، چون الگوریتم‌های تکاملی دارای جمعیتی از موجودات هستند و روی بخش‌های مختلفی از جمعیت به طور موازی کار می‌کنند، احتمال کمتری برای قرارگرفتن در بهینه‌های محلی دارند. این قابلیت الگوریتم‌های تکاملی اجازه می‌دهد که کار بهینه‌سازی را به طور موازی روی چندین بخش جمعیت انجام داد. از بین الگوریتم‌های تکاملی می‌توان به موارد زیر اشاره کرد:

- الگوریتم ژنتیک[1] (Holland, 1975).
- الگوریتم تبرید شبیه‌سازی‌شده[2] (Alfonzetti et al., 2006).
- الگوریتم بهینه‌سازی ازدحام ذرات[3] (Kennedy & Eberhart, 1995).
- الگوریتم رقابت استعماری[4] (Atashpaz-Gargari & Lucas, 2007).
- الگوریتم کلونی مورچگان[5] (Dorigo & Gambardella, 1997).
- الگوریتم کلونی زنبور عسل[6] (Karaboga & Basturk, 2007).

طبق تحقیقاتی که انجام شده الگوریتم‌های تکاملی برای خوشه‌بندی مناسب هستند (Kao et al., 2008; Niknam et al., 2011)؛ ولی تا به حال الگوریتم‌های تکاملی آشوب‌گونه برای خوشه‌بندی استفاده نشده‌اند که بدانیم تأثیرات آن‌ها در مسئلهٔ خوشه‌بندی چیست. آشوب در لغت به معنی درهم‌ریختگی، آشفتگی و بی‌نظمی است و مترادف آن در مکانیک، تلاطم است. این واژه به معنی فقدان هر گونه ساختار بانظم است و معمولاً در محاورات، آشوب و آشفتگی نشانهٔ بی‌نظمی و سازمان‌نیافتگی به نظر آمده و جنبهٔ منفی در بر دارد؛ اما در واقع با پیدایش نگرش جدید و روشن‌شدن ابعاد علمی و نظری آن، امروزه دیگر بی‌نظمی و آشوب به مفهوم سازمان‌نیافتگی، ناکارآیی و درهم‌ریختگی تلقی نمی‌گردد؛ بلکه وجود جنبه‌های غیرقابل‌پیش‌بینی و اتفاقی در پدیده‌های پویا است که ویژگی‌های خاص خود را دارند (Schuster & Just, 2006). تئوری آشوب مطالعهٔ رفتار سیستم‌های پویا است که به شرایط اولیه حساسیت بالایی دارند. تغییر اندکی در شرایط اولیهٔ چنین سیستم‌هایی باعث اثرات بسیار در آینده خواهد شد (Kellert, 1993).

هدف در این پژوهش این است که با بررسی الگوریتم‌های موجود درزمینهٔ خوشه‌بندی، راهکاری برای ایجاد تغییرات در الگوریتم‌های موجود برای بالابردن سرعت هم‌گرایی و بهبود کیفیت خوشه‌بندی ارائه گردد به‌گونه‌ای که محدودیت‌های موجود را پوشش دهد. از نمونهٔ محدودیت‌های موجود می‌توان موارد زیر را برشمرد:

- کارآیی برای پایگاه داده‌های با حجم بالا.
- کشف خوشه‌ها با اشکال مختلف.
- عدم‌حساسیت به ترتیب داده‌های ورودی.
- قابلیت تفسیر و استفاده.

1. Genetic Algorithm
2. Simulated Annealing Algorithm
3. Particle Swarm Optimization Algorithm
4. Imperialist Competitive Algorithm
5. Ant Colony Algorithm
6. Ant Bee Colony Algorithm

کارهای گذشته

در سال ۱۹۹۹ یک الگوریتم ژنتیک ترکیبی جدید به‌منظور تقسیم‌بندی بهینه سراسری یک دادهٔ دریافتی به تعداد مشخصی خوشه توسط کریشنا[1] و همکارانش معرفی شده است (Krishna & Murty, 1999). در این روش یک الگوریتم ژنتیک ترکیب‌شده با گرادیان نزولی کلاسیک در خوشه‌بندی یعنی k-means استفاده شده است. این الگوریتم k-means با ژنتیک (GKA) نامیده می‌شود. در این روش، از الگوریتم k-means در GKA به‌عنوان یک عملگر جستجو به‌جای عملگر آمیزش استفاده می‌شود. همچنین یک عملیات جهش مغرضانهٔ خاص برای خوشه‌بندی تعیین می‌گردد که فاصله بر مبنای جهش نامیده می‌شود. از تئوری مارکوف محدودشده نیز در روش مذکور استفاده شده و نشان داده شده است که GKA به بهینهٔ سراسری همگرا می‌شود. جهش به کار برده شده در روش GKA به این صورت است که تغییرات جهش یک مقدار به فاصلهٔ نقطهٔ داده از مراکز خوشه بستگی دارد. در بسیاری از فراخوانی‌ها هر آلل به نقطهٔ داده بستگی دارد و این مقدار نشان‌دهندهٔ خوشه‌ای است که به آن تعلق دارد. یک عملیات مثل احتمال تغییرات یک آلل برای تعداد اعضای یک خوشه درصورتی‌که مراکز خوشه به داده نزدیک‌تر باشد؛ تعریف شده است. برای اعمال عملگر جهش برای آلل $s_w(i)$ با توجه به الگوی x_i از $d_j = d(x_i, c_j)$ که از فاصلهٔ اقلیدسی بین x_i و c_j است استفاده می‌شود.

موالیک[2] و همکارانش در سال ۲۰۰۰ یک الگوریتم به نام خوشه‌بندی ژنتیک برای مسئلهٔ خوشه‌بندی ارائه کرده‌اند. مزیت این روش استفاده از الگوریتم k-means در ژنتیک است. روش مذکور بر روی ۴ مجموعهٔ داده آزمایش شده است (Maulik & Bandyopadhyay, 2000). در این روش هر رشته دنباله‌ای از اعداد حقیقی است که نشان‌دهندهٔ k مرکز خوشه است. برای یک فضای nبعدی طول یک کروموزوم N*K است که N موقعیت اول نشان‌دهندهٔ Nبعد مرکز خوشهٔ اول است و N موقعیت بعدی نیز مرکز خوشهٔ بعدی و بقیه نیز به همین صورت هستند. جمعیت اولیه به صورت تصادفی تولید می‌شود. برازندگی بر حسب معیار فاصله محاسبه می‌شود به گونه‌ای که ابتدا خوشه‌ها با توجه به مراکز کدگذاری شده در کروموزوم موردنظر تشکیل می‌شوند. پس از تشکیل خوشه‌ها مرکز خوشه‌ها در کروموزوم با میانگین نقاط درون خوشه جایگزین می‌شود.

الگوریتمی مبتنی بر کلونی مورچگان برای مسئلهٔ خوشه‌بندی توسط شلوکار[3] و همکارانش در سال ۲۰۰۴ ارائه شد. در این الگوریتم عوامل توزیعی به کار گرفته می‌شود که در آن از روش واقعی مورچه‌ها برای پیداکردن کوتاه‌ترین مسیر از لانهٔ خود به سمت منبع غذا و بازگشت استفاده شده است (Shelokar et al., 2004). هدف این الگوریتم اختصاص بهینه از N شیء عضو $\mathcal{R}^n$ در K خوشه است به‌گونه‌ای که مجموع مربعات فاصلهٔ اقلیدسی بین شیء و مرکز خوشه‌ای که به آن تعلق دارد کمینه باشد. الگوریتم برای ایجاد راه‌حل‌ها R عامل در نظر می‌گیرد. یک عامل با یک رشتهٔ راه‌حل خالی S با طول N شروع می‌شود به‌گونه‌ای که هر عنصر از رشته با یکی از نمونه‌های آزمایشی مطابق است. مقدار اختصاص‌یافته به یک نمونه از رشتهٔ راه‌حل S بیانگر شمارهٔ خوشه‌ای است که نمونهٔ آزمایشی به S اختصاص داده است. برای ساخت یک راه‌حل، عامل از اطلاعات دنباله‌ای فرومون برای اختصاص هر یک از المان‌های رشتهٔ S به یک خوشهٔ مناسب استفاده می‌کند.

در ابتدای الگوریتم، ماتریس فرومون τ، با مقدار کوچک τ_0 مقداردهی اولیه می‌گردد. مقدار دنباله $\tau_{i,j}$ در مکان (i,j) نشان‌دهندهٔ فرومون جمع‌شده از نمونه iام از خوشهٔ jام است. برای مسئلهٔ تقسیم N نمونه درون K خوشه، اندازهٔ ماتریس فرومون $N \times K$ است. در هر مرحله از تکرار، هر یک از عامل‌ها یا مورچه‌های نرم‌افزار همانند دنباله‌ای از

1. Krishna
2. Mualik
3. Shelokar

راه‌حل‌ها را که از فرآیند ارتباطات فرومون برای به دست آوردن یک بخش نزدیک به بهینه از N نمونهٔ دریافتی به K گروه به گونه‌ای که هدف مورد نظر حاصل گردد استفاده می‌کنند. پس از تولید یک نسل از R راه‌حل‌های آزمایشی، یک جستجوی محلی برای بهبود برازندگی این راه‌حل‌ها اجرا می‌شود. ماتریس فرومون با توجه به کیفیت راه‌حل‌های تولیدشده توسط عامل‌ها به‌روز می‌شود. هدایت توسط ماتریس فرومون بهبودیافته صورت می‌گیرد. عامل‌ها راه‌حل‌های بهبودیافته را می‌سازند و مراحل بالا را تا رسیدن به تعداد از قبل معین‌شده تکرار می‌کنند.

یک روش مبتنی بر الگوریتم بهینه‌سازی زنبور عسل برای هدف خوشه‌بندی در سال ۲۰۰۸ توسط فتحیان و همکارانش پیشنهاد شده است (Fathian et al., 2007). این الگوریتم در چند گام در ادامه آورده شده است:

گام ۱: نمایش رشته: برای راه‌حل انتخابی مسئله از یک کروموزوم استفاده می‌شود که هر ژن درون کروموزوم بیانگر یک پارامتر از راه‌حل انتخابی است. در این روش، یک کروموزوم به عنوان مجموعه‌ای از k مرکز خوشهٔ ابتدایی و هر ژن به عنوان ابعاد مرکز خوشه در نظر گرفته شده است.

گام ۲: تعیین پارامترهای ورودی مدل: الگوریتم با سه پارامتر تعیین‌شده توسط کاربر و یک الگوریتم از پیش تعیین‌شده شروع می‌شود.

گام ۳: تولید تصادفی مجموعه‌ای از راه‌حل‌های اولیه: یک مجموعه از مراکز خوشه‌های اولیه به صورت تصادفی از نقاط مجموعه داده تولید می‌شوند. هر راه‌حل نشان‌دهنده‌ی مراکز خوشه k است.

گام ۴: انتخاب ملکه: بعد از تولید راه‌حل‌های اولیهٔ مذکور، راه‌حل‌های اولیه بر اساس تابع کارآیی تخمین زده می‌شود و بهترین راه‌حل به عنوان ملکه قرار می‌گیرد.

گام ۵: جفت‌گیری: از تبرید شبیه‌سازی‌شده برای انتخاب مجموعه‌ای از راه‌حل‌های موجود در فضای جستجو برای انجام جفت‌گیری بین بهترین راه‌حل و راه‌حل‌های انتخابی استفاده می‌شود.

گام ۶: فرآیند تولید مثل.

گام ۷: نوزادان انتخابی و ملکه توسط کارگران با ژل‌های سلطنتی تغذیه می‌شوند.

گام ۸: اگر بهترین راه‌حل جدید از ملکه بهتر باشد، ملکه با راه‌حل تولیدی جایگزین می‌شود.

گام ۹: معیارهای پایانی بررسی می‌گردد. اگر معیارهای پایانی ارضاء شدند الگوریتم پایان می‌یابد در غیر این صورت راه‌حل‌های قبلی دور ریخته می‌شوند و راه‌حل‌های جدید تولید شده و به مرحله ۵ برمی‌گردند.

یک الگوریتم تکاملی ترکیبی برای ترکیب الگوریتم تبرید شبیه‌سازی‌شده با بهینه‌سازی انبوه ذرات اصلی توسط نیک‌نام و همکارانش در سال ۲۰۰۸ پیشنهاد شده است (Niknam & Amiri, 2010). در الگوریتم پیشنهادی از تبرید شبیه‌سازی‌شده به عنوان یک جستجوگر محلی اطراف Gbest استفاده می‌شود؛ به عبارت دیگر Gbest در هر تکرار تغییر می‌کند. این الگوریتم PSO-SA نامیده می‌شود. متغیرهای کنترلی مراکز خوشه‌ها هستند.

یک تکنیک ترکیبی مبتنی بر ترکیب الگوریتم‌های k-means، جستجوی ساده Nelder-Mead و بهینه‌سازی ازدحام ذرات توسط کائو و همکارانش در سال ۲۰۰۸ پیشنهاد شده است. این الگوریتم پیشنهادی K-NM-PSO نامیده می‌شود. جستجوی K-NM-PSO برای مراکز خوشه از یک مجموعهٔ دادهٔ دلخواه برای الگوریتم k-means است؛ اما با استفاده از این الگوریتم می‌توان به صورت مؤثر بهینهٔ سراسری را یافت (Kao et al., 2008).

یک الگوریتم جدید بر اساس ترکیب دو الگوریتم خوشه‌بندی K-means و رقابت استعماری بهبودیافته توسط نیک‌نام و همکارانش در سال ۲۰۱۲ پیشنهاد می‌گردد که این ترکیب K-MICA نامیده می‌شود و از روش *حداکثرسازی انتظار اصلاح‌شده (EM)* برای تعیین تعداد خوشه‌ها استفاده شده است (Niknam et al., 2011).

روش پیشنهادی

الگوریتم‌های بهینه‌سازی به روش هوشمند، پیشرفت قابل‌ملاحظه‌ای در مقابل روش‌های ریاضی از خود نشان داده‌اند. در این تحقیق از روش جدیدی مبتنی بر ترکیب الگوریتم‌های تکاملی (هم‌چون ژنتیک، بهینه‌سازی انبوه ذرات و کلونی زنبور)، پدیدهٔ آشوب و الگوریتم k-means، برای خوشه‌بندی اطلاعات استفاده خواهد شد. در واقع هدف این است که داده‌های موجود را بتوان به‌نحوی با حداقل خطا خوشه‌بندی نمود. زمانی خوشه‌بندی ایده‌آل میسر می‌گردد و هدف نهایی برآورده شود و حداقل شباهت بین داده‌های موجود در دسته‌های مجزا، هم‌چنین حداکثر شباهت بین داده‌های موجود در یک دسته برقرار باشد.

در این راستا از مزایای الگوریتم k-means، آشوب و الگوریتم‌های تکاملی بهره گرفته خواهد شد.

روش k-means علی‌رغم معایبش که در ادامه توضیح داده می‌شود، به دلیل سادگی و سهولت شبیه‌سازی آن بسیار کاربرد دارد.

- *وابستگی به شرایط اولیه:* مقادیر اولیه در این روش پاسخ نهایی را دست‌خوش تغییرات قرار می‌دهد و خوشه‌های به‌دست‌آمده بنا به مقادیر اولیه بسیار متفاوت خواهند بود.
- *همگرایی به نقاط بهینهٔ محلی:* در صورتی که تابع هدف دارای نقاط بهینهٔ زیادی باشد استفاده از این روش جواب اصلی را نتیجه نخواهد داد.

جهت مرتفع‌نمودن این معایب، در تحقیقات اخیر ایده‌های متفاوتی بیان گردیده است که تا حد قابل‌قبولی کارساز بوده است. الگوریتم‌های تکاملی ارائه‌گردیده در این زمینه را می‌توان به شرح زیر برشمرد؛ ژنتیک، اجتماع ذرات و زنبور مصنوعی.

با ترکیب‌نمودن الگوریتم k-means و الگوریتم‌های تکاملی آشوب‌گونه، می‌توان به هدف مورد نظر که خوشه‌بندی اطلاعات می‌باشد رسید. در این مسیر دقت جواب نهایی به طور چشمگیری افزایش یافته است. همچنین نگرانی در خصوص همگرایی پاسخ به بهینهٔ محلی تا حدود بسیار زیادی برطرف گردیده است و بدون در نظر گرفتن تعداد متغیرها و مقادیر اولیه جواب مطلوب به دست خواهد آمد. در ادامه الگوریتم‌های تکاملی آشوب‌گونه شرح داده خواهد شد.

الگوریتم تکاملی مبتنی بر آشوب

در این بخش الگوریتمی ترکیبی از الگوریتم‌های تکاملی و پدیدهٔ آشوب ارائه شده است. در روش‌های پیشنهادی از الگوریتم‌های تکاملی ژنتیک، بهینه‌سازی گروه ذرات، تکاملی تفاضلی، کلونی زنبور و رقابت کشورهای استعماری استفاده شده است. آشوب شاخه‌ای از حوزه‌های تحقیقات غیرخطی پویا است که به طور گسترده مورد مطالعه قرار گرفته است. موارد کاربردی زیادی در سیستم‌های واقعی، چه ساخت انسان و چه طبیعی، با استفاده از روش‌های نوین در حوزهٔ غیرخطی است که به‌ظاهر اتفاقی به نظر می‌رسد. هرچند این ظاهر تصادفی، منشأ اتفاقی ندارد و به‌کلی نتیجهٔ یک پردازش قطعی و تعریف شده است. بیش از صد سال است که زیست‌شناسان نمونه‌هایی از جمعیت گونه‌های مختلف را مورد بررسی قرار داده‌اند. به‌تازگی دانشمندان قادر به ایفاکردن مدل‌های آشوب‌گونه در جمعیت‌های خاصی بوده‌اند (Liz & Ruiz-Herrera, 2012). به عنوان مثال، مطالعه روی سیاه‌گوش کانادایی نشان داد رفتار آشوب‌گونه در رشد جمعیت وجود دارد (Lai, 1996). به همین منظور در این پژوهش تغییری در الگوریتم‌های تکاملی مبتنی بر جمعیت انجام گرفته و پدیدهٔ آشوب با الگوریتم‌های تکاملی ترکیب شده تا از مولدهای آشوبی در تولید جمعیت الگوریتم‌ها استفاده شود. در این مقاله از الگوریتم‌های تکاملی آشوب‌گونه برای خوشه‌بندی، با هدف

دست‌یابی به خوشه‌بندی با پایداری بالا استفاده شده است. فرمول تولید جمعیت با پدیدهٔ آشوب در ادامه ذکرشده است (Neal, 1993).

$$f(x) = 1 - 2 \times \left| u - \frac{1}{2} \right| \quad (۱)$$

در فرمول بالا u یک عدد تصادفی در محدودهٔ [۱ و ۰] است.

الگوریتم‌های پیشنهادی

در روش پیشنهادی از الگوریتم‌های تکاملی آشوب‌گونه به‌منظور خوشه‌بندی داده‌ها استفاده شده است. در روش‌های پیشنهادی پس از تولید جمعیت به روش آشوب‌گونه، الگوریتم k-means روی آخرین عضو جمعیت، جهت بهبود عملکرد اجرا می‌گردد. در شکل‌های ۱ تا ۴ شبه‌کد الگوریتم‌های پیشنهادی نشان داده شده است.

— ذره را با استفاده از فرمول زیر تولید کن:

$$Paticle(i) = 1 - 2 \times \left| u - \frac{1}{2} \right|$$

— $Paticle(i)$ ذره iام جمعیت است و u عددی بین [۱ و ۰] است.

— اجرای الگوریتم k-means بر روی آخرین فرد جمعیت.

— برای تمام ذرات، سرعت و موقعیتی تصادفی ایجاد کن.

— **تا زمانی که** شرایط خاتمه محقق نشده است:

- یک واحد به t اضافه کن.
- مقدار تابع هدف را به ازای هر ذره محاسبه کن.
- به ازای i از یک تا n (n تعداد جمعیت):
  - بهینهٔ محلی را بیاب و در $x^{i,best}[t]$ ذخیره کن.
- **مقدار بعدی i.**
- بهینهٔ سراسری را محاسبه کن و در $x^{gbest}[t]$ قرار بده.
- **به ازای** i از یک تا n:
  - **به ازای** j از یک تا d:

$$v_j^i[t+1] = wv_j^i[t] + c_1 r_1 \left(x_j^{i,best}[t] - x_j^i[t]\right) + c_2 r_2 \left(x_i^{gbest}[t] - x_j^i[t]\right)$$
$$x_j^i[t+1] = x_j^i[t] + v_j^i[t+1]$$

  - **مقدار بعدی** j.

مقدار بعدی i.

شکل (۱): الگوریتم خوشه‌بندی بهینه‌سازی ازدحام ذرات آشوب‌گونه.

- تولید جمعیت اولیه[1] شامل n کروموزوم به صورت آشوب‌گونه.
- اجرای الگوریتم k-means روی آخرین فرد جمعیت.
- بررسی تابع ارزیابی f(x) برای هر کروموزوم x در جمعیت.
- ایجاد یک جمعیت جدید بر اساس تکرار گام‌های زیر:
  - انتخاب دو کروموزوم والد از یک جمعیت بر اساس میزان برازندگی آن‌ها.
  - در نظر گرفتن مقدار مشخصی برای احتمال اعمال عملگر ترکیب[2] (تقاطع) و سپس انجام عملیات ترکیب بر روی والدین به‌منظور ایجاد فرزندان (اگر هیچ ترکیب جدیدی صورت نگیرد، فرزندان همان والدین خواهند بود).
  - در نظر گرفتن احتمال جهش و سپس تغییر فرزندان در هر مکان.
  - جایگزینی فرزندان جدید در جمعیت.
- استفاده از جمعیتی جدید برای اجراهای بعدی الگوریتم.
- توقف اجرای الگوریتم در صورت مشاهدهٔ شرایط توقف و برگرداندن بهترین جواب در جمعیت فعلی.
- رفتن به مرحله ۲.

شکل (۲): الگوریتم خوشه‌بندی ژنتیک آشوب‌گونه.

- تولید جمعیت اولیه به صورت آشوب‌گونه.
- اعمال الگوریتم k-means بر روی آخرین فرد جمعیت.
- **تا زمانی که** یک امپراتوری باقی‌مانده است مراحل زیر را تکرار کن:
  - **تا زمانی که** هیچ مستعمره‌ای در امپراتوری وجود ندارد:
    - انتقال مستعمرات به سمت کشورهای امپریالیستی آن‌ها.
    - **اگر** هزینهٔ مستعمره‌ای از امپریالیست خود بیشتر بود:
      - موقعیت مستعمره با امپریالیست را تعویض کن.
    - تابع هزینه را برای تمام امپراتوری‌ها را محاسبه کن.
    - ضعیف‌ترین مستعمره از ضعیف‌ترین امپراتوری را به امپراتوری قوی بده.
  - ضعیف‌ترین امپراتوری را حذف کن.

شکل (۳): الگوریتم خوشه‌بندی رقابت استعماری آشوب‌گونه.

1. Initial population
2. crossover

- تولید جمعیت اولیه به صورت آشوب‌گونه.
- اجرای الگوریتم k-means بر روی آخرین فرد جمعیت.
- محاسبگی شایستگی هر فرد جمعیت و انتخاب بهترین فرد جمعیت.
- تا زمانی که شرایط توقف برقرار نشده مراحل زیر را تکرار کن.
  - **جهش:** برای هر فرد جمعیت مثل x یک بردار مانند u از جهش بردار x تولید کن.
  - **ترکیب:** برای هر فرد جمعیت مثل x آن را با بردار u ترکیب کن و یک فرزند با نام z تولید کن.
  - **انتخاب**: اگر شایستگی فرزند یعنی z بهتر از شایستگی والد یعنی x باشد آن‌گاه z را جایگزین x کن.

شکل (٤): الگوریتم خوشه‌بندی تکامل تفاضلی آشوب‌گونه

- تولید جمعیت اولیه به صورت آشوب‌گونه.
- اجرای الگوریتم k-means روی آخرین فرد جمعیت.
- محاسبگی شایستگی هر یک از زنبورهای جمعیت اولیه.
- تا زمانی که شرایط توقف برقرار نشده گام‌های زیر تکرار می‌گردد:
  - انتخاب مکان‌های مناسب جهت جستجو در همسایه‌ها
  - استفاده از زنبورها جهت یافتن مکان مناسب و تخمین برازندگی
  - انتخاب بهترین زنبورهای در هر مسیر
  - زنبورهای باقی‌مانده به جستجوی تصادفی می‌پردازند و برازندگی هر یک از آن‌ها محاسبه می‌گردد.

شکل (٥): الگوریتم خوشه‌بندی کلونی زنبور آشوب‌گونه

روش‌های محاسبه کیفیت خوشه‌بندی

روش‌های مختلفی برای محاسبهٔ کیفیت خوشه‌بندی وجود دارد، در این پژوهش از معیارهای فیشر (FM)، اطلاعات متقابل نرمال شده (NMI) و AR به‌منظور ارزیابی کیفیت خوشه‌بندی روش‌های مختلف استفاده شده است. در ادامه معیارهای استفاده شده در پژوهش مورد بررسی قرار می‌گیرند:

معیار F-Measure از ایده‌های دقت و فراخوانی به‌منظور بازیابی اطلاعات استفاده می‌کند. هر کلاس i به‌صورت مجموعه‌ای از n_i بخش مطلوب برای یک تحقیق و پرس‌وجو در نظر گرفته می‌شود. هر خوشه j (که توسط الگوریتم تولید شده) به‌صورت مجموعه‌ای از n_j بخش بازیابی‌شده توسط یک پرس‌وجو در نظر گرفته می‌شود (Coley, 1999). n_{ij} تعداد عناصر کلاس i در خوشهٔ j را می‌دهد. برای هر کلاس i و خوشهٔ j معیارهای دقت و فراخوانی به صورت زیر تعریف می‌شوند:

$$R(i,j) = \frac{n_{ij}}{n_i} \quad (۲)$$

$$P(i,j) = \frac{n_{ij}}{n_j} \quad (۳)$$

مقدار F-Measure متناظر نیز به‌صورت زیر محاسبه می‌گردد (Perruchet & Peereman, 2004):

$$FM = 2 \times \frac{P.R}{P+R} \quad (۴)$$

واضح است که هر چقدر مقدار F-Measure بیشتر باشد، کیفیت خوشه‌بندی نیز بیشتر می‌شود.

برای استفاده از معیار NMI، ابتدا تمام نمونه‌های دیگر متعلق به مجموعهٔ دادهٔ D که در خوشهٔ C_i قرار ندارند؛ به صورت یک خوشهٔ مستقل D/C_i نمایش داده می‌شود. حال یک خوشه‌بندی شامل دو خوشهٔ C_i و D/C_i ایجاد شده است که آن را P_1 می‌نامیم که $P_1=\{C_i, D/C_i\}$ می‌باشد. اکنون خوشه‌بندی $P(D)$ که روی داده‌های نمونه‌برداری-شده اعمال شده است، نیز باید به‌صورت دو خوشه‌ای ارائه شود تا درنهایت نتایج حاصل از این دو خوشه‌بندی طی فرآیندی با هم منطبق شوند. برای این منظور همهٔ خوشه‌ها در $P(D)$ به دو خوشهٔ C^* و D/C^* تقسیم می‌شوند. خوشه C^* از اجتماع همه خوشه‌هایی که بیش از ۵۰٪ از نمونه‌هایشان در خوشهٔ C_i وجود دارند، تشکیل می‌شود و مابقی خوشه‌ها نیز در خوشهٔ D/C^* قرار می‌گیرند. این خوشه‌بندی را P_2 می‌نامیم که $P2=\{C^*, D/C^*\}$ است. حال از اطلاعات متقابل[1] نرمال‌شده (NMI) [۲۱] که معیار متداول برای ارزیابی شباهت بین دو افراز (نتیجهٔ خوشه‌بندی) است، برای اندازه‌گیری شباهت بین دو خوشه‌بندی P_1 و P_2 استفاده می‌شود. از آن‌جایی که معیار اطلاعات متقابل نرمال‌نشده (MI)، وابسته به اندازهٔ خوشه‌هاست، معمولاً از معیار NMI استفاده می‌شود. رابطهٔ NMI بین دو خوشه‌بندی P_1 و P_2 به‌صورت زیر محاسبه می‌شود:

$$NMI(P_1,P_2) = \frac{MI(P_1,P_2)}{\frac{-1}{2m}\left(\sum_{i=0}^{l} P_i \log\frac{P_i}{m} + \sum_{j=0}^{l} P_j \log\frac{P_j}{m}\right)} \quad (۵)$$

که اطلاعات متقابل، $MI(P_1,P_2)$ از رابطه زیر به دست می‌آید.

$$MI(P_1,P_2) = \sum_{i=0}^{l}\sum_{j=0}^{l} \frac{r_{ij}}{m^2}.\log\frac{mp_{ij}}{r_{ij}}, \quad (۶)$$

$$r_{ij} = p_i.p_j \quad , \quad p_i = p_{i0} + p_{i1} \quad , \quad p_j = p_{0j} + p_{1j}$$

که در این رابطه، p_{11} نشان‌دهندهٔ تعداد نمونه‌های مشترک موجود در خوشه‌های C^* و C_i است. p_{10} نشان‌دهندهٔ تعداد نمونه‌های مشترک موجود در D/C^* و C_i است. p_{01} نشان‌دهندهٔ تعداد نمونه‌های مشترک موجود در C^* و D/C_i است. p_{00} نشان‌دهندهٔ تعداد نمونه‌های مشترک موجود در D/C^* و D/C_i است. هم‌چنین m تعداد کل نمونه‌هاست. در واقع p_i و p_j به ترتیب بیانگر کل نمونه‌های موجود در C_i و C^* هستند. *معیار NMI* نشان‌دهندهٔ میزان شباهت خوشه‌بندی P_1 و P_2 می‌باشد.

نتایج تجربی

در این بخش نتایج به‌دست‌آمده از الگوریتم پیشنهادی بر روی داده‌های معروف، مورد بررسی قرار خواهد گرفت. آن‌گاه، پاسخ نهایی الگوریتم‌های پیشین با الگوریتم موردبحث در این تحقیق، با هم مقایسه شده و با سه معیار NMI، FM، AR مورد ارزیابی قرار می‌گیرد.

به طور کلی مجموعهٔ داده‌های مختلفی برای انجام فرآیند خوشه‌بندی وجود دارد؛ که هر کدام دارای ویژگی‌های خاص خود هستند. لذا برای دستیابی به اهداف موجود؛ مجموعهٔ داده‌های Iris، Wine، Glass، Breast Cancer، Inosphere، Galaxy و Half-ring برگزیده می‌شوند. این مجموعهٔ داده از سایت معروف یادگیری ماشین (UCI) انتخاب شده‌اند.

1. Mutual Information (MI)

مجموعهٔ دادهٔ Wine از آزمایشگاه MCI گرفته شده است که بر اساس فعل و انفعالات شیمیایی شراب به دست آمده است. این مجموعهٔ داده شامل ۱۷۸ نمونه است که هر نمونه ۱۳ ویژگی پیوسته دارد. این مجموعهٔ داده در ۳ کلاس متفاوت طبقه‌بندی می‌شود.

با توجه به جدول ۱ برای مجموعهٔ دادهٔ Wine، جواب بهینه برای الگوریتم‌های Chaotic GA مقدار ۱۶۲۹۲.۳۲، Chaotic ICA مقدار ۱۶۲۹۲.۳۵ و Chaotic ABC مقدار ۱۶۲۹۲.۳۸ می‌شود؛ در حالی که الگوریتم‌های دیگر تحت هیچ شرایطی توانایی رسیدن به این مقدار را ندارند. همان‌طور که می‌دانیم انحراف معیار یکی از شاخص‌های پراکندگی است که نشان می‌دهد به طور متوسط داده‌ها چه مقدار از مقدار متوسط فاصله دارند. اگر مقدار انحراف معیار کوچک باشد، نشانهٔ این است که داده‌ها به میانگین نزدیک هستند و پراکندگی اندکی دارند و انحراف معیار بزرگ بیانگر پراکندگی قابل‌توجه در داده‌ها است. انحراف معیار الگوریتم Chaotic ICA و Chaotic ABC از الگوریتم‌های دیگر کمتر است. همان‌طور که در جدول قابل‌مشاهده است الگوریتم‌های Chaotic PSO، Chaotic ICA، Chaotic DE، Chaotic GA و Chaotic ABC در مقایسه با الگوریتم‌های PSO، ICA، DE، GA و ABC زمان اجرای کمتری دارند.

جدول (۱): نتایج میانگین ۱۰۰ بار اجرای الگوریتم‌های تکاملی موجود بر روی مجموعهٔ دادهٔ Wine.

الگوریتم‌های تکاملی	تابع هزینه			انحراف معیار	زمان اجرا
	حداقل هزینه	متوسط هزینه	حداکثر هزینه		
Chaotic PSO	16314.17	16352.85	16381.54	±(15.98344)	1.8298s
Chaotic ICA	16292.35	16294.32	16296.62	±(1.10363)	1.6710s
Chaotic DE	16303.97	16324.65	16346.55	±(11.34668)	1.7786s
Chaotic GA	16292.32	16298.65	16308.73	±(3.97853)	1.8626s
Chaotic ABC	16292.38	16295.39	16301.51	±(2.28436)	3.1025s
PSO	16322.79	16366.69	16443.59	±(23.68992)	2.0696s
ICA	16294.0	16298.28	16307.78	±(2.56136)	1.7847s
SA	18777.46	34923.27	73115.70	±(12856.19664)	0.6742s
DE	16309.41	16359.06	16440.56	±(25.18457)	1.8319s
GA	16294.97	16300.18	16313.53	±(3.84541)	3.8392s
ABC	16311.50	16339.87	16379.03	±(12.39669)	6.2975s
HS	16400.99	16626.69	16828.79	±(94.67190)	5.9578s
K-means	16555.68	16886.12	18294.85	±(685.714)	0.0130s

در جدول ۲ نتایج ارزیابی ۱۰۰ بار اجرای مستقل، هر یک از الگوریتم‌های مورد بررسی با معیارهای NMI، FM و AR قرار داده شده است. با توجه به مطالب ذکرشده بالاتربودن مقدار معیارهای NMI، FM و AR بیانگر کیفیت بالاتر خوشه‌بندی است. با توجه به جدول، الگوریتم Chaotic ICA نتایج بهتری را نسبت به دیگر الگوریتم‌های آشوب‌گونه دارد و یکی از الگوریتم‌های با کیفیت بالا با توجه به معیارهای ذکرشده محسوب می‌شود.

جدول (۲): نتایج ارزیابی داده‌های Wine با معیارهای NMI، FM و AR.

الگوریتم‌های تکاملی	معیار NMI	معیار FM	معیار AR	میانگین
Chaotic PSO	0.424027	0.718891	0.377163	0.506694
Chaotic ICA	0.424152	0.718469	0.377488	0.506703
Chaotic DE	0.424066	0.718569	0.377294	0.506643
Chaotic GA	0.419431	0.716046	0.37068	0.502052
Chaotic ABC	0.417808	0.717771	0.376033	0.503871
PSO	0.424991	0.719348	0.378295	0.507545
ICA	0.422833	0.717690	0.375627	0.505384
SA	0.377444	0.725092	0.303719	0.468752
DE	0.425012	0.719281	0.377949	0.507414
GA	0.421006	0.716142	0.373345	0.503498
ABC	0.422463	0.717911	0.3749483	0.505108
HS	0.426345	0.717424	0.3764747	0.506748
K-means	0.424974	0.705746	0.364192	0.498304

مجموعهٔ دادهٔ Breast Cancer مربوط به اطلاعات غدهٔ سرطانی بیماران مبتلا به سرطان سینه است که در دانشگاه Wisconsin گردآوری شده است. در این مجموعهٔ داده ۶۸۳ نمونه در دو کلاس خوش‌خیم و بدخیم کلاس‌بندی شده است. هر یک از نمونه‌های این مجموعهٔ داده شامل ۹ ویژگی با مقادیر پیوسته است.

مقدار پاسخ بهینهٔ به‌دست‌آمده از حل مسئله بر روی مجموعهٔ دادهٔ Breast Cancer برای الگوریتم Chaotic ICA، ۱۱۱۹.۱۲۴ و برای الگوریتم Chaotic GA، ۱۱۱۹.۸۰۱ خواهد بود. الگوریتم‌های دیگر نسبت به آن دارای مقدار هزینهٔ بیشتری هستند؛ به‌نحوی که انحراف معیار کمتری از الگوریتم‌های دیگر دارند. با مقایسهٔ نتایج به‌دست‌آمده از جدول ۳، این مسئله واضح می‌گردد که عملکرد الگوریتم‌های پیشنهادی مناسب بوده است و سرعت اجرایی الگوریتم‌های آشوب‌گونه نیز بالاتر است.

جدول (۳): نتایج میانگین ۱۰۰ بار اجرای الگوریتم‌های تکاملی موجود بر روی مجموعهٔ دادهٔ Breast Cancer

الگوریتم‌های تکاملی	تابع هزینه			انحراف معیار	زمان اجرا
	حداقل هزینه	متوسط هزینه	حداکثر هزینه		
Chaotic PSO	1129.214	1143.954	1157.590	± (5.13322)	1.9617s
Chaotic ICA	1119.017	1119.124	1119.704	±(0.12401)	1.7867s
Chaotic DE	1121.562	1152.211	1199.244	±(21.2836)	1.7970s
Chaotic GA	1119.413	1119.801	1120.435	±(0.19226)	1.9979s
Chaotic ABC	1123.524	1127.810	1129.270	±(1.44462)	3.4208s
PSO	1130.852	1144.140	1154.574	±(4.51902)	2.1781s
ICA	1119.177	1128.943	1174.003	± (14.96884)	2.0071s
SA	1438.788	1948.092	2695.564	±(273.36210)	1.2121s
DE	1132.111	1182.066	1226.464	±(22.48397)	1.9954s
GA	1121.378	1125.219	1134.412	±(2.58264)	4.1158s
ABC	1124.143	1133.078	1143.261	±(4.14139)	7.7236s
HS	1333.659	1621.6620	1813.978	±(80.60777)	4.4225s
K-means	1128.335	1128.709	1129.270	±(0.46024)	0.0143s

با توجه به نتایج کارآیی روش‌های موردبررسی با استفاده از معیارهای NMI، FM و AR که در جدول ۴ نشان داده شده است، میانگین مقدار معیارهای ذکرشده برای الگوریتم‌های Chaotic ICA و Chaotic GA، به ترتیب مقادیر ۰٫۸۷۸۶۵۵ و ۰٫۸۷۵۲۱۶ خواهد بود که بیانگر کارآیی بالاتر الگوریتم‌های آشوب‌گونه در مقایسه با الگوریتم‌های دیگر است. همچنین با توجه به نتایج رتبه‌بندی بر اساس ارتباطات معنادار الگوریتم‌ها با یکدیگر الگوریتم‌های Chaotic ICA با تمام الگوریتم‌های موجود ارتباط معنادار دارد و پس از آن الگوریتم Chaotic GA با تعداد زیادی از الگوریتم‌های موجود ارتباط معنادار دارد. نتایج ذکرشده نشان‌دهندهٔ عدم‌تصادفی‌بودن روش‌های پیشنهادی است.

جدول (٤): نتایج ارزیابی داده‌های Breast Cancer با معیارهای NMI، FM و AR.

الگوریتم‌های تکاملی	معیار NMI	معیار FM	معیار AR	میانگین
Chaotic PSO	0.765217	0.960358	0.859962	0.861846
Chaotic ICA	0.790138	0.966176	0.879651	0.878655
Chaotic DE	0.750243	0.956735	0.847722	0.851567
Chaotic GA	0.785005	0.964976	0.875667	0.875216
Chaotic ABC	0.731376	0.954300	0.839575	0.841750
PSO	0.756517	0.958201	0.852685	0.855801
ICA	0.777578	0.963084	0.869173	0.869945
SA	0.412089	0.851566	0.429269	0.564308
DE	0.754986	0.957753	0.851088	0.854609
GA	0.776086	0.962916	0.868648	0.869217
ABC	0.766346	0.960636	0.860927	0.862636
HS	0.578951	0.899503	0.655523	0.711326
K-means	0.717068	0.948598	0.820352	0.828673

مجموعهٔ دادهٔ Iris در واقع مجموعه‌ای از داده‌ها است که شامل سه نمونه گل زنبق است. این مجموعه شامل ۵۰ نمونه است که هر نمونه دارای ۴ ویژگی است و در ۳ کلاس متمایز طبقه‌بندی می‌گردد. جدول ۵ نتایج میانگین ۱۰۰ بار اجرای الگوریتم‌های تکاملی موجود بر روی مجموعهٔ دادهٔ Iris را نشان می‌دهد.

جدول (۵): نتایج میانگین ۱۰۰ بار اجرای الگوریتم‌های تکاملی موجود بر روی مجموعهٔ دادهٔ Iris

الگوریتم‌های تکاملی	تابع هزینه			انحراف معیار	زمان اجرا
	حداقل هزینه	متوسط هزینه	حداکثر هزینه		
Chaotic PSO	97.3298	98.7982	99.5574	± (0. 49263)	1.7454s
Chaotic ICA	96.6555	96.6629	96.9222	±(0.02699)	1.5330s
Chaotic DE	96.8732	97.6374	99.0106	±(0.58680)	1.5875s
Chaotic GA	96.6664	96.6801	96.6970	±(0.00628)	1.7363s
Chaotic ABC	96.7286	97.3214	101.5813	±(0.845477)	2.9370s
PSO	97.8824	98.9718	100.3142	±(0.52493)	1.8018s
ICA	96.6554	96.7205	97.4632	±(0.18222)	1.7502s
SA	130.7615	181.1466	280.4573	±(25.10238)	1.0737s
DE	149.3569	106.4994	309.2958	±(32.22541)	1.7631s
GA	96.6661	97.3955	98.3262	±(0.38596)	3.9793s
ABC	97.8159	100.0714	103.8752	±(1.46867)	5.9294s
HS	107.2247	136.8219	148.4858	±(7.54184)	3.4983s
K-means	97.3259	101.5442	123.9695	±(9.71673)	0.7682s

جدول (٦): نتایج ارزیابی داده‌های Iris با معیارهای NMI، FM و AR.

الگوریتم‌های تکاملی	معیار NMI	معیار FM	معیارAR	میانگین
Chaotic PSO	0.767482	0.908629	0.753492	0.809868
Chaotic ICA	0.761604	0.903772	0.739832	0.801736
Chaotic DE	0.760922	0.903494	0.740055	0.801490
Chaotic GA	0.753398	0.899840	0.732701	0.795313
Chaotic ABC	0.750472	0.902291	0.736773	0.796512
PSO	0.765028	0.906767	0.748582	0.806792
ICA	0.765369	0.906324	0.745936	0.805876
SA	0.679596	0.888349	0.564961	0.710969
DE	0.769606	0.907483	0.748006	0.808365
GA	0.781664	0.918301	0.777250	0.825738
ABC	0.769900	0.908755	0.752736	0.810464
HS	0.713079	0.879649	0.643119	0.745283
K-means	0.730627	0.875371	0.681335	0.762444

مجموعهٔ دادهٔ Ionosphere مربوط به اتمسفر است، این داده‌های راداری توسط یک سیستم در لابراتوار گوس‌بی[1] جمع‌آوری شده است. این مجموعهٔ داده دارای ۳۵۱ نمونه است که در ۲ کلاس مختلف طبقه‌بندی شده است. هر یک از نمونه‌های این مجموعهٔ داده شامل ۳۴ ویژگی پیوسته است.

جدول (۷): نتایج میانگین ۱۰۰ بار اجرای الگوریتم‌های تکاملی موجود بر روی مجموعهٔ دادهٔ Ionosphere

الگوریتم‌های تکاملی	تابع هزینه			انحراف معیار	زمان اجرا
	حداقل هزینه	متوسط هزینه	حداکثر هزینه		
Chaotic PSO	319.202	326.7905	346.983	± (7.25921)	2.0872s
Chaotic ICA	1530.967	1531.951	1533.331	±(0.55444)	1.9458s
Chaotic DE	1581.566	1611.289	1656.076	±(14.64452)	2.0105s
Chaotic GA	1531.832	1532.607	1533.745	±(0.42419)	2.1175s
Chaotic ABC	1536.158	1536.158	1536.158	±(9.14E-13)	3.5230s
PSO	1593.872	1657.443	1842.476	±(71.27108)	2.3357s
ICA	1731.838	1834.948	1919.147	± (32.92895)	2.0606s
SA	1992.604	2164.404	2419.121	±(90.13086)	1.2607s
DE	1856.828	1958.693	2063.174	±(41.20099)	2.1532s
GA	1609.814	1655.09	1762.222	±(29.52986)	4.3329s
ABC	1899.199	1951.96	2006.783	±(23.8655)	7.3836s
HS	2267.1851	2422.111	2507.927	±(49.96369)	7.4590s
K-means	2217.282	2274.583	2311.822	±(18.34671)	0.0142s

بعد از بررسی پاسخ الگوریتم‌های آشوب‌گونه بر روی مجموعهٔ دادهٔ Ionosphere بر اساس جدول ۷ این امر مسلم می‌گردد که روش ارائه‌گردیده قابل‌انطباق با سیستم‌های واقعی نیز خواهد بود. لذا می‌توان از آن در اکثر مسائل بهینه‌سازی استفاده نمود. به عبارتی دیگر این اطمینان وجود دارد که الگوریتم پیشنهادی، محدودیت‌های الگوریتم‌های گذشته را تا حد قابل‌قبولی پوشش خواهد داد.

1. goos bay

جدول (۸): نتایج ارزیابی داده‌های Ionosphere با معیارهای NMI، FM و AR.

الگوریتم‌های تکاملی	معیار NMI	معیار FM	معیار AR	میانگین
Chaotic PSO	0.312042	0.576561	0.150641	0.346415
Chaotic ICA	0.132102	0.707732	0.173175	0.337670
Chaotic DE	0.115313	0.694127	0.154845	0.321428
Chaotic GA	0.131504	0.707315	0.173329	0.337383
Chaotic ABC	0.127076	0.702429	0.167902	0.332469
PSO	0.101822	0.701124	0.136161	0.313036
ICA	0.032224	0.839445	0.014900	0.295523
SA	0.026599	0.849208	0.012832	0.296213
DE	0.034372	0.770898	0.023602	0.276291
GA	0.093799	0.675297	0.142661	0.303919
ABC	0.029051	0.853086	0.017010	0.299716
HS	0.063718	0.775883	0.058938	0.299514
K-means	0.002298	0.527865	0.000324	0.176829

مجموعهٔ دادهٔ Half-ring شامل ۴۰۰ نمونه است که در ۲ کلاس طبقه‌بندی می‌گردد. هر یک از نمونه‌های این مجموعهٔ داده دارای ۲ ویژگی است. در جدول ۹ کارآیی الگوریتم‌ها بر اساس تابع هزینه آورده شده‌اند که با توجه به نتایج این جدول، الگوریتم Chaotic ABC هزینهٔ کمتری در مقایسه با الگوریتم‌های دیگر دارد.

جدول (۹): نتایج میانگین ۱۰۰ بار اجرای الگوریتم‌های تکاملی موجود بر روی مجموعهٔ دادهٔ Half-ring.

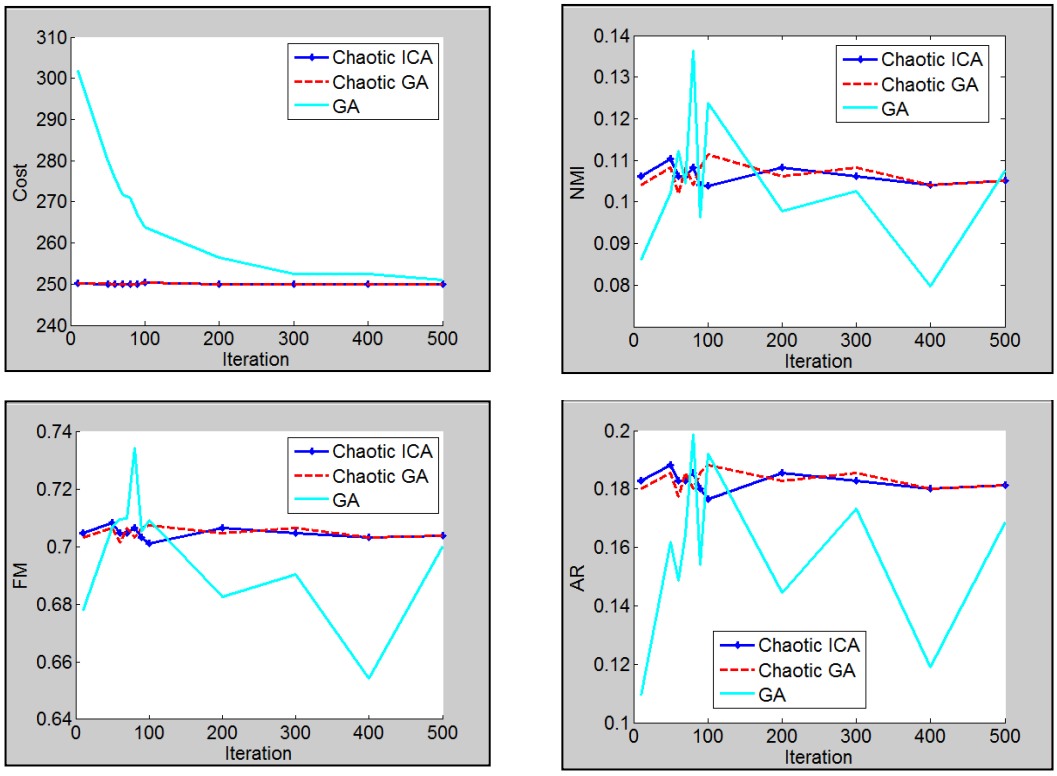
الگوریتم‌های تکاملی	تابع هزینه			انحراف معیار	زمان اجرا
	حداقل هزینه	متوسط هزینه	حداکثر هزینه		
Chaotic PSO	193.0345	193.0917	193.2442	± (0.04052)	1.8360s
Chaotic ICA	193.0277	193.0299	193.0368	±(0.00327)	1.6198s
Chaotic DE	193.0277	193.0454	193.1191	±(0.02313)	1.6667s
Chaotic GA	193.0277	193.036	193.0515	±(0.00412)	1.81264s
Chaotic ABC	193.0277	193.0288	193.0368	±(0.00125)	3.19743s
PSO	193.0402	193.0928	193.2037	±(0.03844)	1.9353s
ICA	193.0277	193.029	193.0373	± (0.00250)	1.7545s
SA	203.4527	286.5792	463.9958	±(54.01495)	1.0461s
DE	193.0277	193.0504	193.1486	±(0.02580)	1.8381s
GA	193.0277	193.031	193.071	±(0.00935)	3.6748s
ABC	193.0277	193.0277	193.0278	±(1.86E-05)	6.8335s
HS	193.0975	193.8955	195.8395	±(0.53657)	3.2754s
K-means	244.9351	268.1763	287.9756	±(7.49415)	0.0112s

الگوریتم‌های آشوب‌گونه با توجه به جدول ۱۰ کارآیی بالایی بر اساس معیارهای NMI، FM و AR بر روی مجموعهٔ دادهٔ Half-ring دارند.

جدول (۱۰): نتایج ارزیابی داده‌های Half-ring با معیارهای NMI، FM و AR.

الگوریتم‌های تکاملی	معیار NMI	معیار FM	معیار AR	میانگین
Chaotic PSO	0.293924	0.770180	0.193460	0.419188
Chaotic ICA	0.297281	0.772064	0.199610	0.422985
Chaotic DE	0.295886	0.771288	0.197055	0.421410
Chaotic GA	0.301586	0.774446	0.207491	0.427841
Chaotic ABC	0.295495	0.771075	0.196340	0.420970
PSO	0.294047	0.770234	0.193683	0.419321
ICA	0.296375	0.771563	0.197952	0.421963
SA	0.301275	0.806328	0.291851	0.466485
DE	0.295894	0.77129	0.197070	0.421418
GA	0.294636	0.770594	0.194767	0.419999
ABC	0.290568	0.771006	0.196110	0.419228
HS	0.289683	0.767231	0.183890	0.413602
K-means	0.002028	0.524445	0.000499	0.175657

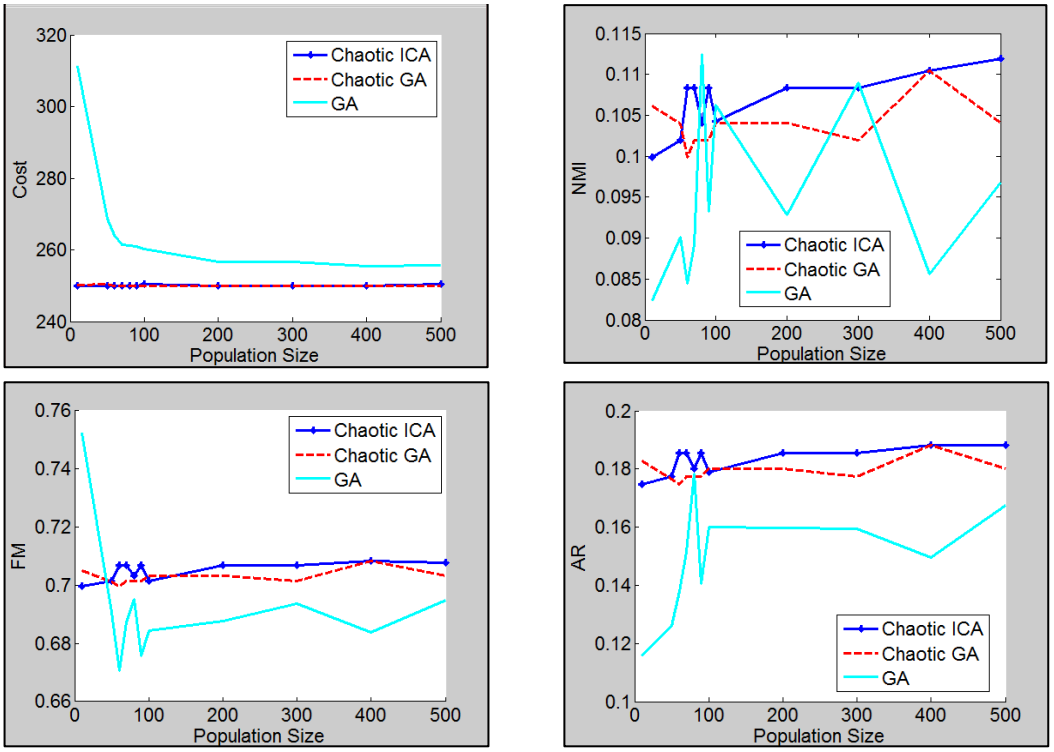
شکل ۶ دقت و کارآیی الگوریتم‌های Chaotic ICA، Chaotic GA و GA که بهترین پاسخ را در میان الگوریتم‌های موردبررسی داشته‌اند را نشان می‌دهند.



شکل ٦: مشخصهٔ همگرایی الگوریتم‌های Chaotic ICA، Chaotic GA و GA بر اساس تعداد تکرارها

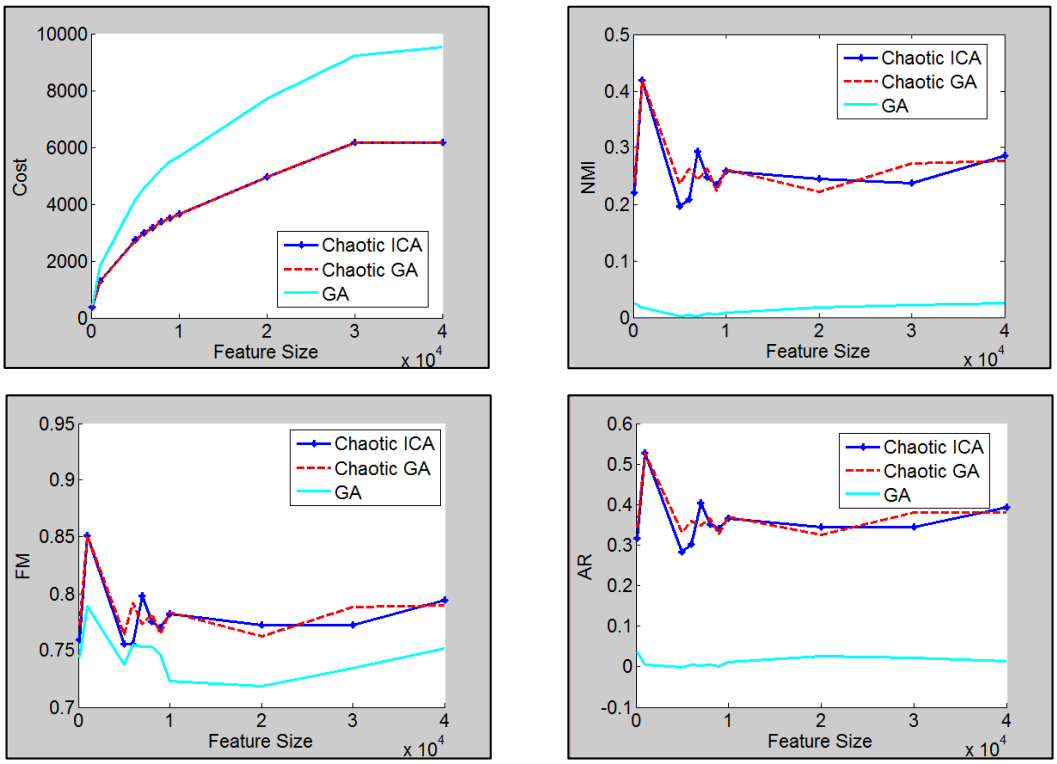
شکل ۶ همگرایی الگوریتم‌های Chaotic ICA، Chaotic GA و GA بر اساس تغییرات در تعداد تکرارها است. نمودارهای همگرایی الگوریتم‌ها با توجه به معیارهای تابع هزینه، NMI، FM و AR است. همان‌طور که مشاهده می‌گردد تغییرات در تکرارها تأثیری بر روی الگوریتم‌های آشوب‌گونه ندارد.

در شکل ۷ همگرایی الگوریتم‌های Chaotic ICA، Chaotic GA و GA نشان داده شده است. الگوریتم‌های آشوب‌گونه به تغییرات در اندازهٔ جمعیت نیز غیرحساس هستند.



شکل ۷: مشخصهٔ همگرایی الگوریتم‌های Chaotic ICA، Chaotic GA و GA بر اساس اندازهٔ جمعیت

در شکل ۸ مشخصه‌های همگرایی الگوریتم‌های Chaotic ICA، Chaotic GA و GA بر اساس تغییر در تعداد ویژگی‌های مجموعهٔ داده است. همان‌طور که مشاهده می‌گردد الگوریتم‌های آشوب‌گونه به تغییرات در اندازهٔ ویژگی‌های مجموعهٔ داده نیز حساس نیستند.



شکل ۸: مشخصهٔ همگرایی الگوریتم‌های Chaotic ICA، Chaotic GA و GA بر اساس تعداد ویژگی‌ها

نتیجه‌گیری

در این مقاله چندین روش برای بهبود کارآیی خوشه‌بندی داده‌ها ارائه گردید. این روش‌ها مبتنی بر الگوریتم‌های تکاملی آشوب‌گونه می‌باشند. نتایج تجربی بر روی مجموعهٔ داده‌های مختلف نشان می‌دهد که روش‌های پیشنهادی در مقایسه با حالات استاندارد خوشه‌بندی دارای کارآیی مناسب‌تری می‌باشند. شبیه‌سازی‌ها بر روی مجموعهٔ داده‌های متنوع اجرا گردیده است؛ لذا پاسخ‌های نهایی به‌دست‌آمده گویای این مطلب می‌باشد که وابستگی به تعداد نمونه، تعداد خوشه‌ها، تعداد پارامترها و تعداد تکرار وجود ندارد و الگوریتم‌های آشوب‌گونه عملکرد مناسبی دارند.

منابع

Alfonzetti, S., Dilettoso, E., & Salerno, N. (2006). Simulated annealing with restarts for the optimization of electromagnetic devices. *IEEE transactions on magnetics*, *42*(4), 1115-1118.

Atashpaz-Gargari, E., & Lucas, C. (2007). Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. 2007 IEEE congress on evolutionary computation,

Coley, D. A. (1999). *An introduction to genetic algorithms for scientists and engineers*. World Scientific Publishing Company.

Dorigo, M., & Gambardella, L. M. (1997). Ant colonies for the travelling salesman problem. *biosystems*, *43*(2), 73-81.

Fathian, M., Amiri, B., & Maroosi, A. (2007). Application of honey-bee mating optimization algorithm on clustering. *Applied Mathematics and Computation*, *190*(2), 1502-1513.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, MI.

Kao, Y.-T., Zahara, E., & Kao, I.-W. (2008). A hybridized approach to data clustering. *Expert Systems with Applications*, *34*(3), 1754-1762.

Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of global optimization*, *39*, 459-471.

Kellert, S. H. (1993). *In the wake of chaos: Unpredictable order in dynamical systems*. University of Chicago press.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings of ICNN'95-international conference on neural networks.

Krishna, K., & Murty, M. N. (1999). Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*(3), 433-439.

Lai, D. (1996). Comparison study of AR models of the Canadian lynx data: A close look at BDS statistic. *Computational statistics & data analysis*, *22*(4), 409-423.

Liz, E., & Ruiz-Herrera, A. (2012). Chaos in discrete structured population models. *SIAM Journal on Applied Dynamical Systems*, *11*(4), 1200-1214.

Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern recognition*, *33*(9), 1455-1465.

Neal, R. W. (1993). *The logistic Lattice in Randon Number Generation, Division of Methematics* University of Texas at San Antonio].

Niknam, T., & Amiri, B. (2010). An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. *Applied soft computing*, *10*(1), 183-197.

Niknam, T., Fard, E. T., Pourjafarian, N., & Rousta, A. (2011). An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering. *Engineering Applications of Artificial Intelligence*, *24*(2), 306-317.

Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics*, *17*(2-3), 97-119.

Schuster, H. G., & Just, W. (2006). *Deterministic chaos: an introduction*. John Wiley & Sons.

Shelokar, P., Jayaraman, V. K., & Kulkarni, B. D. (2004). An ant colony approach for clustering. *Analytica chimica acta*, *509*(2), 187-195.

استناد به این مقاله: منطقی، سجاد و خسروانی‌پور، سارا. (۱۴۰۲). بهبود کارآیی خوشه‌بندی داده‌ها با الگوریتم‌های تکاملی آشوب‌گونه. *فصلنامه پژوهش‌های نوین در شهر هوشمند*، ۱(۴)، ۶-۲۵.