فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Winter, 2021, 2 (4), 35-43
DOR: 20.1001.1.27832570.1400.2.4.6.3

Received: 17 Aug 2021
Accepted: 04 Oct 2021

A New Method for Data Clustering based on The Combination of Genetic Optimization and Firefly Algorithms

Mahsa Afsardeir[1], Mansureh Afsardeir[2*]

1. MSc, Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.
2. MSc, Department of Biomedical Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran.

(Corresponding Author) afsardeir2007@gmail.com

Abstract

Introduction: With the progress of technology and increasing the volume of data in databases, the demand for fast and accurate discovery and extraction of databases has increased. Clustering is one of the data mining approaches that is proposed to analyze and interpret data by exploring the structures using similarities or differences. One of the most widely used clustering methods is the k-means. In this algorithm, cluster centers are randomly selected and each object is assigned to a cluster that has maximum similarity to the center of that cluster. Therefore, this algorithm is not suitable for outlier data since this data easily changes centers and may produce undesirable results. Therefore, by using optimization methods to find the best cluster centers, the performance of this algorithm can be significantly improved. The idea of combining firefly and genetics algorithms to optimize clustering accuracy is an innovation that has not been used before.

Method: In order to optimize k-means clustering, in this paper, the combined method of genetic algorithm and firefly worm is introduced as the firefly genetic algorithm.

Findings: The proposed algorithm is evaluated using three well-known datasets, namely, Breast Cancer, Iris, and Glass. It is clear from the results that the proposed algorithm provides better results in all three datasets. The results confirm that the distance between clusters is much less than the compared approaches.

Discussion and Conclusion: The most important issue in clustering is to correctly determine the cluster centers**.** There are a variety of methods and algorithms that performs clustering with different performance. In this paper, based on firefly metaheuristic algorithms and genetic algorithms a new method has been proposed for data clustering. Our main focus in this study was on two determining factors, namely the distance within the data cluster (distance of each data to the center of the cluster) and the distance that the headers have from each other (maximum distance between the centers of the clusters). In the k-means algorithm, clustering is not accurate since the cluster centers are selected randomly. Employing firefly algorithms and genetics, we try to obtain more accurate centers of the clusters and, as a result, correct clustering.

Keywords: Data Mining, Genetic Algorithms, Firefly Algorithm, Firefly -Genetic Algorithm, K-Means Algorithm.

فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران

روشی نوین جهت خوشه‌بندی داده مبتنی‌بر ترکیب الگوریتم‌های بهینه‌سازی ژنتیک و کرم شب‌تاب

تاریخ دریافت: ۱۴۰۰/۰۵/۲۶
تاریخ پذیرش: ۱۴۰۰/۰۷/۱۲

سال دوم، زمستان ۱۴۰۰
شمارهٔ چهارم، صص: ۳۵ – ۴۳

مهسا افسردیر[۱] منصوره افسردیر[۲*]

۱.کارشناسی‌ارشد، گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، واحد علوم تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.
M.afsardeir2012@yahoo.com
۲.کارشناسی‌ارشد، گروه مهندسی پزشکی، دانشکده فنی مهندسی، واحد دزفول، دانشگاه آزاد اسلامی، دزفول، ایران. (نویسندهٔ مسئول)
afsardeir2007@gmail.com

چکیده: خوشه‌بندی یکی از مسائل مهم در داده‌کاوی است که بدون هدف ازپیش‌تعیین‌شده، داده‌ها را براساس شباهت درون خوشه‌ها تقسیم‌بندی‌می‌کند. از روش‌های متداول خوشه‌بندی الگوریتم k-means است که با دریافت ورودی، داده‌ها را به k خوشه تقسیم‌بندی‌می‌کند. یکی از معایب این روش حساسیت به شرایط اولیه است که منجر به کاهش دقت درخوشه‌بندی می‌شود. از روش‌های بهبود عملکرد k-means می‌توان استفاده از الگوریتم‌های فراابتکاری را نام‌برد. این پژوهش به دو روش بهینه‌سازی ژنتیک وکرم شب‌تاب پرداخته و الگوریتم جدیدی باعنوان الگوریتم ژنتیکی کرم‌شب‌تاب جهت بهینه‌سازی خوشه‌بندی k-means ارائه‌کرده‌است. الگوریتم کرم شب‌تاب از الگوریتم‌های هوش جمعی است که از ویژگی نور چشمک‌زن کرم شب‌تاب الهام‌گرفته‌است ؛ الگوریتم ژنتیک نیز نوعی الگوریتم فراابتکاری است که از تکنیک‌های زیست‌شناسی مانند وراثت و جهش استفاده‌می‌کند. در الگوریتم k-means چون مراکز خوشه‌ها به‌صورت تصادفی انتخاب‌می‌شوند، خوشه‌بندی دقت لازم را ندارد. لذا ما با استفاده از الگوریتم‌های فراابتکاری سعی در به‌دست آوردن مراکز دقیق خوشه‌ها و در نتیجه، دستیابی به خوشه‌بندی صحیح داریم. در روش پیشنهادی، ابتدا الگوریتم k-means را روی داده‌های ورودی اجراکرده و خوشه‌بندی انجام‌می‌شود. سپس مضربی از مراکز خوشه که در این الگوریتم به‌دست‌آمده را به‌عنوان حد پایین و حد بالای الگوریتم پیشنهادی استفاده‌می‌کنیم. جمعیت اولیه به‌صورت تصادفی بین حد پایین و حد بالا تولیدمی‌شود. در حلقه اصلی الگوریتم جمعیت را به دو دسته جمعیت مساوی تقسیم‌می‌کنیم، بر روی دسته اول الگوریتم ژنتیک را اجرامی‌کنیم، بر روی دسته دوم براساس الگوریتم کرم شب‌تاب، موقعیت‌های جدید را به‌دست-می‌آوریم. حال جمعیت قبلی و جمعیت جدید به‌دست‌امده از الگوریتم ژنتیک و جمعیت جدید به‌دست‌امده ازالگوریتم کرم شب‌تاب را تلفیق‌کرده و آن‌ها را ازخوب به بد مرتب‌می‌کنیم و تعداد مورد نیاز از آن‌ها را انتخاب و به ابتدای حلقه می‌رویم. این فرایند را تا برقراری شرط توقف ادامه‌می‌دهیم. در پایان الگوریتم k-means، الگوریتم کرم شب‌تاب، الگوریتم ژنتیک و الگوریتم پیشنهادی بر روی سه مجموعه داده اعمال‌شده و نتایج مقایسه‌شد. نتایج شبیه‌سازی نشان‌می‌دهدکه الگوریتم ژنتیکی کرم شب‌تاب عملکرد بهتری در مقایسه با سایر روش‌ها داشته‌است.

واژه‌های کلیدی: داده‌کاوی، الگوریتم ژنتیک، الگوریتم کرم شب‌تاب، الگوریتم ژنتیکی کرم شب‌تاب، خوشه‌بندی k-means.

۱. مقدمه

با پیشرفت فناوری و افزایش حجم داده در پایگاه داده‌ها، ضرورت کشف و استخراج سریع و دقیق از پایگاه داده‌ها بیش از پیش نمایان‌شده‌است. بنابراین نیاز به طراحی سیستم‌هایی که قادر به کشف دانش و اطلاعات مورد نیاز انسان باشند، به‌خوبی احساس‌می شود. داده‌کاوی مهم‌ترین فناوری برای بهره‌وری مؤثر، صحیح و سریع ازداده‌های حجیم است و اهمیت آن رو به فزونی است[۱]. داده‌کاوی پل ارتباطی میان علم آمار، علم کامپیوتر، هوش مصنوعی، الگوشناسی، فراگیری ماشین داده می‌باشد. داده‌ها اغلب حجیم‌اند و به‌تنهایی قابل‌استفاده نیستند، اما اطلاعات نهفته در داده‌ها قابل‌استفاده است. بنابراین بهره‌گیری از فرآیند داده‌کاوی جهت شناسایی الگوها و مدل‌ها و نیز ارتباط عناصر مختلف در پایگاه داده جهت کشف دانش نهفته در داده‌ها و نهایتاً تبدیل داده به اطلاعات، روزبه‌روز ضروری‌تر می شود. خوشه‌بندی یکی از رویکردهای داده‌کاوی است که در جهت تحلیل و تفسیر داده‌ها مطرح‌می شود. درواقع خوشه‌بندی، به‌دنبال کشف ساختار در داده‌های جمع‌آوری‌شده می‌باشد[۲]. این روش با استفاده از شباهت‌ها یا تفاوت‌ها (مانند فاصله) موجود میان نقاط داده در یک مجموعه به کشف ساختار می‌پردازد. از آنجا که خوشه‌بندی یک روش بدون نظارت محسوب‌می‌شود، در زمینه‌های گوناگون ازجمله پردازش تصویر، مطالعات زمین‌لرزه، بازاریابی، زیست‌شناسی، داده-کاوی از اهمیت ویژه‌ای برخوردار است. یکی از روش‌های پرکاربرد خوشه‌بندی روش k-means است. هدف اصلی k-means افزایش شباهت داخلی خوشه‌ها و کاهش شباهت اشیا بیرون خوشه‌ها است. در این الگوریتم مراکزخوشه‌ها به‌صورت تصادفی انتخاب‌می‌شوند و هر شی را با توجه به بیشترین شباهت آن به مراکز خوشه‌ها، به خوشه‌ها تخصیص‌می‌دهند. ازاین‌رو، این الگوریتم برای داده‌های پرت مناسب نیست زیرا این داده‌ها به‌راحتی مراکز را تغییرمی‌دهند و ممکن است نتایج نامطلوبی حاصل شود. بنابراین با استفاده از روش‌های بهینه سازی جهت یافتن بهترین مراکز خوشه‌ها می‌توان تا حد بسیار زیادی عملکرد این الگوریتم را بهبودبخشید. در سال‌های اخیر الگوریتم‌های هوشمند بسیاری در زمینه بهینه سازی خوشه‌بندی صورت‌گرفته‌است. هدف اصلی در این گونه تحقیقات، انتخاب صحیح مرکز خوشه‌ها برای جداکردن نمونه‌ها از یکدیگر و قراردادن نمونه‌های مشابه در یک خوشه و در نتیجه گروه‌بندی صحیح می‌باشد. بنابراین با توجه به اهمیت روش‌های خوشه‌بندی، با ارائه الگوریتم‌های فرابتکاری و ترکیب این الگوریتم‌ها با یکدیگر می‌توان مسائل خوشه‌بندی را نسبت به حل آن‌ها به‌تنهایی با جواب بهینه‌تری تولیدکرد[۵] [۴] [۳][۱۱]. محاسبه راه‌حل‌های بهینه برای اکثر مسائل بهینه‌سازی که در بسیاری از زمینه‌های کاربردی و عملی مشاهده‌می‌شوند، کار دشواری است. روش‌های حل مسائل بهینه‌سازی مشتمل بردو دسته روش‌های دقیق و روش‌های ابتکاری می‌باشد. روش‌های دقیق راه‌حل‌های بهینه را به‌دست‌آورده و شرایط بهینگی را تضمین‌می‌کنند. روش‌های فرابتکاری راه‌حل‌های با کیفیت بالا را در زمان معقولی، تولیدمی‌کنند، اما تضمینی برای یافتن راه‌حل بهینه سراسری ندارند.

از جمله الگوریتم‌های خوشه‌بندی می‌توان به الگوریتم ژنتیک و الگوریتم کرم شب‌تاب که امروزه به‌طور گسترده در مسائل مختلف استفاده‌می‌شوند، اشاره‌کرد. با ارائه الگوریتم‌های فرابتکاری و ترکیب این الگوریتم‌ها با یکدیگر می‌توان مسائل خوشه‌بندی را نسبت به حل آن‌ها به‌تنهایی با جواب بهینه‌تری تولیدکرد[۱۱].

در کوششی الگوریتم کرم شب‌تاب بهبودیافته را با الگوریتم K-means ترکیب‌کرده‌اند تا بتواند با این الگوریتم جدید، مراکز صحیح و با دقت بالا را برای خوشه‌ها به‌صورت بهینه پیداکند.[۱۶] همچنین الگوریتم کرم شب‌تاب را برای خوشه‌بندی به‌منظور کاهش مشکلاتی که در خوشه‌بندی وجوددارد به‌کاربرده است و در نهایت با الگوریتم زنبور عسل و الگوریتم بهینه‌سازی ازدحام ذرات مقایسه‌کرده‌است.[۱۷] ازجمله الگوریتم های دیگری که برای حل مشکلات خوشه‌بندی ارائه‌شده، الگوریتم K-FA است که از الگوریتم کرم شب‌تاب برای یافتن مراکز خوشه بهینه استفاده‌می‌کند سپس الگوریتم k-means از این مراکز برای اصلاح و مقداردهی مراکز خوشه استفاده‌می‌کند و موجب بهره‌وری بیشتر می‌شود.[۱۸]

در[۱۹] از ترکیب الگوریتم k-means و الگوریتم ژنتیک برای بهبود کیفیت خوشه‌بندی، ثبات در خوشه‌بندی و امکان جستجوی سراسری بهتر استفاده‌شده‌است.

بنابراین ایده ترکیب الگوریتم کرم شب تاب و ژنتیک به‌منظور بهینه‌سازی دقت خوشه‌بندی یک ابتکار و نوآوری است که تاکنون استفاده نشده است. بنابراین هدف از این مقاله ترکیب دو الگوریتم کرم شب‌تاب و الگوریتم ژنتیک جهت افزایش کارایی خوشه‌بندی k-means است. در ادامه، به مبانی k-means و روش‌های بهینه‌سازی ژنتیک، کرم شب‌تاب و شرح روش پیشنهادی این مقاله و نحوه پیاده سازی آن پرداخته‌شده‌است. سپس، نتایج خوشه‌بندی با الگوریتم پیشنهادی را بر روی داده‌های استاندارد مشابه مقایسه‌کرده و در نهایت به نتیجه‌گیری پرداخته‌ایم.

۲. روش پیشنهادی

هدف از الگوریتم k-means بخش‌بندی داده درون k خوشه به‌روشی است که شباهت درون‌خوشه‌ای کم و فاصله بین‌خوشه‌ای بالا باشد. اما به‌دلیل حساس‌بودن به شرایط اولیه دقت خوشه‌بندی را کاهش‌می‌دهد و نمی‌تواند جواب بهینه‌ای ایجادکند. .در این الگوریتم مراکز خوشه ها به‌صورت تصادفی انتخاب‌می‌شوند و هر شی را با توجه به بیشترین شباهت آن به مراکز خوشه‌ها، به خوشه‌ها تخصیص‌می‌دهند. از این رو، این الگوریتم برای داده‌های پرت مناسب نیست زیرا این داده‌ها به‌راحتی مراکز را تغییرمی‌دهند و ممکن است نتایج نامطلوبی حاصل‌شود.

در سال های اخیر الگوریتم های هوشمند بسیاری در زمینه بهینه سازی خوشه‌بندی صورت‌گرفته‌است. هدف اصلی در این گونه تحقیقات، انتخاب صحیح مرکز خوشه‌ها به منظور جداکردن نمونه‌ها از

یکدیگر و قراردادن نمونه های مشـــابه در یک خوشـــه و در نتیجه گروه‌بندی صحیح می‌باشد.

از آن‌جا که روش‌های موجود نمی‌توانند مرکز خوشـــه را با دقت بالایی محاسبه‌کنند بنابراین به خوشه‌بندی دقیق‌تر با استفاده از ترکیب روش‌های بهینه‌ســـازی نیازداریم. ازاین‌رو، می‌توان از یک خوشـــه‌بندی دقیق‌تر برای بهبود شـــناســـایی اشـــیا در تصـــویر و لبه‌یابی دقیق‌تر استفاده‌کرد.

در این مقاله با ترکیب دو الگوریتم ژنتیک و کرم شب‌تاب سعی در بهبود عملکرد الگوریتم k-means داشـــته‌ایم. با توجه به این که هر یک از الگوریتم‌های فراابتکاری دارای نواقص و نقاط قوتی هســـتند با ترکیب این الگوریتم‌ها تحت‌عنوان رو شی به نام الگوریتم ژنتیکی کرم شب‌تاب می‌توان کارایی را افزایش‌داد. به همین منظور، در این تحقیق از ترکیب مزایای الگوریتم ژنتیک و کرم شب‌تاب برای بهبود عملکرد الگوریتم k-means اســـتفاده شـــده اســـت. شـــکل ۱ فرآیند روش پیشـــنهادی را نمـــایش‌می‌دهـــد. در ادامـــه الگوریتم پیشـــنهـــادی بـــه تفصـــیـــل شرح‌داده‌خواهدشد.

۱،۲. خوشه‌بندی k-means

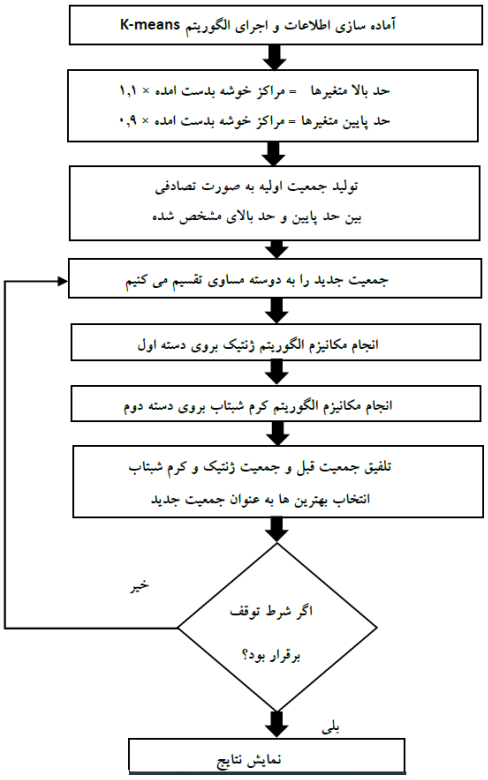
الگوریتم خو شه‌بندی k-means یکی از تق سیم بندی‌های ا سا سی در تجزیه و تحلیل خو شه‌ای ا ست و به‌دلیل سادگی، پیاده سازی آ سان و همگرایی سریع، کاربرد مؤثر و پیچیدگی زمانی خطی به‌طور گسترده‌ای استفاده‌می‌شود [۶]. این الگوریتم ابتدا k نقاط داده را به‌صورت تصادفی بـــه‌عنوان مراکز ثقـــل اولیـــه درنظرمی‌گیرد و ســـپس هر داده را بـــه نزدیکترین خوشـــه اختصـــاص‌داده تا ز مانی که مع یار همگرایی صورت‌پذیرد[۷]. مراحل این الگوریتم به شرح زیر است:

۱)به صورت تصادفی k شی دلخواه را از مجموعه داده D به عنوان مرکز خوشه اولیه انتخاب‌می‌کند.

۲)هر شی را با توجه به بیشترین شباهت آن به مراکز خوشه‌ها، به خوشه‌ها اختصاص‌می‌دهد.

۳)به‌روزرســـانی میانگین خوشـــه‌ای جدی ، مثلاً محاســـبه مقدار میانگین اشیاء برای هر خوشه.

۴)با توجه به مراکز جدید خوشـــه‌ها تا زمانی که هیچ تغییری رخ‌ندهد به گام دوم برگرد.



شکل ۱ : فرآیند روش پیشنهادی

۲،۲. الگوریتم بهینه‌سازی کرم شب‌تاب

الگوریتم کرم شب‌تاب الگوریتمی برگرفته از طبیعت است که رفتار اجتماعی کرم‌های شب‌تاب را شبیه‌سازی‌می‌کند. این الگوریتم توسط یانگ درسال ۲۰۰۹ معرفی‌گردید [۸]. کرم‌های شب‌تاب نورهایی تولید-می‌کنند که الگوی نوری هرکدام با دیگری متفاوت است. آن‌ها به‌منظور جذب جفت و شکار از این نور استفاده‌می‌کنند، میزان این نور رابطه مستقیم با جاذبیت کرم شب‌تاب دارد با درنظرگرفتن میزان نور هر کرم به‌عنوان مقدار تابع هدف، می‌توان رفتار کرم‌های شب‌تاب را به‌صورت یک الگوریتم بهینه‌ساز مدل نمود[۸] [۱۳]. الگوریتم کرم شب‌تاب بر پایه جذب و روشنایی است. این موضوع موجب تقسیم‌بندی خودکار کل جمعیت در زیر گروه‌ها با یک متوسط فاصله معین می‌شود و هر گروه می‌تواند در اطراف یک بهینه محلی ازدحام‌کند. در میان همه این‌ها ، بهترین بهینه سراسری می‌تواند یافت‌شود. دوم، این تقسیم‌بندی اجازه یافتن همه بهینه‌های زمان را می‌دهد. این قابلیت الگوریتم کرم شب‌تاب را مخصوصاً برای مسائل بهینه‌سازی چندکیفیتی غیرخطی مناسب می-سازد. به‌علاوه پارامترها در الگوریتم کرم شب‌تاب می‌تواند برای کنترل تصادفی به نسبت تکرار تنظیم‌شود، به‌گونه‌ای که همگرایی نیز می‌تواند با میزان‌سازی این پارامترها سرعت یابد. این مزایا الگوریتم کرم شب‌تاب را برای قابلیت سروکارداشتن با مسائل پیوستگی، خوشه‌بندی، کلاس-بندی و به‌علاوه بهینه‌ساز ترکیبی مناسب می‌سازد[۱۱].

میزان جذب کرم‌های شب‌تاب (β) نسبی بوده و به فاصله بین دو کرم شب‌تاب (r) و ضریب جذب نور (γ) بستگی دارد که از رابطه (۱) قابل-محاسبه است.

$$\beta(r) = \beta_0 e - e^{-\gamma r^2} \qquad (۱)$$

در رابطه (۱)، β_0 نشان‌دهنده میزان جذب در منبع نور می‌باشد. فاصله بین هر دو کرم شب‌تاب i و j در نقطه x_i و x_j با رابطه(۲) محاسبه‌می-شود.

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2} \qquad (۲)$$

در رابطه (۲)، $x_{i,k}$نشان‌دهنده kامین بخش از مختصات فضایی x_iکرم شب‌تاب i می‌باشد. همچنین حرکت کرم شب‌تاب و جذب i به کرم شب‌تاب j که درخشان‌تر است از رابطه (۳) محاسبه‌می‌شود.

$$x_n^i = x_n^{i-1} + \beta_0 e^{-\gamma r i j^2}(x_{nj}^{i-1} - x_{ni}^{i-1}) + \alpha e_i^{n-1} \qquad (۳)$$

در رابطه (۳)، x_i موقعیت کرم شب‌تاب کم‌نورتر، x_j موقعیت کرم درخشان‌تر، n شماره تکرار، α، عددی تصادفی و e_i^{n-1} ، یک بردار از اعداد تصادفی است که می‌تواند دارای توزیع یکنواخت یا گوسی باشد.

۳،۲. الگوریتم ژنتیک

از الگوریتم ژنتیک در م سائل ج ستجو و بهینه سازی ا ستفاده‌می شود. ابتدا یک نسل اولیه (به‌صورت تصادفی) ایجادمی‌گردد، که در واقع کروموزوم‌های اولیه ه ستند [۹]. هریک از این کروموزوم‌ها جوابی برای م سئله ه ستند اما، نه جواب ا صلی که ما به‌دنبال آن ه ستیم. سپس پد یده جهش با احتمال خیلی کم ممکن است رخ‌دهد. در نهایت کروموزوم‌ها از نظر امتیاز رتبه‌بندی‌می‌شوند، این امتیازدهی معمولاً بر اساس مقدار تابع هدف است. برخی کروموزوم‌ها با هم ترکیب‌شده و نسل بعد را به‌وجودمی‌آورند، احتمال انتخاب کروموزوم‌های با امتیاز بالاتر، بیشتر است اما درعین‌حال، احتمال انتخاب‌شدن برای تمام کروموزوم‌ها حتی کروموزوم‌های با کمترین امتیاز وجوددارد. با نسل جد ید به‌وجودآ مده این مرا حل را تکرارمی‌شود تا به جواب مطلق برسیم[۱۰][۱۲][۱۳].

از نظر الگوریتمی، مراحل اصلی GA به‌شرح زیر است:

۱) به‌طور تصادفی یک گروه اولیه از راه‌حل‌ها ایجادمی‌شود.

۲) ارزش تناسب هر راه‌حل، در گروه، ارزیابیمی‌شود.

۳) با تکرار و اعمال مراحل زیرگروه جدیدی از راه‌حل‌ها را ایجادمی‌شود:

- دو والد از گروه راه‌حل‌ها بر اساس مکانیسم انتخاب، انتخاب‌می‌شوند.
- عمل تقاطع و جهش انجام‌می‌شود.
- کروموزوم‌های شایسته‌تر جایگزین‌می‌شوند.

۴) درصورتی‌که شرط پایان برقرارشده‌با شد، پایان‌می‌یابد. در غیر این‌صورت به مرحله قبل برمی‌گردد.[۱۳][۱۴][۱۵]

در الگوریتم های ژنتیکی, در طی مرحله تولیدمثل ازعملگرهای ژنتیکی ا ستفاده‌می شود. با تأثیر این عملگرها بر روی یک جمعیت, نسل بعدی آن جمعیت تولیدمی‌شود. این عملگرها شامل:

عملگر تقاطع: بر روی یک زوج کروموزوم از نسل مولد عمل‌کرده و یک زوج کروموزوم جدید تولیدمی کند. به عبارت دیگر، عملگر تقاطع در یک لحظه بر روی دو کروموزوم اعمال‌شده و دو نوزاد به‌وسیله ترکیب ساختار دو کروموزوم ایجادمی‌کند. روش‌های تقاطع مختلفی وجوددارد: تقاطع تک‌نقطه‌ای(نقطه‌ای به صورت ت صادفی به‌عنوان نقطه برش در دو کروموزوم والد انتخاب‌شده و از آن نقطه به بعد با هم جابجا می‌شوند)، تقاطع دو نقطه‌ای (به‌طور تصادفی دو نقطه از کروموزوم را انت خاب‌کرده و ت مام ژن های بین این دو نق طه را در دو کروموزوم تعویض‌می‌کنیم) و ... .

عملگر جهش :این عملگر در کروموزوم‌های متفاوت تغییرات تصادفی برنامه‌ریزی نشده‌ای ایجادکرده و ژن‌هایی را که در جمعیت اولیه وجود ندا شته‌اند را وارد جمعیت می‌کند. نقش جهش در الگوریتم ژنتیک بازگرداندن مواد ژنتیکی گم شده و یا پیدانشده داخل جمعیت است. تا از همگرایی زودرس الگوریتم به جواب‌های بهینه محلی جلوگیری‌شود. عملگر های جهش مختلفی وجوددارد، از میان آن ها جهش یکنواخت را به‌صورت مختصر توضیح خواهیم‌داد. در این عملگر، ژنی از کروموزوم به‌طور ت صادفی انتخاب شده‌ا ست و مقدار آن به مقدار تصادفی دیگری تبدیل‌می‌شود. ابتدا یک عدد تصادفی در بازه،[L ۱] که طول کروموزوم موردنظر ا ست، انتخاب شده و ژن موجود در آن مکان از کروموزوم تغییرمی‌کند.

عملگر انتخاب: برای انتخاب بهترین جواب‌ها در تولید مجدد نسل (تولید جمعیت جدید) باید از روشی استفاده‌کرد که تا حد ممکن بهترین جواب را انت خاب‌کند. عملگر انت خاب چرخ رو لت یک روش انتخاب ا ست که در آن عنصری که عدد برازش(تنا سب) بیشتری دا شته با شد، انتخاب‌می‌شود. در واقع به نسبت عدد برازش برای هر عنصر یک احتمال تجمعی نسبت می‌دهیم با این احتمال که شانس انتخاب هر عنصر تعیین‌شود.

۴،۲. الگوریتم ژنتیکی کرم شب‌تاب

هدف از الگوریتم k-means بخش‌بندی داده درون k خوشه به‌طوری که شباهت درون‌خو شه ای بالا، فا صله بین‌خو شه ای بالا با شد اما به‌دلیل حساس‌بودن به شرایط اولیه دقت خوشه‌بندی را کاهش‌داده و جواب بهینه ای به ما نمی‌دهد .جهت بهینه‌سازی خوشه‌بندی k-means در

این مقاله روش ترکیبی الگوریتم ژنتیک و کرم شب تاب با عنوان الگوریتم ژنتیکی کرم شب‌تاب معرفی‌شده‌است. با توجه به این که هریک از الگوریتم‌های فراابتکاری دارای نواقص و نقاط قوتی هستند با ترکیب این الگوریتم‌ها می‌توان کارایی را افزایش‌داد. در الگوریتم پیشنهادی ابتدا اطلاعات موردنظر با استفاده از روش k-means خوشه‌بندی‌می شوند. سپس مضربی از مراکز خوشه که در این روش به‌دست‌آمده‌است را به‌عنوان حد پایین و حد بالای الگوریتم پیشنهادی استفاده‌می‌کنیم. که به‌ترتیب با رابطه (۴) و رابطه (۵) محاسبه‌می‌شود.

(۴) مراکز خوشه به‌دست آمده × ۱/۱= حد بالای خوشه‌ها

(۵) مراکز خوشه به‌دست آمده × ۰/۹ = حد پایین خوشه‌ها

پس از این مرحله جمعیت اولیه به‌صورت تصادفی بین حد پایین و حد بالا تولیدمی‌شود. در حلقه اصلی الگوریتم جمعیت را به دو دسته جمعیت مساوی تقسیم‌می‌کنیم. بر روی دسته اول الگوریتم ژنتیک را اجرامی‌کنیم:

- با احتمال p_c تقاطع تک‌نقطه‌ای را اجرامی‌کنیم.

عملگر تقاطع در یک لحظه بر روی دو کروموزوم اعمال‌شده و دو نوزاد به‌وسیله ترکیب ساختار دو کروموزوم ایجادمی‌کند. مفهوم مهمی که در ارتباط با این عملگر مطرح است نرخ تقاطع p_c است. . اگر تعداد کروموزوم‌های تولیدشده را با x و اندازه جمعیت اولیه را با $pop-size$ نشان‌دهیم، خواهیم‌داشت:

$$p_c = \frac{x}{pop-size} \qquad (۶)$$

با احتمال p_m جهش یکنواخت را اجرامی‌کنیم.

این عملگر در کروموزوم های متفاوت تغییرات تصادفی بر نامه‌ریزی نشده‌ای ایجادکرده و ژن‌هایی را که در جمعیت اولیه وجودنداشته‌اند، وارد جمعیت می‌کند. درباره این عملگر مفهوم مهمی مطرح‌شده که به آن نرخ جهش p_mگویند.

با استفاده از عملگر انتخاب چرخ رولت برای ورود به جمعیت جدید انتخاب‌می‌شود.

برای انتخاب بهترین جواب ها و تولید مجدد نسل (تولید جمعیت جدید) باید از روشی استفاده‌کرد که تا حد ممکن بهترین جواب را انتخاب‌کند. در میان روش‌های مختلف انتخاب ما چرخ رولت پیشنهادی هالند است.

بر روی دسته دوم بر اساس الگوریتم کرم شب‌تاب موقعیت‌های جدید را به‌دست‌می‌آوریم.

میزان جاذبیت هر کرم شب‌تاب را با توجه به ضریب جذب معادله(۷) محاسبه می‌کنیم.

$$\beta(r) = \beta_0 e - e^{-\gamma r^2}$$

که $\beta 0$ ، γ و r به‌ترتیب میزان جذب ازپیش‌تعریف‌شده، ضریب جذب نور و فاصله بین دو کرم شب‌تاب i و j می‌باشند.

سپس حرکت هر کرم شب‌تاب به سمت کرم شب‌تاب جاذب‌تر با استفاده از فرمول (۸) محاسبه‌خواهدشد.

$$x_n^i = x_n^{i-1} + \beta_0 e^{-\gamma r i j^2}(x_{nj}^{i-1} - x_{ni}^{i-1}) + \alpha e_i^{n-1} \qquad (۸)$$

که x_i، موقعیت کرم شب تاب کم‌نورتر، x_j موقعیت کرم شب تاب درخشان‌تر، n، شماره تکرار α عددی تصادفی و e_i^{n-1}، یک بردار از اعداد تصادفی است که می‌تواند دارای توزیع یکنواخت یا گوسی باشد.

حال جمعیت قبلی و جمعیت جدید حاصل از الگوریتم ژنتیک و جمعیت جدید حاصل از الگوریتم کرم شب‌تاب را تلفیق‌کرده و آن‌ها را از خوب به بد مرتب‌می‌کنیم و به تعداد مورد نیاز از آن‌ها انتخاب و به ابتدای حلقه می‌رویم. این فرآیند را تا برقراری شرط توقف ادامه‌می‌دهیم که در این تحقیق حداکثر تعداد تکرار درنظرگرفته‌شده‌است.

۵،۲. ارزیابی الگوریتم پیشنهادی

در این پژوهش از فاصله درون‌خوشه‌ای و فاصله بین مراکز خوشه به عنوان تابع هزینه برای الگوریتم‌ها استفاده‌می‌شود و تابع هزینه نهایی با استفاده از جمع دو این الگوریتم با ضرایب اهمیت به‌ترتیب w1 و w2 محاسبه‌خواهدشد. فاصله درون‌خوشه‌ای با توجه به فرمول فاصله اقلیدسی به‌صورت رابطه (۹) محاسبه‌خواهدشد.

$$\text{فاصله درون خوشه‌ای} = \sum data_i - cluster_i \qquad (۹)$$

که در رابطه (۹) $data_i$ داده iام، $cluster_i$ مرکز خوشه iام و n تعداد جمعیت است. شباهت بین مراکز خوشه به‌صورت رابطه (۱۰) محاسبه‌می‌شود.

$$\text{فاصله برون خوشه‌ای} = \sum_{i=1}^{c}\sum_{j=1}^{c}(cluster_i - cluster_j) \qquad (۱۰)$$

که در این رابطه $cluster_i$ مرکز خوشه iام، $cluster_j$ مرکز خوشه jام و c تعداد مراکز خوشه است. ارزیابی نهایی با با تابع هدف نهایی نیز با استفاده از رابطه (11) و ضریب $w1$ و $w2$ به‌ترتیب برای فاصله درون‌خوشه‌ای و فاصله بین مراکز خوشه محاسبه‌خواهدشد.

$$\text{Cost function} = w1 \times (\text{فاصله درون خوشه‌ای}) + w2 \times (\text{فاصله برون خوشه‌ای}) \qquad (11)$$

۶،۲. پارامترهای روش پیشنهادی

برای یک مقایسه مناسب، باید شرایط حل برای همهٔ الگوریتم‌ها یکسان باشد ازاین‌رو، پارامترهای مربوط به الگوریتم ها در جدول (۲) نشان‌داده‌شده‌است. لازم به ذکر است که الگوریتم پیشنهادی تمام پارامترهای مربوط به الگوریتم‌های ژنتیک و کرم شب‌تاب را دارا خواهدبود و به‌دلیل

احتمالی بودن فضای جستجوی الگوریتم‌ها، هریک از الگوریتم‌ها را ۳۱ بار اجراکرده‌ایم و نتایج در ادامه آمده‌است.

جدول ۱: مقادیر اولیه پارامترهای الگوریتم‌های فراابتکاری

نام الگوریتم	نام پارامتر	اندازه پارامتر
k-means	تعداد تکرار	۱۰۰
	تعداد مراکز خوشه	۳
	اندازه جمعیت	۱۰
الگوریتم ژنتیک	تعداد مراکز خوشه	۳
	درصد جهش	۰/۱
	درصد تقاطع	۰/۹
	اندازه جمعیت	۱۰
الگوریتم کرم شب-تاب	تعداد مراکز خوشه	۳
	ضریب جذب نور	۰/۹
	میزان جذب	۲
	شدت نور	۱
	تعداد تکرار	۱۰۰
الگوریتم ژنتیکی کرم شب‌تاب	تعداد تکرار	۱۰۰
	تعداد مراکز خوشه	۳
	اندازه جمعیت	۱۰
	درصد جهش	۰/۱
	درصد تقاطع	۰/۹
	ضریب جذب نور	۰/۹
	میزان جذب	۲
	شدت نور	۱

۷،۲. پایگاه داده

برای ارزیابی مناسب الگوریتم‌های فراابتکاری از ۳ مجموعه داده شناخته-شده برای مقایسه استفاده‌شده‌است؛ نام و مشخصات این پایگاه داده‌ها در جدول(۲) آمده‌است.

پایگاه داده Brest Cancer : این مجموعه داده مربوط به سرطان سینه است که در سال ۱۹۹۱ از دانشگاه ویسکانسین توسط دکتر ویلیام .اچ. ولبرگ در بیمارستان مدیسون جمع‌آوری‌شده‌است. این مجموعه داده دارای ۶۴۴ نمونه و ۲ کلاس است که هرکدام از این نمونه ها دارای ۴ ویژگی هستند. ویژگی‌های مربوط به این مجموعه داده عبارتند از: ضخامت توده، یکنواختی اندازه سلول، یکنواختی شکل سلول، اندازه سلول اپیتیال، کروماتین مطلوب، هسته آشکار، هسته عادی، چسبندگی حاشیه‌ای و تقسیم غیرمستقیم هسته سلول.

جدول ۲: مقادیر پایگاه داده‌ها

نام پایگاه داده	تعداد نمونه‌ها	تعداد کلاس‌ها	تعداد ویژگی‌ها
مجموعه داده Breast Cancer	۶۴۴	۲	۹
مجموعه داده Iris	۱۵۱	۳	۹
مجموعه داده Glass	۲۱۹	۷	۴

پایگاه داده iris: این مجموعه داده مربوط به اطلاعات سه نمونه گل زنبق است، که در سال ۱۹۸۷ توسط فیشر تهیه‌شده‌است. این مجموعه داده دارای ۱۵۱ نمونه و ۳ کلاس است که هرکدام از این نمونه‌ها دارای ۹ ویژگی هستند. ویژگی‌های مربوط به این مجموعه داده عبارتند از: طول کاسبرگ، پهنای کاسبرگ، طول گلبرگ، پهنای گلبرگ.

پایگاه داده Glass: این مجموعه داده مربوط به اطلاعات شناسایی شیشه است، که در سال ۱۹۸۷ تهیه‌شده‌است. این مجموعه داده دارای ۲۱۹ نمونه و ۷ کلاس است که هرکدام از این نمونه‌ها دارای ۴ ویژگی هستند. ویژگی‌های مربوط به این مجموعه داده عبارتند از: سدیم، منیزیم، آلومینیوم، سیلیکون، پتاسیم، کلسیم، باریوم، آهن و شاخص انکساری.

۸،۲. نتایج و تجزیه و تحلیل داده‌ها

در این قسمت نتایج شبیه سازی مربوط به الگوریتم‌های ارائه شده در قسمت قبل را بررسی‌خواهیم‌کرد. الگوریتم‌های k-mean، ژنتیک، کرم شب‌تاب و الگوریتم ژنتیکی کرم شب‌تاب بر روی ۳ مجموعه داده اجراو نتایج مقایسه شده‌است. همان‌طور که در قسمت ۲-۵ اشاره شد تابع شایستگی مجموع وزن‌دار فاصله درون‌خوشه‌ای و فاصله مراکز خوشه‌ها با ضریب w1 و w2 است. حال برای تحلیل تابع شایستگی؛ هر سه حالت را به‌صورت کلی بررسی‌می‌کنیم. جدول (۳) تا (۵) به‌ترتیب نتایج را بر اساس فاصله بین‌خوشه‌ای و درون‌خوشه‌ای، فاصله درون‌خوشه ای و شباهت بین مراکز خوشه‌ای نمایش‌می‌دهد.

در ادامه الگوریتم‌های ژنتیک ،کرم شب‌تاب و ژنتیکی کرم شب‌تاب برای هریک از مجموعه داده‌های موردنظر ، اجرا شده و نتایج به صورت نمودار همگرایی برای ۱۰۰ بار تکرار در شکل (۲) تا (۵) نشان‌داده‌شده‌است.

جدول ۳: نتایج بر اساس فاصله بین‌خوشه‌ای و درون‌خوشه‌ای

پایگاه داده/ الگوریتم	k-means	الگوریتم ژنتیک	الگوریتم کرم شب‌تاب	الگوریتم ژنتیکی کرم شب‌تاب
مجموعه داده Breast Cancer	۳۲۳/۸۷	۳۲۲/۷۴	۳۲۲/۵۴	۳۲۱/۹۸
مجموعه داده Iris	۳۱/۲۸	۳۰/۹۹	۳۰/۸۵	۳۰/۴۱
مجموعه داده Glass	۶۸/۵۴	۶۸/۲۳	۶۸/۳۲	۶۷/۸۷

جدول۴: نتایج بر اساس شباهت بین مراکز خوشه ای

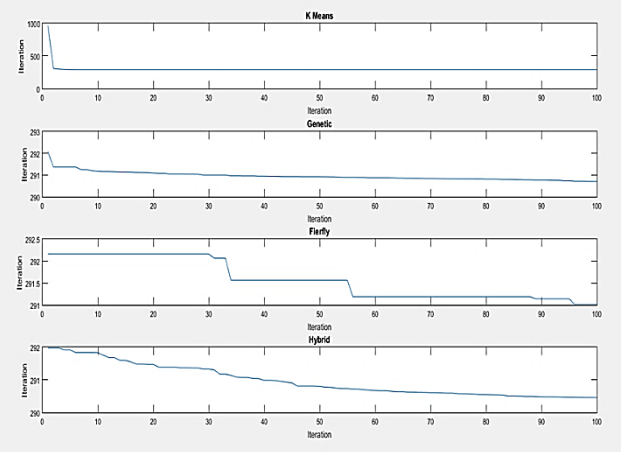
پایگاه داده/ الگوریتم	k-means	الگوریتم ژنتیک	الگوریتم کرم شب‌تاب	الگوریتم ژنتیکی کرم شب‌تاب
مجموعه داده Breast Cancer	۳۲۸/۶۵	۳۲۷/۶۵	۳۲۷/۶۳	۳۲۶/۹۹

مجموعه داده Iris	۲۹/۷۴	۲۹/۶۵	۲۹/۵۱	۲۸/۶۵
مجموعه داده Glass	۶۶/۵۴	۶۶/۲۳	۶۶/۷۸	۶۵/۹۱

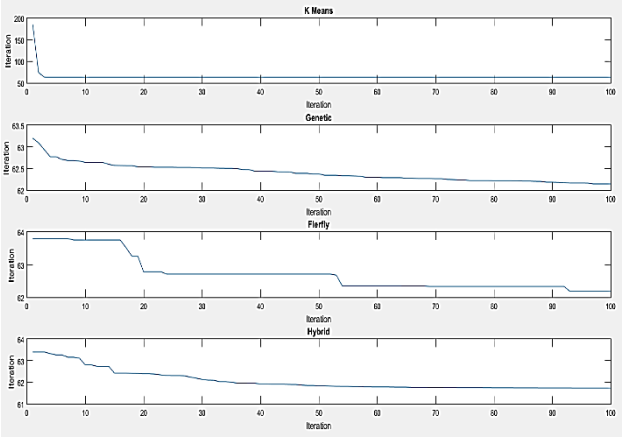
جدول ۵ : نتایج بر اساس فاصله بین‌خوشه‌ای و درون‌خوشه‌ای

پایگاه داده/ الگوریتم	k-means	الگوریتم ژنتیک	الگوریتم کرم شب‌تاب	الگوریتم ژنتیکی کرم شب‌تاب
مجموعه داده Breast Cancer	۴/۳۹	۴/۲۱	۳/۸۷	۳/۵۴
مجموعه داده Iris	۲/۴۳	۲/۳۱	۲/۳۵	۲/۰۹
مجموعه داده Glass	۲/۵۶	۲/۴۵	۲/۳۲	۲/۰۱

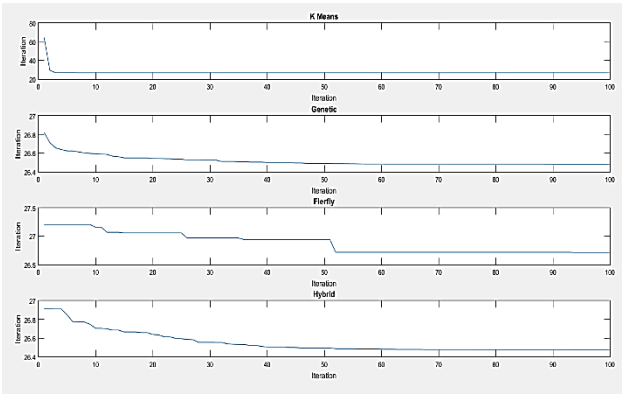
همان‌طور که از نتایج جدول (۳) تا (۵) مشخص است الگوریتم پیشنهادی نتایج بهتری در هر سه نوع تابع هدف به‌دست‌می‌دهد. همچنین از نتایج کاملاً پیداست که فا صله بین‌خوشه‌ای بسیار کمتر از دو نوع دیگر می‌باشد.



شکل۲ : نمودار همگرایی اجرای الگوریتم ها برای مجموعه داده BC



شکل۳ : نمودار همگرایی اجرای الگوریتم‌ها برای مجموعه داده Glass



شکل۴: نمودار همگرایی اجرای الگوریتم ها برای مجموعه داده Iris

۳. نتیجه‌گیری

در خوشه‌بندی مهمترین موضوع، تعیین صحیح مراکز خوشه، برای یک خوشه‌بندی صحیح است. روش‌ها و الگوریتم‌های بسیاری در این زمینه ارائه‌شده که هرکدام با عملکرد متفاوت، خوشه‌بندی را انجام‌می‌دهند. در این مقاله روش پیشنهادی جدیدی بر پایه دو الگوریتم فراابتکاری کرم شب‌تاب و الگوریتم ژنتیک برای خو شه‌بندی داده ارائه شده‌ا ست. بیشترین تمرکز ما در این پژوهش بر دو عامل تعیین‌کنندهٔ فاصله درون‌خو شه‌ای (فا صله هر داده تا مرکز خو شه) و میزان فا صله‌ای که سرخوشه‌ها از یکدیگر دارند (بیشترین فاصله مراکز خوشه‌ها) بوده‌است. در الگوریتمk-means چون مراکز خوشه به‌صورت تصادفی انتخاب‌می‌شود، خوشه‌بندی دقت لازم را ندارد. با استفاده از الگوریتم‌های فراابتکاری کرم شب‌تاب و ژنتیک سعی در به‌د ست‌آوردن مراکز دقیق خوشه‌ها و در نتیجه ، خوشه‌بندی صحیح بودیم. همان ونه که از نتایج شبیه‌سازی‌ها بر روی داده‌های استاندارد UCI پیداست، الگوریتم ژنتیک و ترکیب آن با الگوریتم کرم شب تاب، نسبت به الگوریتمk-means ، الگوریتم کرم شب‌تاب و الگوریتم ژنتیک، خو شه‌بندی را بهبود بخ شیده‌ا ست و دقت پا سخ ارائه شده در م سئله بالاتر از سایر روش‌ها ا ست. همچنین از نتایج م شخص ا ست که روش پی شنهادی فا صله درون‌خو شه‌ای را به صورت بهینه‌تری نسبت به دیگر روش‌ها انجام‌داده‌است. بررسی نتایج حاصل از تکرارهای بلندمدت و کوتاهمدت نشان‌می‌دهد که الگوریتم پیشنهادی به تعداد تکرارهای بهینه‌سازی بلند مدت وابستگی چندانی ندارد. در مجموع ترکیب حاصل از الگوریتم کرم شب تاب و الگوریتم ژنتیک، نسبت به الگوریتم‌های کرم شب‌تاب، ژنتیک و k-means نتایج بهتر و قابل‌قبول‌تری در خوشه‌بندی ارائه‌داده‌است.

مراجع

[1] Yaghini, M., and Ghazanfari, N. "Tabu-KM: a hybrid clustering algorithm based on tabu search approach." International Journal of Industrial Engineering & Production Research 21, no. 2 (2010)

[2] Jiawei, H., Kamber, M., Han, J., Kamber, M. and Pei, J. "Data Mining: Concepts and Techniques Elsevier." (2006).

[3] Taber, R. "Clustering (Xu, R. and Wunsch II, DC; 2009) [Book review]." IEEE Computational Intelligence Magazine 4, no. 3 (2009): 92-95.

[4] Berzal, F. and Matín, N. "Data mining: concepts and techniques by Jiawei Han and Micheline Kamber." ACM Sigmod Record 31, no. 2 (2002): 66-68.

[5] Jain, A. K., and Dubes, R. C. Jain, Anil K., and Richard C. Dubes. "Algorithms for clustering data." Prentice-Hall, Inc., 1988.

[6] Jain, A.K., "Data clustering: 50 years beyond K-means." Pattern recognition letters 31, no. 8 (2010): 651-666.

[7] Hassanzadeh, T. and Meybodi, M.R. "A new hybrid approach for data clustering using firefly algorithm and K-means." In The 16th CSI international symposium on artificial intelligence and signal processing (AISP 2012), pp. 007-011. IEEE, 2012.

[8] Yang, X. S. "Firefly algorithms for multimodal optimization." In International symposium on stochastic algorithms, pp. 169-178. Springer, Berlin, Heidelberg, 2009.

[9] Moscato, P. "On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms." Caltech concurrent computation program, C3P Report 826 (1989): 1989.

[10] Dianati, M., Song, I. and Treiber, M. An introduction to genetic algorithms and evolution strategies. Technical report, University of Waterloo, Ontario, N2L 3G1, Canada, 2002.

[11] Wahid, F., Ghazali, R. & Ismail, L.H. Improved Firefly Algorithm Based on Genetic Algorithm Operators for Energy Efficiency in Smart Buildings. Arab J Sci Eng 44, 4027–4047 (2019).

[12] Mahshwar, Keshva Kaushik, Vikram Arora. A Hybrid Data Clustering Using Firefly Algorithm Based Improved Genetic Algorithms. Sciencedirect. Procedia Computer Science 58 (2015) 249-256.

[13] M A El-Shorbagy, Adel M El-Refaey, A hybrid genetic–firefly algorithm for engineering design problems, Journal of Computational Design and Engineering, Volume 9, Issue 2, April 2022, Pages 706-730,

[14] Mustafa Servet Kiran, Ahmet Babalik. Improved Artificial Bee Colony Algorithm for Continuous Optimization Problems. Journal of Computer and Communications, 2014, 2, 108-116.

[15] Abdullah, A., Deris, S., Mohamad, M.S., Hashim, S.Z.M. (2012). A New Hybrid Firefly Algorithm for Complex and Nonlinear Problem. In: Omatu, S., De Paz Santana, J., González, S., Molina, J., Bernardos, A., Rodríguez, J. (eds) Distributed Computing and Artificial Intelligence. Advances in Intelligent and Soft Computing, vol 151. Springer, Berlin, Heidelberg.

[16] Hassanzadeh. t, meybodi.m, "A new hybrid Approach for Data clustering using firefly Algorithm and k-means"

[17] J.senthilnath, S.N.omkar, "clustering using firefly algorithm:performance study",swarm and Evolutionary Computation, volume1,ISSue3, pp 164-171,September 2011.

[18] Hassanzadeh,Tahereh, Meybodi ,Mohammad Reza, A New Hybrid Approach For Data Clustering Using Firefly Algorithm And K-Means, The 16th CSI International Symposium On Artificial Intelligence And Signal Processing (AISP), 2012,007-011.

[19] Binlu, Fangyuan Ju, An Optimized Genetic K-Means Clustering Algorithm, Computer Science and Information Processing (CSIP),2012,1296-1299.