"Research Article"

10.30495/QJOPM.2020.1867405.2443

A New Clustering Algorithm for Productivity in Data Mining: The Case of UCA Data

*Jhila Nasiri(Ph.D.)[1],Farzin Modarres Khiabani (Ph.D.)[*2],Nima Azarmir Shotorbani (Ph.D.)[3]*

(Receipt: 2020.02.19- Acceptance:2020.07.18)

Abstract

Methods of clustering in data mining have dramatically developed in recent years as a result of the crucial need to categorize data leading to the expansion of data mining techniques and enhanced productivity of clustering methods in management and decision making. Whale optimization algorithm is a new stochastic global optimization method employed to resolve various problems. We already presented a data clustering method based on Whale optimization algorithm in which the initial solutions are randomly selected. What has made K-mean algorithm a highly popular clustering approaches appealing to many researchers is the simplicity and brevity of the stages involved in the process. The present enquiry aimed at employing K-mean algorithm to improve the capability of Whale optimization clustering and proposing the hybrid KWOA algorithm which can find more accurate clusters. The computational results of running the newly proposed algorithm, along with some well-known clustering algorithms, on real data sets from a well-known machine learning repository underscored the promising performance of the proposed algorithm in terms of the quality and standard deviation of the final solutions.

KeyWords: Clustering, Data Mining, Productivity, Swarm intelligence

1Assistant Professor, Department of Mathematics, Tabriz Branch, Islamic Azad University, Tabriz, Iran

2-Associate professor, Department of Mathematics, Tabriz Branch, Islamic Azad University, Tabriz, Iran.

*-Corresponding Author: f.modarres@iauc.ac.ir

3- Assistant professor, Department of Mathematics, Tabriz Branch, Islamic Azad University, Tabriz, Iran.

10.30495/QJOPM.2020.1867405.2443

((مقاله پژوهشی))

ارائه الگوریتم خوشه‌بندی جدید به‌منظور بهره‌وری در عملیات داده‌کاوی (مطالعه داده‌های استاندارد یوسی‌آی)

ژیلا نصیری[۱]، فرزین مدرس خیابانی*[۲]، نیما آذرمیرشتربانی[۳]

(دریافت: ۹۸/۱۱/۳۰–پذیرش نهایی: ۹۹/۰۴/۲۸)

چکیده

روش‌های خوشه‌بندی و بهره‌وری آنها در عملیات داده‌کاوی توسعه زیادی یافته‌اند. نیاز مدیران به داده‌های دسته‌بندی‌شده و بهره‌وری روش‌های خوشه‌بندی در امر مدیریت و تصمیم‌گیری، به گسترش روش‌های داده‌کاوی ضرورت بخشیده است. الگوریتم بهینه‌سازی نهنگ روش عمومی است که در حل مسائل متعددی کاربرد دارد. در این الگوریتم جواب‌های آغازین به‌صورت تصادفی انتخاب می‌شوند. الگوریتم کی–میانگین یک روش خوشه‌بندی پرکاربرد است که به دلیل سادگی و کوتاه بودن مراحل، بسیار موردتوجه محققان قرار می‌گیرد. در این مقاله این مزیت الگوریتم کی– میانگین را برای افزایش توانایی الگوریتم بهینه‌سازی نهنگ در خوشه‌بندی داده‌ها به‌کاررفته است. الگوریتم پیشنهادی ترکیبی از الگوریتم‌های کی–میانگین و خوشه‌بندی نهنگ است. در این پژوهش الگوریتم جدید و چند الگوریتم خوشه‌بندی دیگر را بر روی مجموعه داده‌های واقعی و شناخته شده اجرا شده است. نتایج عددی نشان می‌دهد که الگوریتم جدید ازنظر کیفیت جواب‌ها و انحراف استاندارد مقادیر جواب‌های نهایی، نتایج مطلوبی نشان می‌دهد.

واژه‌های کلیدی: بهره‌وری، خوشه‌بندی، داده‌کاوی، هوش جمعی.

۱. استادیار گروه ریاضی، واحد تبریز، دانشگاه آزاد اسلامی، تبریز، ایران
۲. دانشیار گروه ریاضی، واحد تبریز، دانشگاه آزاد اسلامی، تبریز، ایران
*– نویسنده مسئول f.modarres@iauc.ac.ir
۳. استادیار گروه ریاضی، واحد تبریز، دانشگاه آزاد اسلامی، تبریز، ایران

مقدمه

وقتی داده‌ها بدون برچسب هستند و گروه‌بندی در بین آنها صورت نگرفته، عملیات داده‌کاوی روی این داده‌ها یادگیری بدون نظارت نامیده می‌شود. در این شرایط تقسیم‌بندی داده‌ها، برحسب ویژگی‌ها به کلاس‌های طبیعی خوشه‌بندی نامیده می‌شود (آرماندو[1] و فرمانی، ۲۰۱۴). خوشه‌بندی افرازی به‌سادگی مجموعه داده‌ها را به خوشه‌هایی افراز می‌کند که بیشترین شباهت در عناصر داخل یک خوشه و بیشترین تفاوت در عناصر بین خوشه‌های مختلف وجود دارد. خوشه‌بندی افرازی اطلاعات کمتری ارائه می‌دهد اما برای کار روی داده‌های بزرگ توسعه یافته است (سندر[2]، ۲۰۰۳). در روش‌های خوشه‌بندی افرازی تعداد خوشه‌ها از قبل معلوم است و جستجو برای پیدا کردن خوشه‌های بهینه برحسب توابع هدف مربوطه انجام می‌شود (جین[3]و همکاران، ۱۹۹۹؛ رکاچ و مایمون[4]، ۲۰۱۵). روش‌های قدیمی خوشه‌بندی افرازی مانند روش کی-میانگین[5] در بهینه محلی گیر می‌کنند و پیدا کردن مرکز خوشه‌های اولیه مشکل دیگر این روش‌ها است. برای مقابله با این محدودیت‌ها، محققان دریافته‌اند که الگوریتم‌های تکاملی و الگوریتم‌های الهام گرفته از طبیعت می‌توانند جایگزین‌های خوبی برای روش‌های قدیمی در حل مسائل عملی باشند. نیکنام و همکاران یک روش ترکیبی بر اساس الگوریتم کلونی مورچه و تبرید شبیه‌سازی شده معرفی کردند (نیکنام و همکاران، ۲۰۰۹). احمدی‌فر و مدرس یک الگوریتم خوشه‌بندی با ترکیب الگوریتم‌های تراکم ذرات و کی-میانگین ارائه دادند که از مزیت‌های هر دو الگوریتم استفاده می‌کند. برای جستجوی سرتاسری، ابتدا الگوریتم تراکم ذرات اجرا می‌شود سپس برای همگرایی سریع‌تر جریان خوشه‌بندی با الگوریتم کی-میانگین ادامه می‌یابد (احمدی‌فر و مدرس، ۲۰۰۸). سندیپ و پانکاژ[6] یک روش ترکیبی دیگر پیشنهاد داده‌اند که الگوریتم تراکم ذرات و خوشه‌بندی فازی را دنباله‌دار اجرا کرده و برای خوشه‌بندی داده‌ها استفاده می‌کنند. نتایج عددی نشان می‌دهد روش جدید جواب‌هایی باکیفیت بالاتر به‌دست می‌آورد و همچنین الگوریتم در بهینه محلی گیر نمی‌افتد (سندیپ و پانکاج، ۲۰۱۴). آرماندو و فرمانی برای افزایش کارایی الگوریتم کی-میانگین، آن را با الگوریتم کلونی زنبور مصنوعی ترکیب کردند، الگوریتم ترکیبی حاصل در پیدا کردن جواب‌های سراسری بسیار مؤثر عمل می‌کند (آرماندو و

1. Armando
2. Sander
3. Jain
4. Rokach & Maimon
5. k-mean
6. Sandeep & Pankaj

فرمانی، ۲۰۱۴). کارتیکیان و کریستوفر[1] یک الگوریتم دیگر با ترکیب الگوریتم تراکم ذرات و الگوریتم کلونی زنبور مصنوعی پیشنهاد دادند. آن‌ها الگوریتم پیشنهادی را با چندین الگوریتم خوشه‌بندی پرکاربرد ازنظر کارایی مقایسه کردند (کارتیکیان و کریستوفر، ۲۰۱۴).

الگوریتم بهینه‌سازی نهنگ یکی از روش‌های بهینه‌سازی هوش جمعی است که اخیراً موردمطالعه قرار گرفته است (میرجلیلی و لویس[2]، ۲۰۱۶). الگوریتم نهنگ درواقع، شبیه‌سازی رفتار شکار شبکه حبابی نهنگ‌های گوژپشت برای حل مسائل عددی بهینه‌سازی است. نتایج استفاده از این الگوریتم را با نتایج حاصل از سایر الگوریتم‌های فراابتکاری پرکاربرد نظیر الگوریتم تراکم ذرات روی ۲۹ مساله بهینه‌سازی معیار مقایسه کرده‌اند. نتایج مقایسه برتری الگوریتم فراابتکاری جدید و بهره‌وری آن را در جستجوی سراسری، جستجوی محلی، اجتناب از گیر افتادن در بهینه محلی و همگرایی به جواب بهینه سراسری نشان می‌دهد. الگوریتم نهنگ به دلیل سادگی در پیاده‌سازی، کم بودن تعداد پارامترهای کنترلی و داشتن دو فرمول جداگانه برای جستجوی سراسری و جستجوی موضعی موردتوجه گسترده پژوهشگران قرار گرفته است. ما قبلاً، از این الگوریتم برای حل مساله خوشه‌بندی داده‌ها استفاده کرده‌ایم، نتایج عددی حاصل بسیار مطلوب هستند (نصیری و مدرس، ۲۰۱۸). سؤال اصلی پژوهش حاضر این است که آیا می‌توان الگوریتم خوشه‌بندی جدیدی ارائه داد که خوشه‌بندی داده‌ها را با دقت بالاتری نسبت به روش‌های موجود، انجام داده و بهره‌وری عملیات داده‌کاوی را افزایش دهد؟ با توجه به سؤال مطروحه، هدف اصلی این پژوهش طراحی یک روش جدید در داده‌کاوی سیستم‌ها برای خوشه‌بندی داده‌های عددی و توصیفی است که عملکرد بهتری در مقایسه با روش‌های مشابه داشته باشد. در این مقاله برای بهبود الگوریتم خوشه‌بندی نهنگ، انتخاب حریصانه معمول در الگوریتم‌های هوش جمعی را با قانون انتخاب مقید دب[3] جایگزین کرده‌ایم (بابالیک[4] و همکاران، ۲۰۱۷). از این طریق با حذف جواب‌های نشدنی، روند جستجوی جواب بهینه را تحت تأثیر قرار داده‌ایم. همچنین مرحله اکتشاف در الگوریتم خوشه‌بندی نهنگ نسبتاً طولانی است در این مقاله الگوریتم نهنگ را با الگوریتم کی–میانگین ترکیب نموده‌ایم تا با استفاده از روش سریع کی–میانگین مرحله اکتشاف را کوتاه‌تر کرده و جستجو را هدف‌دار انجام دهیم. در ادامه عملکرد الگوریتم ترکیبی پیشنهادی را روی چند مجموعه داده واقعی اجرا نموده و نتایج را با

1. Karthikeyan & Christopher
2. Levis
3. Deb’ rule
4. Babalik

الگوریتم‌های خوشه‌بندی کی-میانگین، الگوریتم خوشه‌بندی نهنگ و چند الگوریتم پرکاربرد دیگر مقایسه کرده‌ایم.

گروه‌بندی عناصر یک مجموعه داده بدون برچسب، برحسب برخی معیارهای شباهت خوشه‌بندی نامیده می‌شود. بنابراین، عناصر شبیه به هم در یک خوشه و عناصر غیرمشابه در خوشه‌های متفاوت قرار می‌گیرند. هماهنگ با بسیاری از پژوهش‌ها، در این مقاله کمینه کردن مقدار تابع فاصله درون خوشه‌ای را به‌عنوان تابع هدف در نظر گرفته‌ایم. به‌طورکلی مسأله به این صورت بیان می‌شود که فرض کنید $s = \{x_1, x_2, \dots, x_n\}$ یک مجموعه n عنصری از داده‌های m بعدی باشد. هر عنصر x_i با یک بردار به‌صورت $\{x_{i1}, x_{i2}, \dots, x_{im}\}$ نشان داده می‌شود که هر x_{ij} مقدار توصیف j ام عنصر iام را نشان می‌دهد. هدف خوشه‌بندی اختصاص هر عنصر x_i به یکی از k خوشه در مجموعه $z = \{c_1, c_2, \dots, c_K\}$ است به‌طوری که فاصله بین هر عنصر داده و مرکز خوشه کمترین مقدار ممکن باشد، هر عنصر حتماً در یکی از خوشه‌ها قرار گیرد، هیچ خوشه‌ای خالی نباشد و هیچ عنصری هم‌زمان در بیش از یک خوشه حضور نداشته باشد. به‌این‌ترتیب مسأله خوشه‌بندی عبارت است از کمینه کردن فاصله اقلیدسی زیر:

$$d(s,z) = \left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\|x_i - z_k\|^2\right)^{\frac{1}{2}} \quad (۱)$$

که در آن x_i، عنصر داده iامو z_k مرکز خوشه kام را نشان می‌دهند. w_{ik} وزن ارتباطی عنصر داده با خوشه مربوطه است و به‌صورت زیر محاسبه می‌شود:

$$w_{ik} = \begin{cases} 1: \|x_i - z_k\|^2 = \min\limits_{1\le j\le n} \|x_i - z_j\|^2 \\ \text{سایر } 0 \end{cases} \quad (۲)$$

مطابق رابطه (۲) هر عنصر داده به نزدیک‌ترین مرکز خوشه نسبت داده می‌شود.

الگوریتم کی-میانگین، یک روش تکراری قدیمی است که مجموعه داده‌ها را به تعداد از قبل مشخص‌شده‌ای از خوشه‌ها افراز می‌کند. این الگوریتم مرکز خوشه‌های اولیه را به‌صورت تصادفی انتخاب می‌کند. مزیت این روش سرعت اجرای بسیار بالای آن است و عیب آن وابستگی جواب‌های نهایی به مرکز خوشه‌های اولیه است. برای اجرای این الگوریتم ابتدا k عنصر از مجموعه داده

به‌صورت تصادفی به‌عنوان مرکز خوشه‌های اولیه انتخاب می‌شود. هر عنصر داده به نزدیک‌ترین مرکز خوشه اختصاص داده می‌شود. مرکز خوشه جدید با محاسبه میانگین عناصر قرار گرفته در هر خوشه به دست می‌آید. مراحل فوق تا همگرایی به جواب‌های بهینه محلی تکرار می‌شود. بنابراین، این الگوریتم یک روش با جواب بهینه موضعی در همسایگی جواب‌های تصادفی آغازین به دست می‌آورد و با تغییر جواب‌های آغازین تصادفی در هر تکرار الگوریتم، امکان تغییر جواب‌های نهایی وجود دارد (کانونگو[1] و همکاران، ۲۰۰۲).

الگوریتم بهینه‌سازی نهنگ یک الگوریتم جدید فراابتکاری است که برای مسائل بهینه‌سازی عددی معرفی شده است (میرجلیلی و لویس، ۲۰۱۶). الگوریتم رفتار شکار هوشمندانه نهنگ‌های گوژپشت را شبیه‌سازی می‌کند. این روش غذایابی روش شکار شبکه حبابی نامیده می‌شود که فقط در نهنگ‌های گوژپشت مشاهده شده است. به بیان ساده، در رفتار شکار شبکه حبابی نهنگ‌ها ابتدا به عمق ۱۲ متری زیر آب رفته و حباب‌هایی در مسیر حلزونی در امتداد دایره‌های تنگ شونده در اطراف شکار تولید می‌کنند، سپس به دنبال حباب‌ها به طرف سطح آب و به سمت شکار شنا می‌کنند. نهنگ‌های گوژپشت توانایی پیدا کردن شکار و محاصره آن را دارند. الگوریتم نهنگ موقعیت بهترین عامل جستجو را محل شکار یا نزدیک به نقطه بهینه فرض می‌کند و دیگر عامل‌های جستجو تلاش خواهند کرد تا موقعیت خود را در جهت بهترین عامل جستجو اصلاح کنند. این رفتار نهنگ‌ها با روابط زیر فرموله می‌شود:

$$\vec{D} = \left|\vec{C}.\overrightarrow{X^*}(t) - \vec{X}(t)\right| \quad (۳)$$

$$\vec{X}(t+1) = \overrightarrow{X^*}(t) - \vec{A}.\vec{D} \quad (۴)$$

که t شماره تکرار فعلی است، X^* بردار مکان بهترین جواب به‌دست آمده تا تکرار فعلی، $\vec{X}$ بردار مکان هر عامل جستجو، $|\ |$ نماد قدر مطلق و. نماد ضرب داخلی بردارها است. بردارهای ضریب $\vec{A}$ و $\vec{C}$ از روابط زیر به دست می‌آیند.

$$\vec{A} = 2\vec{a}.r - \vec{a} \quad (۵)$$

$$\vec{C} = 2r \quad (۶)$$

که $\vec{a}$ در طول تکرارها از مقدار اولیه ۲ به مقدار ۰ به‌صورت خطی کاهش می‌یابد و r یک عدد تصادفی بین صفر و یک است.

1. Kanungo

روش هجوم شبکه حبابی، استراتژی شبکه حبابی ترکیبی از مکانیسم دایره‌های تنگ شونده و تغییر مکان حلزونی می‌باشد. با کاهش مقدار $\vec{a}$ در معادله (۵) رفتار دایره‌های تنگ شونده شبیه‌سازی می‌شود. توجه کنید دامنه نوسان $\vec{A}$ نیز که یک مقدار تصادفی در بازه $[-a, a]$ است، با کاهش $\vec{a}$ در طول تکرارها کاهش می‌یابد. با تنظیم مقادیر $\vec{A}$ بین ۱- و ۱، موقعیت جدید عامل جستجو می‌تواند در هر نقطه بین موقعیت قبلی آن و مکان بهترین عامل جستجوی فعلی قرار بگیرد. در مکانیسم تغییر مکان حلزونی یک معادله حلزونی بین موقعیت عامل جستجو و شکار برای شبیه‌سازی حرکت مارپیچی نهنگ‌های گوژپشت به‌صورت زیر در نظر گرفته می‌شود:

$$\overrightarrow{D'} = \left|\overrightarrow{X^*}(t) - \vec{X}(t)\right| \quad (۷)$$

$$\vec{X}(t+1) = \overrightarrow{D'}.e^{bl}cos(2\pi l) + \overrightarrow{X^*}(t) \quad (۸)$$

که $\overrightarrow{D'}$ فاصله بین نهنگ و شکار، b عدد ثابت، l عدد تصادفی بین ۱- و ۱ می‌باشد. همچنین ضرب بین بردارها مؤلفه به مؤلفه انجام می‌شود. در واقع نهنگ‌های گوژپشت در امتداد یک مسیر حلزونی شکل شنا می‌کنند و هم‌زمان دایره‌های تنگ شونده‌ای را طی می‌کنند. این هم‌زمانی را با در نظر گرفتن یک احتمال ۵۰ درصد برای هر یک از حرکت‌های مدور یا حلزونی در طول تکرارهای الگوریتم مدل‌سازی کرده‌اند.

جستجوی شکار، اغلب الگوریتم‌های فرا ابتکاری جستجوی سراسری فضای جواب را برای پیدا کردن جواب بهینه با انتخاب تصادفی انجام می‌دهند. در روش شبکه حبابی نیز بهترین موقعیت شناخته شده نیست، بنابراین نهنگ‌های گوژپشت به‌صورت تصادفی به جستجوی شکار می‌پردازند. در مقابل جستجوی موضعی در این مرحله $\vec{A}$ یک بردار با مقادیر تصادفی کمتر از ۱- و یا بزرگتر از ۱ در نظر گرفته می‌شود. به‌این‌ترتیب عامل جستجو قادر به حرکت در فواصل دورتر از مکان قبلی بوده و همه فضای جواب را جستجو می‌کند. به‌عبارت‌دیگر در این مرحله موقعیت عوامل جستجو به جای بهترین موقعیت پیدا شده فعلی توسط یک موقعیت تصادفی بروز رسانی می‌شود. بنابراین تغییر مکان عوامل جستجو از روابط زیر محاسبه می‌شود:

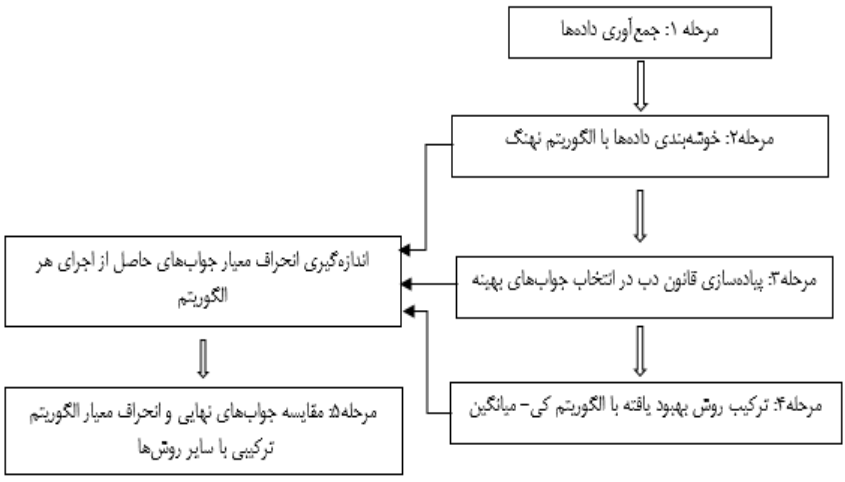
$$\vec{D} = \left|\vec{C}.\overrightarrow{X_{rand}} - \vec{X}\right| \quad (۹)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A}.\vec{D} \quad (۱۰)$$

که $\overrightarrow{X_{rand}}$ موقعیت برداری است که به‌صورت تصادفی انتخاب شده است.

الگوریتم بهینه‌سازی نهنگ با یک مجموعه از جواب‌های تصادفی شروع می‌شود سپس در هر تکرار موقعیت عامل‌های جستجو طبق توضیحات بالا اصلاح می‌کنند. الگوریتم نهنگ یک الگوریتم بهینه‌سازی سرتاسری است. تغییرات تطبیقی بردار جستجوی $\vec{A}$ به این الگوریتم اجازه می‌دهد ابتدا اکتشاف و سپس بهره‌برداری انجام دهد و به‌سادگی بین این دو مرحله جابجا شود. توانایی بالای الگوریتم نهنگ در جستجوی سراسری مربوط به مکانیسم تغییر مکان نهنگ‌ها مطابق رابطه (۱۰) می‌باشد، جستجوی موضعی دقیق و تأکید بر همگرایی به جواب سراسری با دقت بالا توسط روابط (۴) و (۸) برآورده می‌شود. این روابط الگوریتم را قادر به گریز از جواب‌های موضعی کرده و سرعت همگرایی را در طول دوره تکرارها بالا می‌برند. همچنین فقط دو پارامتر داخلی باید قبل از اجرا تنظیم شود.

اهداف این تحقیق عبارت‌اند از: ۱. ارائه یک روش خوشه‌بندی هوشمند ترکیبی با بهره‌وری بالا در مقایسه با روش‌های موجود، ۲. کمینه کردن انحراف معیار جواب‌های روش پیشنهادی در اجراهای متعدد الگوریتم، ۳. دسته‌بندی مفید و هدفمند اطلاعات با روش سریع و هوشمند و با بهره‌وری بالا جهت تسهیل در تصمیم‌گیری مدیران. شکل ۱ مدل این تحقیق را نشان می‌دهد.



شکل شماره ۱: مدل تحقیق

Figure 1: Research model

ابزار و روش

این پژوهش ازنظر نتیجه کاربردی است و ازنظر روش اکتشافی می‌باشد زیرا درباره وجود روش بهتری برای بهبود کیفیت داده‌کاوی و افزایش بهره‌وری تحقیق می‌کند. داده‌های مورداستفاده داده‌های واقعی از نوع عددی و کیفی هستند. روش تحقیق شبیه‌سازی کامپیوتری پدیده طبیعی برای حل مساله بهینه‌سازی می‌باشد. سپس روش پیشنهادی با روش‌های موجود و مشابه مقایسه شده است تا بهبود نتایج حاصل از حل مساله با عملکرد الگوریتم جدید و ضرورت پژوهش در معرض نمایش قرار بگیرد. نتیجه این پژوهش مدیران را در دسته‌بندی بهینه اطلاعات سسیتم با اطلاعات وسیع و اخذ تصمیم‌هایی با بهره‌وری بالا یاری خواهد کرد.

الگوریتم بهینه‌سازی نهنگ یک الگوریتم بهینه‌سازی فرا ابتکاری جدید است که یک الگوریتم قدرتمند و بر اساس هوش جمعی می‌باشد. که برای کار با داده‌های وسیع رایج در سیستم‌های مدیریتی امروزی مناسب می‌باشد. هدف الگوریتم پیدا کردن موقعیت مکانی عامل جستجویی است که بهترین مقدار را برای تابع هدف مفروض به دست آورد. در این بخش، ما تصمیم داریم مساله خوشه‌بندی داده‌ها را با استفاده الگوریتم نهنگ حل کنیم. ما در اینجا فاصله درونی خوشه‌ها را به‌عنوان تابع هدف انتخاب کرده‌ایم که اندازه بین مرکز خوشه و بردارهای داده همان خوشه را با استفاده از فرمول‌های (۱) و (۲) اندازه می‌گیرد.

الگوریتم کی–میانگین یک روش خوشه‌بندی پرطرفدار است که سریع‌تر از الگوریتم خوشه‌بندی نهنگ همگرا می‌شود، اما معمولاً دقت خوشه‌بندی آن پایین‌تر است. به‌عبارت‌دیگر الگوریتم کی–میانگین بعد از تعداد کمتری ارزیابی تابع هدف همگرا می‌شود. این در حالی است که دقت جواب‌ها به مرکز خوشه‌های تصادفی آغازین بستگی دارد و لذا احتمال گیر افتادن در بهینه محلی وجود دارد. در این مقاله برای استفاده آر مزیت الگوریتم کی–میانگین و گریز از گیر افتادن در بهینه محلی، یک الگوریتم خوشه‌بندی با ترکیبی پیشنهاد کرده‌ایم. الگوریتم پیشنهادی ابتدا، الگوریتم کی–میانگین را یک بار اجرا می‌کند که شرط پایان آن ماکزیمم تعداد تکرارهای طی شده یا بدون تغییر ماندن مرکز خوشه‌ها در دو تکرار متوالی می‌باشد. جواب بهینه حاصل از این اجرا به‌عنوان یکی از عامل‌های جستجوی الگوریتم خوشه‌بندی نهنگ در نظر گرفته می‌شود و بقیه جمیعت آغازین به‌صورت تصادفی انتخاب می‌شوند. در ادامه، خوشه‌بندی نهنگ با استفاده از قانون دب به جای قانون انتخاب حریصانه اجرا می‌شود.

از نظر تئوری، الگوریتم خوشه‌بندی ترکیبی جدید یک بهینه‌ساز سرتاسری است که شامل جستجوی سرتاسری و موضعی می‌باشد. مکانیسم توصیف‌شده یک فضای جستجو در همسایگی

بهترین جواب پیدا شده، تعریف می‌کند و به بقیه عوامل جستجو اجازه می‌دهد تا از بهترین جواب فعلی برای بهبود وضعیت خود استفاده کنند. مقدار تطبیقی بردار A به الگوریتم اجازه می‌دهد تا به‌آسانی مراحل اکتشاف و بهره‌برداری را انجام دهد. به‌این‌ترتیب که با کاهش تطبیقی A تعدادی از تکرارها روی اکتشاف تمرکز می‌کنند و در ادامه بقیه تکرارها با بهره‌برداری ادامه می‌دهند. استفاده از قانون دب با کنار گذاشتن جواب‌های نشدنی در هر مقایسه مسیر جستجو را کوتاه می‌کند و حجم محاسبات را کاهش می‌دهد. همچنین استفاده از الگوریتم کی- میانگین سرعت جستجو را بالا می‌برد. در اینجا برای بررسی بهره‌وری و کارکرد الگوریتم ترکیبی جدید از هفت مجموعه داده پایگاه داده دانشگاه ایروین کالیفرنیا[1] استفاده کرده‌ایم (دیو و گرف[2]، ۲۰۱۹). این مجموعه داده‌ها به‌طور گسترده در مقالات برای اندازه‌گیری کارایی الگوریتم‌های جدید استفاده می‌شوند اسامی و اطلاعات این مجموعه داده‌ها در جدول ۱ خلاصه شده است. برای نشان دادن توانایی الگوریتم ترکیبی جدید آن را با الگوریتم خوشه‌بندی نهنگ که در (نصیری و مدرس، ۲۰۱۸) معرفی شده است و ۴ الگوریتم پرکاربرد دیگر مقایسه کرده‌ایم. این الگوریتم‌ها عبارت‌اند از خوشه‌بندی کلونی زنبور (کارابوقا و اوزتورک[3]، ۲۰۰۹)، خوشه‌بندی تراکم ذرات (مروه و انگلهچت[4]، ۲۰۰۳)، خوشه‌بندی با الگوریتم ژنتیک (موالیک و همکاران، ۲۰۰۲) و الگوریتم کی-میانگین. همه روش‌ها را در نرم‌افزار متلب ۲۰۱۳ و بر روی سیستم کامپیوتری با حافظه ۴ گیگابایت و سرعت پردازش ۱/۷۳ گیگاهرتز با سیستم‌عامل ویندوز ایکس پی اجرا کرده‌ایم. شرط پایان الگوریتم کی-میانگین را در هر اجرا سپری شدن ۱۰ تکرار یا بدون تغییر ماندن مرکز خوشه‌ها در دو تکرار متوالی در نظر گرفته‌ایم. مقادیر پارامترهای اولیه همه الگوریتم‌های مورد مقایسه در جدول ۲ آمده است.

جدول شماره ۱: اطلاعات مجموعه داده‌ها

Table 1: Information of data sets

مجموعه داده‌ها	تعداد خوشه‌ها	تعداد عناصر توصیف (بعد داده‌ها)	تعداد عناصر داده
Iris	3	4	150
Win	3	13	178
Cmc	3	9	1473
Balance	3	4	625
Cancer	2	11	569
Glass	6	9	214
Thyroid	3	5	215

1. Irvine University of California
2. Due & Graff
3. Karaboga & Ozturk
4. Merve & Engelhrecht

جدول شماره ۲: تنظیم پارامتر الگوریتم‌های مورد مقایسه
Table 2: Parameter setting of compared algorithms

پارامتر	مقدار	پارامتر	مقدار	پارامتر	مقدار
ژنتیک		کی- میانگین		نهنگ	
جمیعت	50	تعداد تکرار	10	جمیعت	50
ضریب تقاطع	0.8			ضریب b	1
نسبت جهش	0.001			تعداد تکرارها	30
تعداد تکرار	1000				
کلونی زنبور		تراکم ذرات		ترکیبی نهنگ	
تعداد زنبورها	10	جمیعت	50	جمیعت	50
محدودیت	10	$c_1 = c_2$	2	ضریب b	1
تعداد تکرار	1000	ω	1	تعداد تکرارهای کی- میانگین	10
		تعداد تکرار	1000	تعداد تکرار نهنگ	300

یافته‌ها

جدول ۳ نتایج شبیه‌سازی الگوریتم‌های خوشه‌بندی را با مجموعه داده‌های مفروض نمایش می‌دهد. مقادیر گزارش‌شده برای هر الگوریتم خوشه‌بندی، میانگین مقدار تابع معیار فاصله درون خوشه‌ای برای ۲۰ اجرای هر الگوریتم و انحراف معیار استاندارد آن‌ها می‌باشند. همچنین رتبه هر الگوریتم در مقایسه با سایر الگوریتم‌ها برای هر مجموعه داده، برحسب مقدار تابع فاصله درون خوشه در جدول ۳ آمده است. با توجه به نتایج عددی روش جدید کمترین مقدار میانگین تابع فاصله را برای داده‌های iris، win، cmc، glass،cancer و thyroid به دست می‌آورد. یعنی در حل این مسائل الگوریتم جدید، در شش مساله از هفت مورد انتخابی عملکرد الگوریتم جدید در مقایسه با سایر روش‌ها در رتبه اول قرار می‌گیرد. و فقط در مجموعه داده balance الگوریتم جدید در رتبه دوم قرار دارد. در مساله iris بعد از الگوریتم ترکیبی جدید، الگوریتم نهنگ در رتبه دوم قرار دارد که به این معنی است که تغییرات انجام‌گرفته روی الگوریتم نهنگ معمولی کیفیت جواب را تحت تأثیر قرار می‌دهد و جواب بهینه بهتری را به دست می‌آورد. همچنین مقدار انحراف معیار برای روش جدید ۰/۰۶ به دست می‌آید که در مقایسه با سایر روش‌ها بسیار جزئی و قابل‌چشم‌پوشی است به این معنی که روش جدید تقریباً، در همه اجراها به جواب بهینه یکسانی همگرا می‌شود. در مجموعه داده win نیز که یک مجموعه داده ۱۳ بعدی است روش ترکیبی جدید در رتبه اول قرار دارد و همچنین مقدار انحراف معیار استاندارد جواب‌ها ۱۱/۹۳ به دست آمده که در مقایسه با مقدار مشابه در الگوریتم

نهنگ که برابر با ۴۷/۴۵ محاسبه شده است، مقدار کمتری دارد و بهبود عملکرد الگوریتم هوش جمعی نهنگ را در ترکیب با روش کی–میانگین تأیید می‌کند و توانایی بالای الگوریتم جدید را در کار با داده‌های با بعد بالا نشان می‌دهد. در مساله cmc روش جدید کمترین مقدار تابع فاصله درون خوشه را به دست می‌آورد و در رتبه اول قرار می‌گیرد. مقدار انحراف معیار جواب‌های الگوریتم پیشنهادی در این مساله برابر ۱/۷۱ محاسبه شده که در مقایسه با انحراف معیار سایر روش‌ها بسیار جزئی است. با توجه به ۹ بعدی بودن داده‌های این مجموعه بار دیگر توانایی الگوریتم در حل مسائل با بعد بالا تأیید می‌شود.

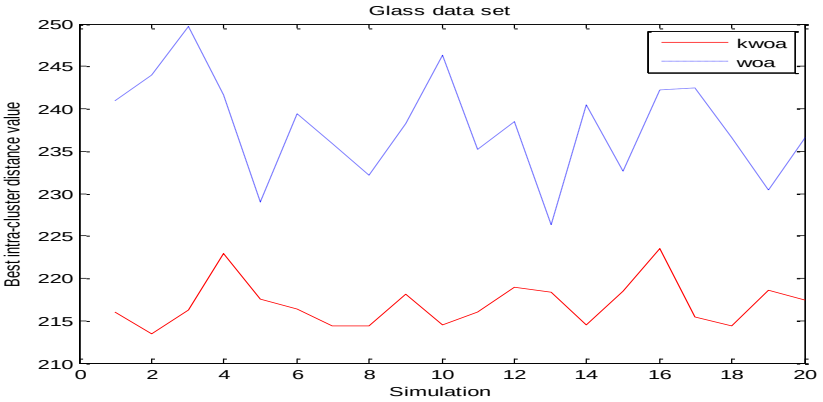
در مجموعه داده ۱۱ بعدی cancer نیز روش جدید کمترین مقدار تابع فاصله و انحراف استاندارد را محاسبه می‌کند و در رتبه اول قرار دارد. نکته جالب‌توجه این است که مقادیر تابع فاصله درون خوشه و انحراف معیار استاندارد الگوریتم نهنگ ترکیبی به ازای همه مجموعه داده‌های موردمطالعه کمتر از مقادیر الگوریتم نهنگ می‌باشد. این نتیجه با کنار گذاشتن جواب‌های نشدنی توسط قانون دب و هدایت بهتر الگوریتم به سوی جواب‌های با کیفیت بالا میسر شده است. رتبه میانگین الگوریتم ترکیبی جدید در سطر آخر جدول ۳ برابر با ۱/۱۴ درج شده است. این نتیجه حاصل ترکیب الگوریتم نهنگ با روش کی–میانگین می‌باشد. همچنین مقادیر انحراف معیار استاندارد جواب‌ها در مقایسه با جواب‌های خوشه‌بندی نهنگ و کی–میانگین بسیار کمتر می‌شود. نتیجه اینکه الگوریتم پیشنهادی این مقاله، به جواب‌هایی با کیفیت بالا و انحراف معیار استاندارد جزئی و قابل‌چشم‌پوشی در بیشتر مجموعه داده‌ها همگرا می‌شود. به‌این‌ترتیب برتری روش جدید در مقایسه با الگوریتم‌های شناخته‌شده و پرکاربرد آشکار است و می‌تواند به‌عنوان یک روش اکتشافی فراابتکاری مؤثر و توانا برای به دست آوردن جواب بهینه یا جواب‌های نزدیک به بهینه در مسائل داده‌کاوی جهت افزایش بهره‌وری مورد استفاده قرار بگیرد. بعد از ترکیبی جدید الگوریتم خوشه‌بندی نهنگ با رتبه میانگین ۲/۵۷ قرار دارد. الگوریتم خوشه‌بندی تراکم ذرات در رتبه سوم قرار می‌گیرد. این به آن معنی است که الگوریتم جدید نتایج عددی فوق‌العاده در حل مسائل از خود نشان می‌دهد.

جدول شماره ۳: نتایج عددی الگوریتم خوشه‌بندی ترکیبی نهنگ و مقایسه با سایر الگوریتم‌ها
Table 3: Numerical results of hybrid whale clustering and compared with other algorithms

مجموعه داده	معیار	ژنتیک	کلونی زنبور	تراکم ذرات	کی-میانگین	نهنگ	ترکیبی نهنگ
Iris	میانگین انحراف معیار رتبه	119.51 10.28 6	132.75 5.61 5	98.74 0.22 3	103.74 11.55 4	97.05 0.43 2	**96.75** 0.06 1
Win	میانگین انحراف معیار رتبه	16565.82 269.48 4	16595.30 65.90 5	16317.56 12.80 2	17018.89 722.06 7	16319.71 47.45 3	**16312.99** 11.93 1
Cmc	میانگین انحراف معیار رتبه	6125.82 144.46 5	6203.49 70.56 6	5573.30 12.44 2	5890.63 47.07 4	5641.90 84.40 3	**5537.48** 1.71 1
Balance	میانگین انحراف معیار رتبه	1433.62 3.13 5	1436.51 1.51 6	**1425.51** 0.29 1	1426.91 3.80 4	1426.83 2.45 3	1426.81 2.63 2
Cancer	میانگین انحراف معیار رتبه	4531.15 227.70 6	3089.67 7.17 5	3039.5 2.31 2	3061.45 0.76 5	3051.12 8.02 3	**3027.58** 0.17 1
Glass	میانگین انحراف معیار رتبه	237.92 14.32 4	252.26 10.37 6	247.91 0.56 5	239.62 15.80 3	237.14 22.83 2	**231.32** 3.14 1
Thyroid	میانگین انحراف معیار رتبه	1889.42 6.00 4	1885.72 8.93 3	1896.00 0.02 5	2014.51 76.45 6	1885.26 13.02 2	**1877.81** 12.24 1
	میانگین رتبه‌بندی	4.85	5.14	2.85	4.57	2.57	**1.14**

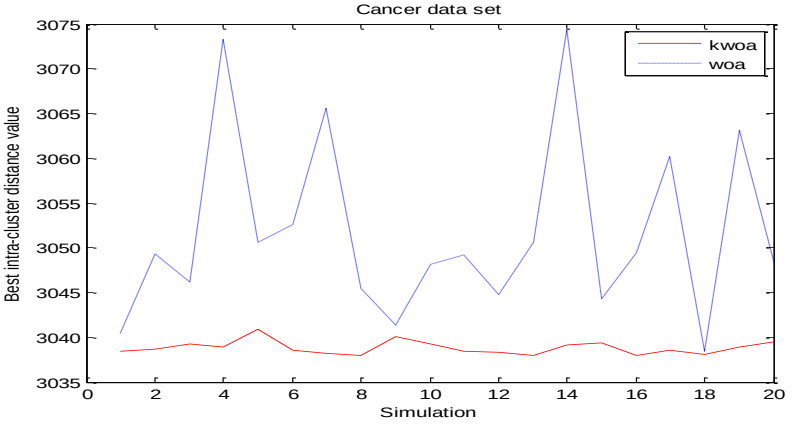
در شکل ۲ و ۳، کمترین مقدار تابع فاصله درون خوشه برای الگوریتم پیشنهادی و الگوریتم خوشه‌بندی نهنگ در ۲۰ شبیه‌سازی با جواب‌های اولیه تصادفی روی مجموعه داده Glass و Cancer نمایش داده‌شده است. کم بودن مقادیر تابع فاصله جواب‌های بهینه و اختلاف اندک آنها با هم در طول ۲۰ اجرا الگوریتم ترکیبی نهنگ آشکار است. با توجه به این شکل‌ها مشاهده می‌کنیم که چگونه با تغییراتی روی جواب آغازین تصادفی و روش انتخاب بهترین عامل جستجو در هر مرحله، عملکرد الگوریتم بهینه‌سازی نهنگ را در خوشه‌بندی تحت تأثیر قرار داده و کیفیت جواب‌ها

را بهبود می‌بخشد. در کنار نتایج عددی فوق‌العاده حاصل از اجرای الگوریتم، تعداد کم پارامترهایی که باید از قبل تعیین شوند، ساده بودن پیاده‌سازی و اجرا، توانایی گریز از بهینه محلی و رسیدن به بهینه سراسری از مزیت‌های الگوریتم پیشنهادی هستند. این موارد منجر به عملکرد الگوریتم در به دست آوردن مقدار تابع فاصله کوچکتر می‌شود همچنین پراکندگی مقادیر تابع هدف در بازه کوچک‌تری قرار دارد.



شکل شماره ۲: اجرای الگوریتم ترکیبی جدید و الگوریتم نهنگ روی داده glass

Figure 2: Running the kwoa algorithm and woa algorithm on glass data



شکل شماره ۳. اجرای الگوریتم ترکیبی جدید و الگوریتم نهنگ روی داده cancer

Figure 3: Running the kwoa algorithm and woa algorithm on cancer data

بحث و نتیجه‌گیری

هدف اصلی این پژوهش پاسخ دادن به این سؤال بود که آیا می‌توان روش خوشه‌بندی هوشمند جدیدی ارائه داد که عملکردی بهتر از روش‌های موجود داشته باشد؟ در این مقاله الگوریتم جدید، سرعت الگوریتم کی–میانگین در کنار جستجوی کامل خوشه‌بندی نهنگ به کار گرفته است تا جواب بهتری را با انحراف معیار کمتری در اجراهای مختلف به دست آورد. همچنین، از قانون انتخاب دب در انتخاب جواب‌های بهتر را جایگزین انتخاب معمول در الگوریتم‌های فرا ابتکاری شده است. به‌این‌ترتیب با حذف جواب‌های نشدنی و بی‌کیفیت در روند جستجو از یک‌سو موجب تسریع فرایند جستجو و از سوی دیگر باعث جلوگیری از جستجو در مناطق پرت فضای جواب شده است. این امر منجر به کاهش غیرقابل‌چشم‌پوشی در انحراف معیار استاندارد در مقادیر تابع هدف مساله شده است. در این پژوهش الگوریتم ترکیبی پیشنهادی روی چند مجموعه داده واقعی و شناخته شده اجرا شد. نتایج عددی بسیار مطلوب هستند. به‌عبارت‌دیگر، الگوریتم پیشنهادی در اکثر اجراها به‌سرعت با جواب بهینه سراسری همگرا می‌شود که مقادیر کم انحراف معیار استاندارد مؤید این ادعا است. نتایج مقایسه این الگوریتم با الگوریتم‌های پرکاربرد درزمینه[۶] خوشه‌بندی، برتری این الگوریتم را در حل مساله و در رقابت با سایر روش‌ها تأیید می‌کند. همچنین نتایج عددی نشان می‌دهد این الگوریتم برای حل مسائل با بعد بالا نیز بسیار مناسب است. لذا این پژوهش یک روش کاربردی جدید برای داده‌کاوی با بهره‌وری بالا در اختیار مدیران و تصمیم‌گیرندگان قرار می‌دهد. استفاده از این الگو مدیران را قادر می‌سازد تا با رعایت تمام محدودیت‌های ارادی و اجباری سازمان‌ها و با صرف انرژی و هزینه اندک به دسته‌بندی مفید و هدفمند اطلاعات دسترسی داشته باشند. احاطه بر اطلاعات و سادگی کاربر روی داده‌های گذشته سازمان و مدل‌سازی داده‌های آتی در پیش‌بینی شرایط آینده و مدیریت هوشمندانه با بهره‌وری بالا را به دنبال خواهد داشت.

ترکیب الگوریتم نهنگ با سایر الگوریتم‌ها برای حل مسائل کاربردی داده‌کاوی در حوزه مدیریت و استفاده از الگوریتم پیشنهادی این پژوهش در مطالعه موردی سازمان‌ها و ادارات داخلی در پژوهش‌های آتی پیشنهاد می‌گردد.

تعارض منافع

نویسندگان هیچ‌گونه تعارض منافعی برای اعلام ندارند.

References

Armando, G., & Farmani, M. R. (2014). Clustering analysis with combination of artificial bee colony algorithm and k-means technique. International Journal of Computer Theory and Engineering, (6)2, 141-145.

Sander, J. (2003). Coursteme homepage for principles of knowledge discovery in data. Available: http://www.cs.ualberta.ca/-joerg

Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: a review. ACM Compute. (31)3, 264-323.

Rokach, L., & maimon, O. (2005). Clutering methods. Maimon, Data mining and Knowledg Discovery Handbooks, Springer, New York, 1-432.

Niknam, T., Amiri, B., Olamaie, J., & Arefi, A. (2009). An efficient hybrid evolutionary optimization algorithm based on PSO and SA for clustering. Journal of Zhejiang University Science, 10(4), 512-519.

Ahmadyfard, A., & Modaress, H. (2008). Combining PSO and K-means to enhance data clustering. In: International symposium on telecommunications, 688-691.

Sandeep, U. M., & Pankaj, G. G. (2014). Hybrid particle swarm optimization (HPSO) for data clustering. International Journal of Computer application, 97(19), 1-15.

Karthikeyan, S., & Christopher, T. (2014). A Hybrid Clustering approach using Artificial Bee Colony (ABC) and particle swarm optimization. International Journal of Computer Applications,100(15), 1-6.

Mirjalili, S., & Lewi, A. (2016). The whale optimization algorithm. Advances in Engineering Software, 95, 51-67.

Nasiri, J., & Khiyabani, F. M. (2018). A whale optimization algorithm (WOA) approach for clustering. Cogent Mathematics & Statistics, 5(1).

Babalik, A., Cevahir, C. A., & Servet, K. M. (2017). A modification of tree-seed algorithm using Deb's rules for constrained optimization. Applications Soft Computing, 63(3), 289-305.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., & Wu, A.Y. (2002). An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions Pattern Analysis and Machine Intelligence, 24(7), 881-892.

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/

Karaboga, D., & Ozturk, C. (2009). A novel clustering approach: artificial bee colony (ABC) algorithm. Applications Soft Computing, 11(1), 652–657.

Van der Merve, D.W., & Engelhrecht, A.P. (2003). Data clustering using particle swarm optimization. Conference of evolutionary computation CEC'03, 215-220.

Mualik, U., & Bandyopadhyay, S. (2002). Genetic algorithm-based clustering technique. Pattern Recognition, 33, 1455-1465.