

استفاده از تکنیک خوشه‌بندی سیستم کلونی مورچگان بهبودیافته با هدف خوشه‌بندی داده‌های زلزله ایران

بهروز مینائی¹، محمد فتحیان²، احمدرضا جعفریان مقدم^{3*} و مهدی نصیری⁴

¹ دانشیار دانشکده مهندسی کامپیوتر - دانشگاه علم و صنعت ایران

² دانشیار دانشکده مهندسی صنایع - دانشگاه علم و صنعت ایران

³ دانشجوی دکترای دانشکده مهندسی صنایع دانشگاه علم و صنعت ایران و مدیر پروژه توسعه نرم‌افزار شرکت مهندسی شبکه پویا داده نوین

⁴ دانشجوی دکترای مهندسی کامپیوتر دانشگاه علم و صنعت ایران

(تاریخ دریافت 90/3/7، تاریخ دریافت روایت اصلاح‌شده 90/4/25، تاریخ تصویب 90/6/12)

چکیده

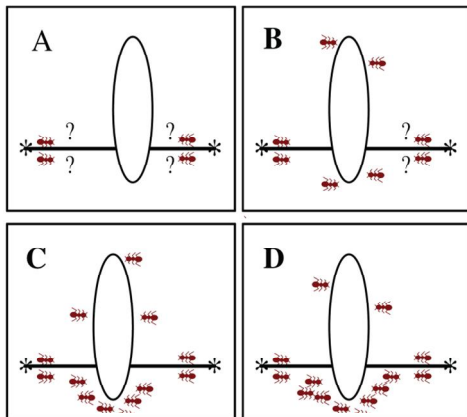
تکنیک خوشه‌بندی از مهم‌ترین تکنیک‌های داده‌کاوی و شاخه‌ای از تحلیل آماری چند متغیره بوده و روشی برای گروه‌بندی داده‌های مشابه در خوشه‌های یکسان است. با بزرگ‌تر شدن بانک‌های داده‌ای، تلاش محققان برای یافتن روش‌های خوشه‌بندی کارا و مؤثر متمرکز شده است تا از این راه بتوانند زمینه تصمیم‌گیری سریع و منطبق با واقعیت را فراهم آورند. بدین منظور، در این مقاله تکنیک خوشه‌بندی بهبودیافته سیستم کلونی مورچگان (IASC) با هدف ارائه یک الگوریتم خوشه‌بندی سریع و با دقت بالا پیشنهاد شده است. نتایج حاصل از اجرای الگوریتم روی داده‌های زلزله ایران، نشان از دقت و سرعت الگوریتم و کاهش زمان اجرا دارد. همچنین الگوریتم پیشنهادی قادر است داده‌های پرت را شناسایی کند.

واژه‌های کلیدی: تحلیل خوشه‌بندی، سیستم کلونی مورچگان، الگوریتم‌های فراابتکاری، زلزله

مقدمه

تکنیک خوشه‌بندی از مهم‌ترین تکنیک‌های داده‌کاوی است که امروزه اهمیت آن در دنیای واقعی بر کسی پوشیده نیست. با بزرگ‌تر شدن بانک‌های داده‌ای، تلاش محققان برای یافتن روش‌های خوشه‌بندی کارا و مؤثر متمرکز شده است تا از این راه بتوانند زمینه تصمیم‌گیری سریع و منطبق با واقعیت را فراهم آورند. تحلیل خوشه‌بندی، شاخه‌ای از تحلیل آماری چند متغیره بوده و روشی برای گروه‌بندی داده‌های مشابه در خوشه‌های یکسان است [1 و 2]. تکنیک‌های خوشه‌بندی سعی دارند با کشف روابط موجود در بین داده‌های جدید، روش خوشه‌بندی خود را بهبود بخشند. از این رو تکنیک‌های خوشه‌بندی به تکنیک‌های یادگیرنده نیز شهرت یافته‌اند [3 و 4]، به نحوی که قادرند پس از تعیین خوشه داده‌های مختلف، خوشه داده جدید را که به مجموعه اضافه می‌شود با صرف کمترین زمان مشخص کنند [5]. انواع تکنیک خوشه‌بندی عبارتند از: خوشه‌بندی سلسله مراتبی [4 و 6]، خوشه‌بندی مختلط [7 و 8]، خوشه‌بندی شبکه یادگیرنده [9-12]، خوشه‌بندی بر اساس تابع هدف و خوشه‌بندی افزایشی [13] و [14]. سیستم کلونی مورچگان (ACS²) که اولین بار در [15] بر اساس نتایج آزمایشات ارائه‌شده در [16] و [17] ارائه شد، برای حل مسائل گسسته و از جمله مسئله TSP³، کاربرد وسیع یافت. امروز دامنه کاربرد این الگوریتم گسترش یافته است و به تازگی برای حل مسائل مختلف از جمله خوشه‌بندی داده‌ها مورد استفاده قرار گرفته است. هر چند کارهای انجام‌شده برای خوشه‌بندی داده‌ها با استفاده از ACS محدود است، اما نتایج ارائه‌شده گواهی بر قدرت و سرعت الگوریتم‌های مبتنی بر ACS است. در [18] الگوریتم خوشه‌بندی مبتنی بر ACS پیشنهاد داده‌اند که در آن از استراتژی مورچه‌های مطلوب استفاده شده است. در این مقاله، از الگوریتم SA⁴ برای کاهش تعداد شهرهای بازدید شده توسط مورچه‌ها و از استراتژی انتخاب مسابقه⁵ برای یافتن بهترین مسیرها بهره گرفته شده تا الگوریتم سریع‌تری ارائه شود. در [19] برای خوشه‌بندی داده‌های ارائه‌شده، الگوریتمی است که در آن هر شهر (داده) بیانگر یک مورچه است و از طرفی

تکنیک خوشه‌بندی از مهم‌ترین تکنیک‌های داده‌کاوی است که امروزه اهمیت آن در دنیای واقعی بر کسی پوشیده نیست. با بزرگ‌تر شدن بانک‌های داده‌ای، تلاش محققان برای یافتن روش‌های خوشه‌بندی کارا و مؤثر متمرکز شده است تا از این راه بتوانند زمینه تصمیم‌گیری سریع و منطبق با واقعیت را فراهم آورند. تحلیل خوشه‌بندی، شاخه‌ای از تحلیل آماری چند متغیره بوده و روشی برای گروه‌بندی داده‌های مشابه در خوشه‌های یکسان است [1 و 2]. تکنیک‌های خوشه‌بندی سعی دارند با کشف روابط موجود در بین داده‌های جدید، روش خوشه‌بندی خود را بهبود بخشند. از این رو تکنیک‌های خوشه‌بندی به تکنیک‌های یادگیرنده نیز شهرت یافته‌اند [3 و 4]، به نحوی که قادرند پس از تعیین خوشه داده‌های مختلف، خوشه داده جدید را که به مجموعه اضافه می‌شود با صرف کمترین زمان مشخص کنند [5]. انواع تکنیک خوشه‌بندی عبارتند از: خوشه‌بندی سلسله مراتبی [4 و 6]، خوشه‌بندی مختلط [7 و 8]، خوشه‌بندی شبکه یادگیرنده [9-12]، خوشه‌بندی بر اساس تابع هدف و خوشه‌بندی افزایشی [13] و [14]. سیستم کلونی مورچگان (ACS²) که اولین بار در [15] بر اساس نتایج آزمایشات ارائه‌شده در [16] و [17] ارائه شد، برای حل مسائل گسسته و از جمله مسئله TSP³، کاربرد وسیع یافت. امروز دامنه کاربرد این الگوریتم گسترش یافته است و به تازگی برای حل مسائل مختلف از جمله خوشه‌بندی داده‌ها مورد استفاده قرار گرفته است. هر چند کارهای انجام‌شده برای خوشه‌بندی داده‌ها با استفاده از ACS محدود است، اما نتایج ارائه‌شده گواهی بر قدرت و سرعت الگوریتم‌های مبتنی بر ACS است. در [18] الگوریتم خوشه‌بندی مبتنی بر ACS پیشنهاد داده‌اند که در آن از استراتژی مورچه‌های مطلوب استفاده شده است. در این مقاله، از الگوریتم SA⁴ برای کاهش تعداد شهرهای بازدید شده توسط مورچه‌ها و از استراتژی انتخاب مسابقه⁵ برای یافتن بهترین مسیرها بهره گرفته شده تا الگوریتم سریع‌تری ارائه شود. در [19] برای خوشه‌بندی داده‌های ارائه‌شده، الگوریتمی است که در آن هر شهر (داده) بیانگر یک مورچه است و از طرفی



شکل 1: رفتار مورچه‌های واقعی

در الگوریتم مورچگان، هر مورچه بیانگر یک جواب⁷ از مسئله است. فرایند انتخاب یک مسیر توسط مورچه را قانون انتقال وضعیت⁸ می‌نامند. اگر مورچه k ام در شهر r باشد و بخواهد به شهر s انتقال یابد انتخاب شهر s با استفاده از این قانون به صورت معادله (1) خواهد بود:

$$s = \begin{cases} \arg \max_{u \in J_k(r)} \{ \tau(r, u) \cdot \eta(r, u)^\beta \}, & \text{if } q \leq q_0 \\ S, & \text{otherwise} \end{cases} \quad (1)$$

که در این معادله $\tau(r, u)$ بیانگر مقدار فرومون روی کمان ru بوده و $\eta(r, u)$ بیانگر معکوس فاصله بین دو نقطه r و u است. β و q_0 ($0 \leq q_0 \leq 1$) پارامترهای الگوریتم هستند. q یک مقدار تصادفی در بازه $0 \leq q \leq 1$ است و همچنین مجموعه $J_k(r)$ مجموعه شهرهای (نقاطی) است که توسط مورچه k ام بعد از عبور از نقطه r هنوز بازدید نشده‌اند.

در صورتی که مقدار تصادفی تولیدشده q کوچک‌تر و یا مساوی مقدار پارامتر q_0 باشد، عمل استخراج⁹ برای انتخاب یک شهر (نقطه) با استفاده از معادله (1) انجام می‌گیرد. این بدین معنی است که از بین نقاط موجود، یک نقطه با بیشترین میزان فرومون انتخاب خواهد شد.

در صورتی که q بزرگ‌تر از پارامتر q_0 باشد، در این صورت انتخاب نقطه بعدی برای انتقال به صورت احتمالی انجام خواهد شد. در واقع عمل اکتشاف¹⁰ نقطه جدید صورت می‌گیرد. این احتمال با استفاده از معادله (2) محاسبه می‌شود:

مورچه‌ها خصوصیات و ویژگی‌های متفاوتی دارند. در [20] الگوریتم خوشه‌بندی بر پایه سیستم مورچگان ارائه شده است و در [21] الگوریتم ارائه‌شده در [20] با الگوریتم خوشه‌بندی k میانگین ترکیب شده تا الگوریتم قوی‌تری ارائه شود. [22] الگوریتم‌های ارائه‌شده در [21] و [20] را مورد توجه قرار داده و یک الگوریتم دو مرحله‌ای را ارائه کرده است. در مرحله اول الگوریتم، داده‌ها با استفاده از الگوریتم خوشه‌بندی بر پایه سیستم مورچگان و الگوریتم k میانگین مورچگان، خوشه‌بندی شده و در مرحله دوم با استفاده از الگوریتم مبتنی بر ACS، قوانین انجمنی در هر یک از خوشه‌ها را استخراج کرده است. الگوریتم ارائه‌شده در این مقاله، بسیار شبیه به الگوریتم ارائه‌شده در مرحله اول [22] است. الگوریتم پیشنهادی، اصلاحاتی را روی الگوریتم ذکرشده اعمال کرده که منجر به بهبود عملکرد و نتایج الگوریتم شده است.

ادامه مقاله به این ترتیب بخش‌بندی شده است؛ بخش دوم مروری بر رفتار مورچگان و الگوریتم مبتنی بر این رفتار دارد. الگوریتم پیشنهادی در بخش سوم ارائه شده است و بخش چهارم به ارزیابی عملکرد الگوریتم پیشنهادی با اجرای آن روی داده‌های زلزله ایران اختصاص دارد. نتایج حاصل از اجرای الگوریتم در بخش پایانی مقاله مورد توجه قرار گرفته است.

1. مروری بر رفتار و الگوریتم مورچگان [23]

سیستم کلونی مورچگان (ACS) برگرفته از رفتار مورچگان برای یافتن غذا است. هنگامی که یک مورچه برای یافتن غذا جستجوی خود را آغاز می‌کند، ردپای خود را با استفاده از یک ماده شیمیایی به نام فرومون⁶ بر جای می‌گذارد. این ماده به مرور زمان تبخیر می‌شود. بنابراین مسیری که مورچه‌های زیادی از آن عبور کنند، مقدار فرومون بیشتری خواهد داشت و از طرفی مورچه‌ها مسیری را که فرومون بیشتری داشته باشد، با احتمال بیشتری انتخاب خواهند کرد. این فرایند در نهایت به انتخاب یک مسیر توسط همه مورچگان منجر خواهد شد. شکل (1) این فرایند را نشان داده است که در نهایت اغلب مورچه‌ها مسیر کوتاه‌تر را انتخاب کرده‌اند.

$E = \{O_1, \dots, O_n\}$ بیانگر مجموعه n داده از بانک داده

است که هر داده k ویژگی دارد.

α اهمیت نسبی میزان تبخیر فرمون (میزان فرمون

باقیمانده)، $\alpha \geq 0$

β اهمیت نسبی میزان فاصله بین دو نقطه (قابلیت دید)،

$\beta \geq 0$

ρ پارامتر کاهش فرمون، $0 < \rho < 1$

Q یک ثابت

n تعداد داده‌ها

m تعداد مورچه‌ها

nc تعداد خوشه‌ها

T مجموعه داده‌های ویزیت شده. حداکثر طول آرایه n, T

خواهد بود. $T = \{O_a, O_b, \dots, O_t\}$ و نقاط a, b, t نقاطی

هستند که توسط مورچه‌ها ویزیت شده است. این آرایه در

الگوریتم، حکم یک فهرست تابو¹³ را دارد که مانع از

انتخاب نقاط تکراری می‌شود.

T_m مجموعه T که توسط مورچه m تشکیل شده است.

$D_{\text{mean}}(T_m)$ متوسط فاصله بین همه داده‌های متعلق به

مورچه m با مرکز آن خوشه است.

$O_{\text{center}}(T)$ داده‌ای که بیانگر مرکز همه داده‌های مجموعه

T است و به صورت معادله (6) محاسبه می‌شود. N_T تعداد

داده‌ها در مجموعه T است.

$$O_{\text{center}}(T) = \frac{1}{n_T} \sum_{O_i \in T} O_i \quad (6)$$

TWCV مجموع واریانس درون خوشه‌ها که با استفاده

از معادله (7) محاسبه می‌شود:

$$\sum_{k=1}^{nc} \sum_{i \in k} (O_i, O_{\text{center}}(T_k))^2 \quad (7)$$

SC: ضریب سیلهوت¹⁴ بین دو خوشه است [24] که با

استفاده از معادلات (8) و (9) محاسبه می‌شود که در

آن a : متوسط فاصله بین داده i با سایر داده‌های درون یک

خوشه و b : متوسط فاصله بین داده I با سایر داده‌ها در

خوشه دیگر هستند:

$$p_k(r, s) = \begin{cases} \frac{\tau(r, s) \cdot \eta(r, s)^\beta}{\sum_{u \in J_k(r)} \tau(r, u) \cdot \eta(r, u)^\beta}, & \text{if } s \in J_k(r) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

در الگوریتم مورچگان، مقدار فرمون روی کمان‌ها در

دو موقعیت به هنگام می‌شود. با عبور هر مورچه از روی

کمان، مقدار فرمون آن کمان با استفاده از معادله (3) به

هنگام می‌شود که این تغییر را تغییر محلی¹¹ می‌نامند:

$$\tau(r, s) = (1 - \rho) \cdot \tau(r, s) + \rho \cdot \Delta\tau(r, s) \quad (3)$$

که در آن $0 < \rho < 1$ پارامتر تبخیر فرمون بوده و

$\Delta\tau(r, s) = \tau_0$ است که یک مقدار ثابت خواهد بود.

بعد از اتمام همه تورها توسط مورچه‌ها، فرمون روی همه

کمان‌ها به هنگام خواهد شد. این تغییر با استفاده از

معادله (4) محاسبه می‌شود که به آن، به هنگام کردن

عمومی¹² گویند:

$$\tau(r, s) = (1 - \alpha) \cdot \tau(r, s) + \alpha \cdot \Delta\tau(r, s) \quad (4)$$

به طوری که

$$\Delta\tau(r, s) = \begin{cases} \frac{1}{L_{gb}}, & \text{if } (r, s) \in \text{global-best-tour} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

که در آن $0 < \alpha < 1$ پارامتر میزان کاهش فرمون

است و L_{gb} کوتاه‌ترین مسیر از نقطه آغازین تا نقطه فعلی

است. بنابراین با افزایش میزان فرمون روی کمان‌هایی که

توسط مورچه‌های بیشتری بازدید شده است، احتمال

انتخاب آن مسیر توسط سایر مورچه‌ها را افزایش خواهد

داد.

2. الگوریتم پیشنهادی

همان طور که اشاره شد، الگوریتم پیشنهادی در

این مقاله شبیه الگوریتم ارائه شده در [22] است. الگوریتم

[22] قادر به شناسایی داده‌های پرت نبوده و همچنین در

ایجاد و ادغام خوشه‌ها اشکلات اساسی دارد. الگوریتم

پیشنهادی در این بخش سعی در رفع این نواقص دارد. در

الگوریتم پیشنهادی IASC متغیرها و پارامترهای زیر مورد

توجه است:

مجموعه T_{mi} و T_{mj} محاسبه می‌شود. در صورتی که این ضریب مقداری کمتر از α داشته باشد، آنگاه این دو خوشه ادغام می‌شوند. بعد از اجرای رویه *Agglomerate*، رویه *Agglomerate_obj* بار دیگر اجرا خواهد شد. سپس رویه *Remove* داده‌ها غیر مشابه در خوشه‌ها را حذف خواهد کرد. اگر فاصله بین یک داده از مجموعه T_m بیشتر از $D_{mean}(T_m)$ باشد، آن داده از مجموعه T_m جدا خواهد شد.

بعد از اجرای رویه‌های الگوریتم مقدار *TWCV* درون خوشه‌ها محاسبه می‌شود. اگر مقدار *TWCV* تغییر نکند، الگوریتم خاتمه خواهد یافت، در غیر این صورت مراحل *Divide*، *Agglomerate_obj*، *Agglomerate* تا جایی تکرار خواهد شد تا مقدار *TWCV* بدون تغییر بماند. در پایان الگوریتم داده‌هایی که درون خوشه‌ای قرار نگرفته‌اند، به عنوان داده‌های پرت ارائه خواهند شد و همچنین هر خوشه و داده‌های آنها با توجه به نقاط ویزیت‌شده توسط هر مورچه m تعیین می‌شود.

3. اجرای الگوریتم پیشنهادی

در این بخش برای بررسی عملکرد الگوریتم پیشنهادی، نمونه‌ای واقعی از داده‌های زلزله در ایران مورد توجه قرار گرفته است. این داده‌ها از مرجع [25] جمع‌آوری شده است و 514 داده دارد. این داده‌ها 7 ویژگی شامل سال وقوع زلزله، شدت زلزله بر اساس مقیاس‌های امواج درونی (M_b)، امواج سطحی (M_s)، گشتاوری (M_w) و مقیاس محلی (M_l) و همچنین طول و عرض موقعیت جغرافیایی زمین لرزه است. الگوریتم پیشنهادی در نرم‌افزار *MATLAB* پیاده‌سازی و تحت سیستم با مشخصات *Intel® Core™2 Duo 2.93Ghz* و *2 GB RAM* اجرا شد.

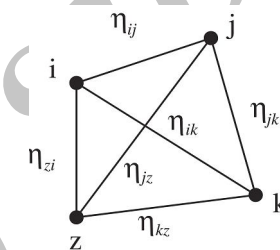
جدول (1) نتایج اجرای الگوریتم ارائه‌شده در [22] با عنوان *ASCA* و نتایج اجرای الگوریتم پیشنهادی (*IASC*) را روی بانک داده ذکر شده با تعداد مورچه‌های مختلف و $\rho = 0.2$ نشان داده است. الگوریتم داده‌های پرت را به خوبی و به درستی تعیین کرده است. نتایج ارائه‌شده در جدول (1) نشان از دقت و قدرت الگوریتم پیشنهادی دارد. طبق نتایج این جدول، تعداد داده‌های پرت در همه اجراها برای الگوریتم پیشنهادی یکسان است که این موضوع

$$\text{if } (a < b) : SC = \frac{a}{b} - 1 \quad (8)$$

$$\text{if } (a > b) : SC = 1 - \frac{a}{b} \quad (9)$$

ضریب *SC* در بازه $[0, 1]$ قرار می‌گیرد. هر چه مقدار *SC* به عدد 1 نزدیک‌تر باشد، بیانگر تفکیک بیشتر بین خوشه‌ها است.

در این الگوریتم هر یک از نقاط بانک داده به صورت یک نقطه در فضا فرض می‌شود که توسط کمان‌هایی به یکدیگر متصل می‌شوند. شکل (2) این موضوع را نشان داده است. اعداد روی کمان‌ها می‌تواند میزان فراوانی توالی آیت‌م¹⁵ ijz در بانک داده و یا یک مقدار فرمون باشد.



شکل 2: شبکه‌ای از داده‌ها

الگوریتم *IASC* چهار رویه *Divide*، *Agglomerate_obj*، *Agglomerate* و *Remove* دارد. در الگوریتم (1) مراحل اجرای هر یک نشان داده شده است. در ابتدا همه پارامترهای الگوریتم تعریف شده و نیز همه داده‌ها در یک خوشه قرار می‌گیرند. سپس رویه *Divide* بر اساس میزان فرمون و برخی معیارهای دیگر، داده‌ها را در تعدادی خوشه تقسیم می‌کند. در این مرحله برخی از داده‌ها ممکن است به خوشه‌ای تعلق نگیرند. بعد از رویه *Divide*، رویه *Agglomerate_obj* داده‌ها را در خوشه‌های مناسب قرار می‌دهد. در الگوریتم (1) مراحل این رویه نشان داده شده است. در این رویه فاصله بین یک داده با مرکز مجموعه T_m محاسبه می‌شود. اگر این فاصله کمتر از $D_{mean}(T_m)$ باشد، داده در مجموعه C قرار می‌گیرد. در صورتی که مجموعه C بیش از دو عضو داشته باشد، عضوی که کمترین فاصله با مرکز T_m را دارد، به مجموعه T_m افزوده می‌شود.

سپس رویه *Agglomerate* اجرا می‌شود. در این رویه خوشه‌هایی که بیشترین شباهت را با یکدیگر داشته باشند، با هم ادغام می‌کند. در این ضریب *SC* بین دو

با آنکه ASCA تعداد خوشه‌های بسیار زیادی را ارائه کرده است، اما TWCV نیز برای اجراهای مختلف مقداری بیشتر نسبت به الگوریتم پیشنهادی ارائه کرده است. از آنجا که الگوریتم پیشنهادی تعداد خوشه‌های بسیار کمتری نسبت به ASCA ارائه کرده و از طرفی با توجه به اینکه مقدار TWCV در الگوریتم پیشنهادی در اغلب اجراها بسیار کمتر از ASCA است، به این نتیجه خواهیم رسید که الگوریتم پیشنهادی از دقت و درستی بسیار بالاتری نسبت به ASCA بهره‌مند است.

بیانگر این مطلب است که الگوریتم داده‌های پرت را به خوبی و به درستی تعیین کرده است.

هر چند زمان اجرای الگوریتم پیشنهادی بیشتر از زمان اجرای الگوریتم ASCA است، اما الگوریتم پیشنهادی تعداد خوشه‌های به مراتب کمتری نسبت به ASCA ارائه کرده است. نتایج نشان می‌دهد که ASCA نسبت به تعداد مورچه‌ها حساس است، اما الگوریتم پیشنهادی حساسیت بسیار کمتری نسبت به پارامتر تعداد مورچه‌ها دارد.

الگوریتم 1: الگوریتم IASC و رویه‌های آن

Procedure IASC algorithm

Initialize the parameters
Group all objects as a cluster
Do
 Divide for all ant m.
 Agglomerate_obj for all ant m.
 Agglomerate for all ant m.
 Agglomerate_obj for all ant m.
 Remove for all ant m.
 Determine the non-clustered objects as outlier objects
 Calculating TWCV
While (TWCV is not change)
 Determine clusters by visited objects by each ant m.

Procedure Divide

Lay pheromone on the path by η_{ij} for all i and j, $i \neq j$

Calculating $\bar{\tau}$.

Updating pheromone by

$$\tau_{ij} = (1 - \rho)\tau_{ij} + \Delta\tau_{ij} \text{ Where } \Delta\tau_{ij} = \begin{cases} \frac{1}{d_{ij}} & \text{if } \tau_{ij} > \bar{\tau} \\ 0 & \end{cases}$$

Calculating $\bar{\tau}_i$ for all $i=1, 2, \dots, n$ and Sort them.

Select m object from sorted i.

Each ant m starts from one of m selected object.

Each ant m collects object j if $\tau_{ij} \geq \bar{\tau}$.

Procedure Agglomerate_obj

Let $C = \phi$.

If O_j satisfied with the following equation:

$$D(O_{center}(T_m), O_j) < D_{mean}(T_m)$$

where $j \in \{n - T_m \mid m = 1, \dots, m\}$

Add O_j to C

If $(C < 2)$ Assign O_j to T_m

Else Assign O_j to T_m if the distance of $O_{center}(T_m)$ and O_j is minimum.

Procedure Agglomerate

Let $C = \phi$.

Determine parameter α ; $0 \leq \alpha \leq 0.5$.

If T_m satisfied with the following equation:

$$SC(T_{m_i}, T_{m_j}) \leq \alpha$$

Agglomerate T_{m_i} and T_{m_j} as a cluster.

Procedure Remove

Remove object j from Ant m if $D(O_{center}(T_m), O_j) > D_{mean}(T_m)$.

where $j \in T_m$

4. نتیجه گیری

در این مقاله، الگوریتم خوشه‌بندی بهبودیافته سیستم کلونی مورچگان (IASC) برای ارائه یک الگوریتم خوشه‌بندی سریع و با دقت بالا پیشنهاد شد. الگوریتم پیشنهادی قادر است داده‌های پرت را شناسایی کرده و برای داشتن تحلیلی بهتر و منطبق با واقعیات، آنها را از

انجام تحلیل خارج کند. همچنین این الگوریتم حساسیت بسیار کمی نسبت به پارامترهای خود از جمله تعداد مورچه‌ها دارد. در این مقاله الگوریتم پیشنهادی روی داده‌های زلزله ایران پیاده‌سازی و اجرا شد. نتایج حاصل از اجرای الگوریتم نشان از دقت بالا و سرعت اجرای الگوریتم دارد.

جدول 1: نتایج اجرای الگوریتم‌های ASCA و الگوریتم پیشنهادی IASC

	IAS C	ASC A	IAS C	ASC A	IAS C	ASC A	IAS C	ASC A	IAS C	ASC A	IAS C	ASC A
تعداد مورچه	20		30		40		50		60		70	
تعداد خوشه	6	20	11	30	13	40	17	47	19	56	21	64
زمان اجرا (s)	18	5/4	21/6	7/7	34/15	10/26	29/92	12/12	26/71	14/25	38/11	16/32
TWCV	47/4	81/55	31/8	71/51	32/33	59/08	32/57	53/4	32/83	45/09	44/89	40/55
داده‌های پرت	6	-	6	-	6	-	6	-	6	-	6	-

مراجع

- 1 - Kamber, H. (2006). *Data Mining: Concepts and Techniques*. 2th. Ed. Elsevier.
- 2 - Aldenderfer, M.S. and Blashfield, R.K. (1986). *Cluster Analysis*. Newbury Park: Sage Publications.
- 3 - Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.
- 4 - Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- 5 - Jain, A. K., Duin, R. P. W. and Mao, J. (2000). "Statistical pattern recognition: A review." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, PP. 4–37.
- 6 - Hartigan, J.A. (1975). *Clustering Algorithms*. Wiley, New York.
- 7 - McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.
- 8 - McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.
- 9 - Grossberg, S. (1976). "Adaptive pattern classification and universal recoding I: Parallel development and coding of neural feature detectors." *Biological Cybernetics*, Vol. 23, PP. 121–134.
- 10 - Lippmann, R. P. (1987). "An introduction to computing with neural nets." *IEEE Transactions on Acoustics, Speech, Signal Processing*, PP. 4–22.
- 11 - Tsao, E. C. K., Bezdek, J. C. and Pal, N. R. (1994). "Fuzzy Kohonen clustering networks." *Pattern Recognition*, Vol. 27, PP. 757–764.
- 12 - Kohonen, T. (2001). *Self-Organizing Maps*, 3th. Ed. Springer-Verlag, Berlin.
- 13 - Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithm*. Plenum Press, New York.
- 14 - Yang, M. S. (1993). "A survey of fuzzy clustering." *Mathematical and Computer Modelling*, Vol. 18, PP. 1–16.

- 15 - Colorni, A., Dorigo M. and Maniezzo, V. (1991). "Distributed Optimization by Ant Colonies." *actes de la première conférence européenne sur la vie artificielle*, Paris, France, Elsevier Publishing, PP. 134-142.
- 16 - Goss, S., Aron, S., Deneubourg, J. L. and Pasteels, J. M. (1989). "The self-organized exploratory pattern of the Argentine ant." *Naturwissenschaften*, Vol. 76, PP. 579-581.
- 17 - Deneubourg, J. L., Aron, S., Goss, S. and Pasteels, J. M. (1990). "The self-organizing exploratory pattern of the Argentine ant." *Journal of Insect Behavior*, Vol. 3, PP. 159.
- 18 - Tsai, C. F., Wu, H. C. and Tsai, C. W. (2002). "A new clustering approach for data mining in large databases." *In Proceedings of the international symposium on parallel architectures, algorithms and networks (ISPAN'02)*, IEEE Computer Society, PP. 1087-4089.
- 19 - Yang, X. B., Sun, J. G. and Huang, D. (2002). "A new clustering method based on ant colony algorithm." *In Proceedings of the 4th world congress on intelligent control and automation*, PP. 2222-2226.
- 20 - Kuo, R. J., Wang, H. S., Hu, T. L. and Chou, S. H. (2005), "Application of Ant K-Means on Clustering Analysis." *Computers and Mathematics with Applications*, Vol. 50, PP. 1709-1724.
- 21 - Kuo, R. J. and Shih, C. W. (2007). "Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan." *Computers and Mathematics with Applications*, Vol. 54, PP. 1303-1318.
- 22 - Kuo, R. J., Lin S. Y. and Shih, C. W. (2007). "Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan." *Expert Systems with Applications*, Vol. 33, PP. 794-808.
- 23 - Talbi, El. (2009). *Metaheuristics: from design to implementation*, John Wiley & Sons, Inc., Hoboken, New Jersey.
- 24 - Tan, P. N., Steinbach, M. and Kumar, V. (2005). *Introduction to Data Mining*, Addison-Wesley.
- 25 - websites: www.ngdir.ir, www.bhrc.ac.ir & www.geophysics.ut.ac.ir.

واژه‌های انگلیسی به ترتیب استفاده در متن

- 1- Improved Ant System-based Clustering algorithm
- 2- Ant Colony System
- 3- Traveling Salesman Problem
- 4- Simulated Annealing
- 5- Tournament selection strategy
- 6- Pheromone
- 7- Solution
- 8- State Transition Rule
- 9- Exploitation
- 10- Exploration
- 11- Local Updating
- 12- Global Updating
- 13- Tabu List
- 14- Silhouette Coefficient
- 15- Item