تعیین تعداد گروه در مجموعه داده‌های ژئوشیمیایی با استفاده از شاخص‌های بازشناسی الگوی مبتنی بر تفکیک و تراکم خوشه‌ها

سعید اسمعیل اوغلی[*1]، سیدحسن طباطبایی[2]، هوشنگ اسدی هارونی[3]

۱- دانشجوی دکتری، دانشکده مهندسی معدن، دانشگاه صنعتی اصفهان

۲- دانشیار، دانشکده مهندسی معدن، دانشگاه صنعتی اصفهان

۳- استادیار، دانشکده مهندسی معدن، دانشگاه صنعتی اصفهان

(دریافت: مهر ۱۳۹۵، پذیرش: دی ۱۳۹۷)

چکیده

تقسیم‌بندی مجموعه داده به زیرمجموعه‌های همگن، هدفی اساسی در تحلیل داده‌های ژئوشیمیایی است که اغلب از ابزار خوشه‌بندی برای نیل به آن استفاده می‌شود. مهم‌ترین چالش عملی موجود در این راستا، تخمین تعداد حقیقی گروه‌های نهان در مجموعه داده است که به طور سنتی از اطلاعات ژئوشیمیایی توصیفی، دانش کارشناسی یا به کارگیری یک شاخص آماری خاص برای حل آن استفاده می‌شود. خروجی این روش‌ها اغلب ناپایدار و همراه با عدم‌قطعیت است، لذا رویکردی که این مقاله برای حل مسئله تعیین تعداد خوشه در داده‌ها پیشنهاد می‌کند، اجرای گستره‌ای از شاخص‌های موجود و تولید توزیعی از پاسخ‌های ممکن و نهایتاً استخراج جواب نهایی از آن است. شاخص‌های به کار رفته در این زمینه، مبتنی بر روابط بازشناسی الگو و بر مبنای بیشینه‌سازی پارامتر تفکیک بین گروهی و کمینه‌سازی پارامتر تراکم درون گروهی هستند. جهت آزمون رویکرد پیشنهادی، مجموعه داده شبیه‌سازی شده دوبعدی با چهار خوشه مصنوعی تولید گشته و با اجرای ۳۰ شاخص پرکاربرد بر روی آن، بالاترین فرکانس موجود در توزیع پاسخ‌ها منطبق بر جواب حقیقی مسئله به دست آمده است. این راهکار عیناً بر روی یک مجموعه داده ژئوشیمیایی حقیقی و چندمتغیره، شامل داده‌های خاک کانسار مس- طلای دالی شمالی واقع در استان مرکزی اجرا شده است که نتایج به دست آمده نشان دهنده معنی‌دار بودن و انطباق پاسخ نهایی با فرآیندهای زمین‌شناسی و کانه‌زایی محدوده است.

کلید واژه‌ها

داده‌های ژئوشیمیایی، خوشه‌بندی، تعداد گروه، تفکیک خوشه‌ها، تراکم خوشه‌ها، کانسار دالی شمالی

[*] عهده‌دار مکاتبات: s.esmaeiloghli@mi.iut.ac.ir

۱- مقدمه

در حین انجام مطالعات ژئوشیمی اکتشافی، مجموعه داده‌های حجیمی گردآوری می‌شود که حاوی مشاهدات جزئی از متغیرهای گوناگون است. بسته به ماهیت مطالعه، این داده‌ها می‌توانند از نمونه‌های خاک و سنگ (اکتشافات سطحی) یا مغزه‌های حفاری (اکتشافات زیرسطحی) حاصل شده باشند. همچنین ممکن است داده‌ها دارای ماهیت عددی (عیار عنصر) یا اسمی (جنس سنگ یا زون آلتراسیون) باشند. تمامی انواع ذکر شده، نهایتاً منجر به تولید مجموعه داده‌ای چندبعدی می‌شوند که تجزیه و طبقه‌بندی آن برای هوش انسانی کاری بسیار دشوار است[۱، ۲]. این در حالی است که ژئوشیمیست اکتشافی، نیازمند استخراج اطلاعات مفید و طبقه‌بندی شده از مجموعه داده خام است. این نوع تخلیص داده‌ها با اهداف گوناگونی چون کشف الگوی وابستگی متغیرها و نمونه‌ها با یکدیگر، شناسایی ارتباط ژنتیکی عناصر ژئوشیمیایی، تفکیک فرآیندهای آلتراسیونی و فازهای کانه‌زایی و..... صورت می‌پذیرد. برای این هدف، الگوریتم‌ها و ابزار ریاضیاتی متنوعی در حوزه یادگیری ماشین و داده‌کاوی پیش‌بینی شده است، که مفیدترین، سریع‌ترین و کم‌هزینه‌ترین آنها، روش‌های خوشه‌بندی هستند. فرآیند خوشه‌بندی که با اسامی دیگری چون رده‌بندی عددی و طبقه‌بندی خودکار نیز شناخته می‌شود، شامل تقسیم‌بندی مجموعه‌ای از داده‌ها به گروه‌ها یا خوشه‌هایی است، به نحوی که اعضای درون هر خوشه بیش‌ترین تشابه را با یکدیگر داشته باشند و بین اعضای خوشه‌های مختلف نیز بیش‌ترین تباین وجود داشته باشد. این فرآیند از دیدگاه تحلیل سیستم، نگاشتی از فضای هتروژن داده‌ها به فضای هموژن خوشه‌ها است، که خروجی این سیستم می‌تواند در جداسازی و تفسیر فرآیندهای ژئوشیمیایی نقش بسیار مهمی ایفا نماید.

روش‌های خوشه‌بندی علی‌رغم کاربردهای مفید در مطالعات ژئوشیمیایی، با مشکلاتی از لحاظ اجرایی همراه‌اند که نتایج به دست آمده از آنها را با ابهام مواجه می‌سازد. بهره‌برداری صحیح و بهینه از روش‌های خوشه‌بندی، مستلزم آگاهی از ساختار نهان و تعداد گروه‌های موجود در مجموعه داده است. اغلب الگوریتم‌های خوشه‌بندی فرض معلوم بودن تعداد خوشه‌ها را داشته و مقدار آن را به عنوان ورودی دریافت می‌کنند[۳]. بنابراین، کاربرانی که مایل به استخراج دسته‌بندی‌های مختلف نهفته در داده‌ها هستند، ابتدا باید تعداد گروه‌های موجود در مجموعه داده را محاسبه نمایند. در ادبیات موضوع، بحث فراوانی بر مسئله تعیین تعداد گروه در داده‌ها شده و طیف وسیعی از راهکارهای مبتنی بر روش‌های هوشمند، فراابتکاری، بازشناسی الگو، فرآیندهای تصادفی و..... برای حل آن ارائه شده است. در این میان، بیش‌ترین اقبال به سمت شاخص‌های مبتنی بر بازشناسی الگو بوده است. دان[۱]، شاخصی بر اساس فاصله بین خوشه‌ها و قطر هر خوشه معرفی کرده است[۴]. میلیگان و کوپر[۲]، با تولید یک مجموعه داده شبیه‌سازی شده با تعداد خوشه معلوم، کارآیی ۳۰ شاخص مختلف را مورد ارزیابی قرار داده‌اند[۵]. روسیو[۳]، ضریب سیلووت[۴] [۶] و تیبشیرانی و همکاران[۵] نیز آماره گپ[۶] [۷] را بدین منظور ابداع نموده‌اند. عملکرد تمام شاخص‌های بازشناسی الگو بر دو پارامتر اساسی استوار است:

- تفکیک خوشه‌ها[۷]، که معیاری از جدایش و تفریق گروه داده‌ها از یکدیگر بوده و تأمین کننده بیشینه واریانس بین خوشه‌ها است؛
- تراکم خوشه‌ها[۸]، که معیاری از فشردگی اعضای هر گروه بوده و تأمین کننده بیشینه کوواریانس درون خوشه‌ها است.

انتخاب نادرست تعداد گروه در مجموعه داده‌های ژئوشیمیایی، نتایج حاصل از خوشه‌بندی داده‌ها و تفسیر آنها را کاملاً تحت تأثیر قرار می‌دهد. این در حالی است که در اکثر قریب به اتفاق مطالعات ژئوشیمیایی، معیار صحیحی برای انتخاب تعداد بهینه و صحیح خوشه‌ها در نظر گرفته نشده و یا این که صرفاً بر اساس نتایج یک شاخص خاص اقدام به تعیین تعداد گروه شده است. در این میان، زارع مطلق و همکاران[۹] [۸] از تعدادی شاخص محدود برای ارزیابی کیفیت نتایج دسته‌بندی در آنالیز خوشه‌ای مبتنی بر گراف بهره برده‌اند. اما با توجه به این که شاخص‌های مختلف، بسته به ساختار موجود در ماتریس داده، پاسخ‌های متفاوتی ارائه می‌دهند، لذا استفاده از یک یا چند شاخص خاص نمی‌تواند رهیافت مؤثری در زمینه محاسبه تعداد مناسب گروه در مجموعه داده ژئوشیمیایی ارائه کند. این مقاله در نظر دارد تا با اجرای طیفی از شاخص‌های مطرح شده در حوزه بازشناسی الگو، ضمن

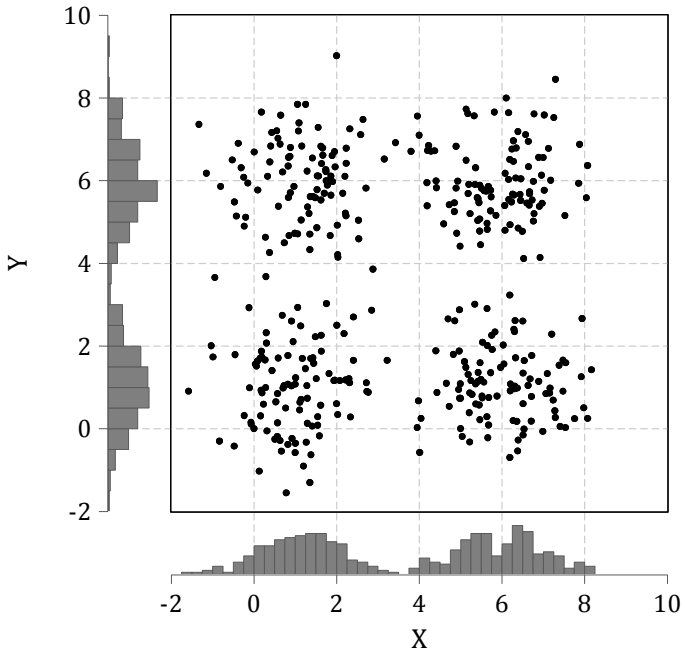
ارائه توزیعی از جواب‌های ممکن، بهترین حالت ممکن را بر مبنای توزیع فرکانسی پاسخ‌های تولید شده انتخاب نماید.

۲- مواد و روش‌ها

تعداد ۳۰ شاخص برای محاسبه تعداد گروه در مجموعه داده مورد استفاده قرار می‌گیرد. این روش‌ها بر روی دو مجموعه داده شبیه‌سازی شده و واقعی اجرا شده و جواب بهینه به دست آمده مورد بحث قرار می‌گیرد.

۲-۱- مجموعه داده شبیه‌سازی شده

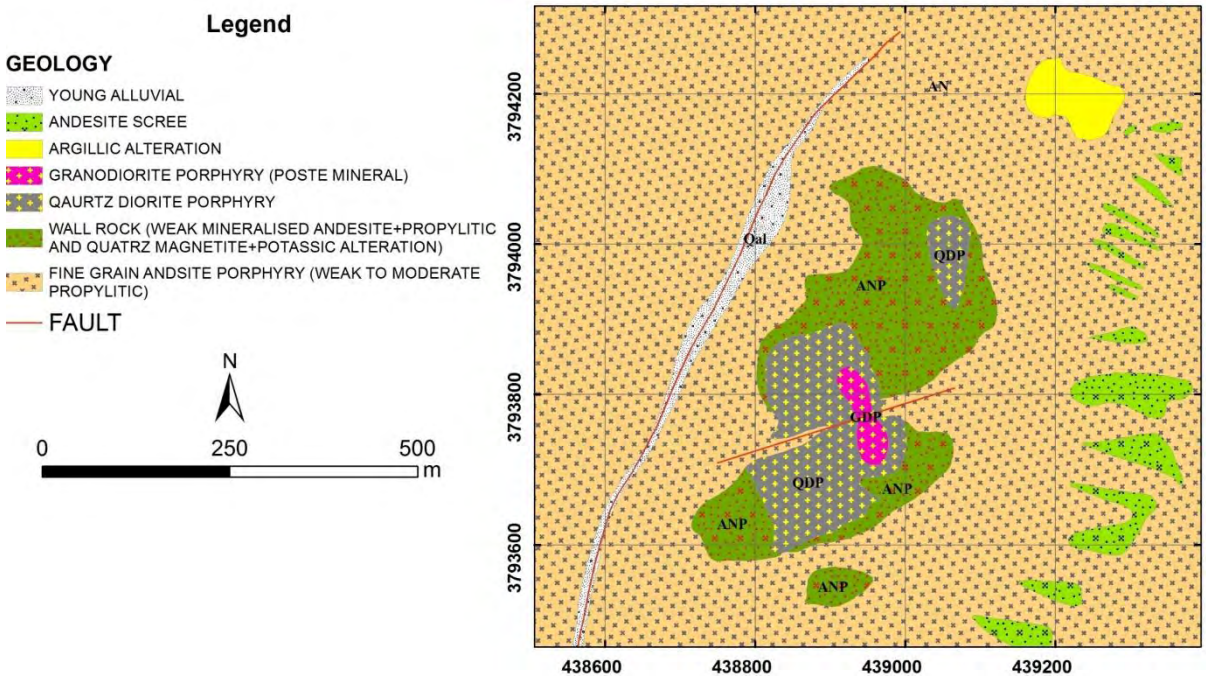
برای بررسی قابلیت شاخص‌ها در تشخیص صحیح تعداد گروه در داده‌ها، یک مجموعه داده شبیه‌سازی شده متشکل از ۴ خوشه مصنوعی مجزا و فاقد همپوشانی در محیط MathWorks MATLAB ساخته شده است. این مجموعه داده دارای ۴۰۰ نقطه داده دوبعدی (هر خوشه شامل ۱۰۰ داده با ویژگی‌های X و Y) است. خوشه‌ها تحت شرایط توزیع گوسی با مرکزیت جرم نقاط (۱،۱)، (۱،۶)، (۶،۱) و (۶،۶) و انحراف استاندارد واحد تولید شده و در فضای اقلیدسی دو بعدی پراکنده شده‌اند (شکل ۱).



شکل ۱: نمودار پراکندگی و توزیع فرکانسی مجموعه داده‌ی شبیه‌سازی شده در فضای دوبعدی

۲-۲- مجموعه داده ژئوشیمیایی

مجموعه داده واقعی به کار رفته در این پژوهش، شامل ۱۴۹ نمونه خاک برداشت شده از محدوده کانسار مس-طلای دالی شمالی در استان مرکزی است. محدوده دالی از جمله مناطق کانه‌زایی شده در کمربند ولکانیکی ارومیه-دختر است و شامل دو محدوده تپه جنوبی و تپه شمالی است که با روندی شمال شرقی- جنوب غربی و با فاصله ۱/۷ کیلومتر از یکدیگر واقع شده‌اند[۹، ۱۰]. کانه‌زایی دالی شمالی در یک کمپلکس نفوذی گرانودیوریتی (تونالیتی)، کوارتزدیوریتی و سنگ دیواره آندزیتی نهشته شده است. سنگ‌های آندزیتی، میزبان سنگ‌های کوارتزدیوریت پورفیری هستند که بخش اعظم کانه‌زایی در آنها و کنتاکت با آندزیت‌ها مشاهده می‌شود (شکل ۲) [۹]. نمونه‌های موجود، در یک شبکه اکتشافی ۵۰m×۵۰m برداشت شده و با روش ICP-MS، برای ۴۵ عنصر ژئوشیمیایی آنالیز شده‌اند. از این میان، تعداد ۸ عنصر Au، Cu، Ba، Mn، Sr، Ti، V و Zn که دارای تغییرپذیری دینامیک و معنی‌داری از لحاظ عیار بودند برای این پژوهش انتخاب شده‌اند. در جدول ۱، پارامترهای آماری مجموعه داده انتخابی درج شده است.



شکل ۲: نقشه زمین‌شناسی مقیاس ۱:۱۰۰۰ محدوده کانسار دالی شمالی

جدول ۱: آماره‌های توصیفی عناصر ژئوشیمیایی در نمونه‌های برداشت شده از محدوده کانسار دالی شمالی

عنصر	واحد	تعداد	میانگین	انحراف استاندارد	ضریب تغییرات	کمینه	میانه	بیشینه
Au	ppb	۱۴۹	۲۸۵/۷۰	۴۸۶/۷۰	۱۷۰/۳۵	۲	۴۴	۲۸۶۷
Ba	ppm	۱۴۹	۹۶/۰۱	۳۸/۷۴	۴۰/۳۵	۳۹	۹۰	۳۶۸
Cu	ppm	۱۴۹	۸۶۰/۸۰	۱۰۵۱/۹۰	۱۲۲/۲۰	۴۹	۳۲۰	۵۴۰۳
Mn	ppm	۱۴۹	۵۶۹/۸۰	۱۶۳/۴۰	۲۸/۶۷	۸۳	۵۵۷	۱۰۰۶
Sr	ppm	۱۴۹	۱۲۷/۷۵	۶۳/۰۲	۴۹/۳۳	۴۰	۱۲۰	۳۴۵
Ti	ppm	۱۴۹	۱۱۰۹/۷۰	۵۴۱/۹۰	۴۸/۸۳	۲۹۸	۹۹۱	۳۰۱۱
V	ppm	۱۴۹	۹۵/۴۳	۲۳/۱۳	۲۴/۲۴	۴۴	۹۴	۱۶۸
Zn	ppm	۱۴۹	۹۷/۷۰	۲۲/۹۶	۲۳/۵۰	۴۵	۹۷	۱۹۶

۲-۳- شاخص‌های تعیین تعداد بهینه گروه در مجموعه داده

تمامی شاخص‌هایی که در حوزه بازشناسی الگو برای محاسبه تعداد مناسب گروه در مجموعه داده معرفی شده‌اند، اطلاعات موجود در زمینه تفکیک بین خوشه‌ای و تراکم درون خوشه‌ای را با هم ترکیب نموده و ضمن مد نظر قرار دادن سایر فاکتورها از قبیل خواص هندسی و آماری داده‌ها، ابعاد الگوها (تعداد داده‌ها)، ابعاد ویژگی‌ها (تعداد متغیرها) و معیارهای شباهت، تعداد بهینه خوشه‌ها را تعیین می‌کنند[۳]. در این مقاله برای تخمین تعداد بهینه گروه در داده‌ها از ۳۰ شاخص پرکاربرد در این زمینه استفاده شده است. مؤلفه‌های بنیادین استفاده شده در روابط ریاضی این شاخص‌ها به شرح جدول ۲ است.

در ادامه به تشریح اجمالی شاخص‌ها پرداخته می‌شود:

۱. شاخص کالینسکی- هاراباز[۱۰] **(CH)** [۱۱]، که با رابطه ۱ تعریف می‌شود:

$$\mathrm{CH}(q)=\frac{\mathrm{trace}(B_q)/(q-1)}{\mathrm{trace}(W_q)/(n-q)} \qquad (۱)$$

مقدار q به ازای بیشینه مقدار CH، تعداد بهینه خوشه را مشخص می‌نماید [۱۱].

۲. شاخص Duda [۱۲]، که از رابطه ۲ پیروی می‌کند:

$$\mathrm{Duda}=\frac{Je(2)}{Je(1)}=\frac{W_k+W_l}{W_m} \qquad (۲)$$

که $Je(2)$ و $Je(1)$، مجموع مربعات خطای درون گروهی هستند، هنگامی که داده‌ها به ترتیب به دو و یک خوشه تقسیم شده باشند. فرض می‌شود که خوشه‌های C_k و C_l در قالب C_m ادغام شده‌اند[۱۲]. بهترین تعداد خوشه، کم‌ترین مقدار q است که به ازای آن رابطه ۳ برقرار باشد [۱۳]:

$$\text{Duda} \geq 1-\frac{2}{\pi p}-z\sqrt{\frac{2[1-(8/\pi^2 p)]}{n_m p}} = \text{Critical Value(Duda)} \quad (۳)$$

که z امتیاز نرمال استاندارد است. به صورت تجربی ثابت شده است که بهترین نتیجه به ازای $z=3.20$ حاصل می‌شود [۵].

جدول ۲: مؤلفه‌های بنیادین معادلات شاخص‌های تعیین تعداد گروه.

مؤلفه	شرح	رابطه
n	تعداد مشاهدات	–
p	تعداد متغیرها	–
q	تعداد خوشه‌ها	–
X	ماتریس $n \times p$ داده‌ها	$X=\{x_{ij}\},\ i=1,2,\dots,n,\ j=1,2,\dots,p$
$\bar{X}$	ماتریس $q \times p$ میانگین خوشه‌ها	–
$\bar{x}$	مرکز جرم ماتریس X	–
n_k	تعداد اعضای خوشه C_k	–
c_k	مرکز جرم خوشه C_k	–
x_i	بردار p بعدی مشاهدات مربوط به عضو i ام در خوشه C_k	$\|x\|=(x^{\top}x)^{1/2}$
W_q	ماتریس تفکیک درون گروهی برای داده‌های خوشه‌بندی شده در q خوشه	$W_q=\sum_{k=1}^{q}\sum_{i\in C_k}(x_i-c_k)(x_i-c_k)^{\top}$
B_q	ماتریس تفکیک بین گروهی برای داده‌های خوشه‌بندی شده در q خوشه	$B_q=\sum_{k=1}^{q}n_k(c_k-\bar{x})(c_k-\bar{x})^{\top}$
N_t	تعداد کل جفت نقاط در مجموعه داده	$N_t=\frac{n(n-1)}{2}$
N_w	تعداد کل جفت نقاط متعلق به خوشه یکسان	$N_w=\sum_{k=1}^{q}\frac{n_k(n_k-1)}{2}$
N_b	تعداد کل جفت نقاط متعلق به خوشه‌های مختلف	$N_b=N_t-N_w$
S_w	مجموع فواصل درون گروهی	$S_w=\sum_{k=1}^{q}\sum_{\substack{i,j\in C_k\\ i<j}}d(x_i,x_j)$
S_b	مجموع فواصل بین گروهی	$S_b=\sum_{k=1}^{q-1}\sum_{l=k+1}^{q}\sum_{\substack{i\in C_k\\ j\in C_l}}d(x_i,x_j)$

۳. شاخص Pseudo t^2 [۱۲]، که از رابطه ۴ محاسبه می‌شود:

$$\text{Pseudo } t^2=\frac{V_{kl}}{(W_k+W_l)/(n_k+n_l-2)} \quad (۴)$$

که $V_{kl}=W_m-W_k-W_l$ و $C_m=C_l \cup C_k$ است. بهترین تعداد خوشه، کم‌ترین مقدار q است، به نحوی که[۱۳]:

$$\text{Pseudo } t^2 \leq \frac{1-\text{Critical Value(Duda)}}{\text{Critical Value(Duda)}} \times (n_k+n_l-2) \quad (۵)$$

۴. شاخص C [۱۴]، که با رابطه ۶ تعریف می‌شود:

$$\text{Cindex}=\frac{S_w-S_{\min}}{S_{\max}-S_{\min}},\ S_{\min}\neq S_{\max},\ \text{Cindex}\in(0,1) \quad (۶)$$

که $S_{\min}$، مجموع N_w تا از کوچک‌ترین فواصل میان جفت نقاط و $S_{\max}$، مجموع N_w تا از بزرگ‌ترین فواصل میان جفت نقاط در کل داده‌ها است. کم‌ترین میزان Cindex برای تعیین تعداد مناسب گروه در داده‌ها مورد استفاده قرار می‌گیرد[۵].

۵. شاخص گاما[11] [۱۵]، که طبق رابطه ۷ تعریف شده و مقایسه‌ای میان تمام عدم شباهت‌های درون گروهی و بین گروهی انجام می‌دهد.

$$\text{Gamma} = \frac{s(+) - s(-)}{s(+) + s(-)} \qquad (۷)$$

که $s(+)$ تعداد مقایسه‌های سازگار یا به عبارتی تعداد دفعاتی است که دو نقطه هم‌گروه، فاصله‌ای بیش از دو نقطه غیرهم‌گروه دارند. $s(-)$ نیز تعداد مقایسات ناسازگار است. بیشینه مقدار Gamma، تعداد صحیح خوشه‌ها را معرفی می‌کند[۵].

۶. شاخص بیل[12] [۱۶]، طبق رابطه ۸، از آزمون F برای بررسی فرض وجود q_1 یا q_2 خوشه ($q_2>q_1$) در داده‌ها استفاده می‌کند.

$$\text{Beale} = F \equiv \frac{V_{kl}/W_k + W_l}{[(n_m-1)/(n_m-2)]2^{2/p} - 1} \qquad (۸)$$

تعداد صحیح خوشه‌ها از مقایسه F با توزیع $F_{p,(nm-2)p}$ به دست می‌آید. فرض صفر در این روش آن است که صرفاً یک گروه در مجموعه داده وجود دارد. این فرض زمانی رد می‌شود که مقدار F به طرز معنی‌داری بزرگ باشد[۱۳].

۷. معیار خوشه‌بندی مکعبی[13] **(CCC)** [۱۷]، که از رابطه ۹ تبعیت می‌کند:

$$\text{CCC} = \ln\left[\frac{1-\text{E}(R^2)}{1-R^2}\right]\frac{\sqrt{np^*/2}}{[0.001+\text{E}(R^2)]^2} \qquad (۹)$$

که:

$$R^2 = 1 - \frac{\text{trace}(X^\top X - \bar{X}^\top Z^\top Z\bar{X})}{\text{trace}(X^\top X)} \qquad (۱۰)$$

که Z ماتریس شاخص خوشه از مرتبه $n\times q$ است، به طوری که اگر مشاهده i ام متعلق به خوشه k ام باشد، $z_{ik}=1$ و در غیر این صورت $z_{ik}=0$ است. لذا:

$$\text{E}(R^2) = 1 - \left[\frac{\sum_{j=1}^{p^*}\frac{1}{n+u_j} + \sum_{j=p^*+1}^{p}\frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p}u_j^2}\right]\left[\frac{(n-q)^2}{n}\right]\left[1+\frac{4}{n}\right] \qquad (۱۱)$$

که در رابطه ۱۱:

$$u_j = \frac{s_j}{c} \qquad (۱۲)$$

$$c = \left(\frac{v^*}{q}\right)^{1/p^*} \qquad (۱۳)$$

$$v^* = \prod_{j=1}^{p^*} s_j \qquad (۱۴)$$

که در آنها، s_j ریشه دوم مقدار ویژه j ام ماتریس $X^\top X/(n\text{-}1)$ است. از بیشینه مقدار CCC برای تعیین تعداد بهینه خوشه در مجموعه داده استفاده می‌شود[۵].

۸. شاخص Ptbiserial [۱۸، ۱۹]، یک ضریب همبستگی Point-Biserial میان ماتریس عدم شباعت و یک ماتریس تناظر (شامل درایه‌های صفر و یک) است. زمانی که دو داده متناظر، همراه هم در یک خوشه جای گرفته باشند مقدار صفر، و در غیر این صورت مقدار یک اختصاص می‌یابد. این شاخص با رابطه ۱۵ تعریف می‌شود[۱۹]:

$$\text{Ptbiserial} = \frac{(\bar{S}_b - \bar{S}_w)(N_w N_b/N_t^2)^{1/2}}{s_d} \qquad (۱۵)$$

که $_b=S_b/N_b$ ، $_w=S_w/N_w$ و s_d نیز انحراف استاندارد تمام فواصل است. بیشینه مقدار Ptbiserial برای تعیین مناسب‌ترین تعداد خوشه در مجموعه داده به کار می‌رود [۵].

۹. شاخص Gplus [۱۹، ۲۰] ، که با رابطه ۱۶ بیان می‌شود:

$$\text{Gplus} = \frac{2s(-)}{N_t(N_t-1)} \qquad (۱۶)$$

کمینه مقدار Gplus، ملاک تعیین بهترین تعداد خوشه در داده‌ها است[۵].

۱۰. شاخص دیویس- بولدین[14] **(DB)** [۲۱]، تابعی از مجموع نسبت تراکم درون گروهی به تفکیک بین گروهی است:

$$\text{DB}(q) = \frac{1}{q}\sum_{k=1}^{q}\max_{k\neq 1}\left(\frac{\delta_k+\delta_l}{d_{kl}}\right) \qquad (۱۷)$$

که d_{kl}، فاصله بین مراکز جرم خوشه‌های C_k و C_l است که اغلب به صورت نورم ۲ (فاصله اقلیدسی) محاسبه می‌شود. δ_k، ضریب تراکم خوشه C_k است که به ازای نورم ۲، معادل انحراف استاندارد فواصل اعضای خوشه C_k با مراکز جرم خوشه است. مقدار q به ازای کمینه شاخص DB، تعداد بهینه گروه‌ها را مشخص می‌نماید[۵، ۲۱].

۱۱. شاخص فری[15] [۲۲]، که از رابطه ۱۸ به دست می‌آید:

$$\text{Frey} = \frac{\bar{S}_{b_{j+1}} - \bar{S}_{b_j}}{\bar{S}_{w_{j+1}} - \bar{S}_{w_j}} \qquad (۱۸)$$

این شاخص، نسبتی از اختلاف امتیازات دو سطح (مرحله) متوالی در سلسله مراتب خوشه‌بندی است. صورت کسر، تفاضل میانگین‌های فواصل بین گروهی را در دو سطح از سلسله مراتب خوشه‌بندی (سطوح j و $j+1$) محاسبه می‌کند. مخرج کسر نیز همین فرآیند را برای

فواصل درون گروهی تکرار می‌نماید. مقدار عددی Frey، اغلب در حدود عدد ۱ نوسان می‌کند. بهترین نتیجه زمانی رخ می‌دهد که خوشه‌بندی تا جایی ادامه پیدا کند که مقدار Frey به کم‌تر از واحد برسد. در این حالت، آخرین سطح خوشه‌بندی (ماقبل رسیدن به زیر عدد ۱) به عنوان بهترین سطح تقسیم‌بندی شناخته می‌شود[۵].

۱۲. معیار هارتیگان[۱۶] [۲۳]، که از رابطه ۱۹ پیروی می‌کند:

$$\text{Hartigan} = \left[\frac{\text{trace}(W_q)}{\text{trace}(W_{q+1})} - 1\right](n-q-1)\ ,\ q \in \{1,2,\dots,n-2\} \quad (۱۹)$$

بیشینه اختلاف موجود میان سطوح سلسله مراتبی، معرف مناسب‌ترین تعداد گروه در مجموعه داده است [۵].

۱۳. شاخص Tau [۱۹]، [۲۰]، طبق رابطه ۲۰ از تناظر دو ماتریس محاسبه می‌شود، که ماتریس اول شامل فواصل بین نقاط است و ماتریس دوم شامل درایه‌های صفر و یک است که هم‌خوشه بودن یا نبودن دو نقطه را بیان می‌کند.

$$\text{Tau} = \frac{s(+)-s(-)}{\left\{\left[\frac{N_t(N_t-1)}{2}-t\right]\left[\frac{N_t(N_t-1)}{2}\right]\right\}^{1/2}} \quad (۲۰)$$

که t تعداد مقایسه‌های دو جفت داده را نشان می‌دهد. تعداد بهینه خوشه‌ها بر اساس بیشینه مقدار Tau تعیین می‌شود[۵].

۱۴. شاخص راتکوفسکی[۱۷] [۲۴]، که از طریق رابطه ۲۱ محاسبه می‌شود:

$$\text{Ratkovsky} = \frac{\bar{S}}{q^{1/2}} \quad (۲۱)$$

که:

$$\bar{S} = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\frac{BGSS_j}{TSS_j}} \quad (۲۲)$$

که در آن، $BGSS$ به مجموع مربعات بین گروهی و TSS به مجموع مربعات کلی هر متغیر اشاره می‌کند. تعداد بهینه خوشه، مقداری از q است که به ازای آن، بیشینه شاخص راتکوفسکی محاسبه می‌شود[۵].

۱۵. شاخص اسکات[۱۸] [۲۵]، که از رابطه ۲۳ به دست می‌آید:

$$\text{Scott} = n\log\frac{\det(T)}{\det(W_q)} \quad (۲۳)$$

که T، مجموع مربعات کل است. بیشینه اختلاف میان سطوح سلسله مراتبی، تعداد صحیح گروه‌ها را مشخص می‌کند[۵].

۱۶. شاخص ماریوت[۱۹] [۲۶]، که از رابطه ۲۴ تبعیت می‌کند:

$$\text{Marriot} = q^2\det(W_q) \quad (۲۴)$$

در مورد این شاخص نیز بیشینه اختلاف میان سطوح سلسله مراتبی جهت تعیین بهترین سطح تقسیم‌بندی به کار می‌رود[۵].

۱۷. شاخص بال[۲۰] [۲۷]، با رابطه ۲۵ تعریف می‌شود:

$$\text{Ball} = \frac{W_q}{q} \quad (۲۵)$$

بیش‌ترین اختلاف میان سطوح متوالی سلسله مراتب خوشه‌بندی برای یافتن جواب بهینه مورد استفاده قرار می‌گیرد[۵].

۱۸. شاخص Trcovw [۵]، که طبق رابطه ۲۶ حاصل می‌شود:

$$\text{Trcovw} = \text{trace}[\text{cov}(W_q)] \quad (۲۶)$$

بیشینه اختلاف امتیاز میان سطوح سلسله مراتبی به عنوان بهترین پاسخ در نظر گرفته می‌شود[۵].

۱۹. شاخص Tracew [۵]، که به صورت رابطه ۲۷ محاسبه می‌شود:

$$\text{Tracew} = \text{trace}(W_q) \quad (۲۷)$$

با کاهش تعداد خوشه‌ها، Tracew به طور یکنواخت افزایش می‌یابد. بیشینه مقدار تفاضل دوم امتیازات سطوح سلسله‌مراتبی، تعداد مناسب گروه‌ها در مجموعه داده را مشخص می‌کند[۵].

۲۰. شاخص فریدمن[۲۱] [۲۸]، که از رابطه ۲۸ به دست می‌آید:

$$\text{Friedman} = \text{trace}(W_q^{-1}B_q) \quad (۲۸)$$

بیشینه اختلاف در مقادیر متوالی شاخص فریدمن، معیار تعیین پاسخ بهینه است[۵].

۲۱. شاخص مک‌کلین[۲۲] [۲۹]، که از رابطه ۲۹ پیروی می‌کند:

$$\text{McClain} = \frac{\bar{S}_w}{\bar{S}_b} = \frac{S_w/N_w}{S_b/N_b} \quad (۲۹)$$

کمینه مقدار شاخص مک‌کلین برای مشخص نمودن بهترین تعداد گروه در مجموعه داده به کار می‌رود[۳].

۲۲. شاخص رابین[۲۳] [۲۸]، که مطابق رابطه ۳۰ تعیین می‌شود:

$$\text{Rubin} = \frac{\det(T)}{\det(W_q)} \quad (۳۰)$$

کم‌ترین مقدار تفاضل دوم میان سطوح متوالی خوشه‌بندی به منظور تشخیص تعداد گروه در مجموعه داده استفاده می‌شود[۵].

۲۳. شاخص کرزانوفسکی- لای[۲۴] **(KL)** [۳۰]، که با رابطه ۳۱ تعریف می‌شود:

$$\mathrm{KL}(q)=\left|\frac{\mathrm{DIFF}_q}{\mathrm{DIFF}_{q+1}}\right| \quad (۳۱)$$

که:

$$\mathrm{DIFF}_q=(q-1)^{2/p}\mathrm{trace}(W_{q-1})-q^{2/p}\mathrm{trace}(W_q) \quad (۳۲)$$

بهترین مقدار q به ازای بیشینه شاخص KL حاصل می‌شود[۳].

۲۴. ضریب سیلووت[۶]، که به صورت رابطه ۳۳ محاسبه می‌شود:

$$\mathrm{Sillhouette}=\frac{\sum_{i=1}^{n}S(i)}{n}\ ,\ \mathrm{Sillhouette}\in[-1,+1] \quad (۳۳)$$

که:

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i);b(i)\}} \quad (۳۴)$$

که $a(i)$، متوسط عدم شباهت عضو i ام با تمام اعضای خوشه C_r است:

$$a(i)=\frac{\sum_{j\in\{C_r\setminus i\}}d_{ij}}{n_r-1} \quad (۳۵)$$

$$b(i)=\min_{s\neq r}\left\{d_{i_{C_s}}\right\} \quad (۳۶)$$

که:

$$d_{i_{C_s}}=\frac{\sum_{j\in C_c}d_{ij}}{n_s} \quad (۳۷)$$

بیش‌ترین مقدار ضریب سیلووت به منظور تعیین مناسب‌ترین تعداد گروه در مجموعه داده مورد استفاده قرار می‌گیرد[۳].

۲۵. آماره گپ [۷]، که با رابطه ۳۸ بیان می‌شود:

$$\mathrm{Gap}(q)=\frac{1}{B}\sum_{b=1}^{B}\log W_{qb}-\log W_q \quad (۳۸)$$

که B، تعداد مجموعه داده مرجعی است که با استفاده از تابع یکنواخت تولید شده است [۷]. W_{qb} نیز ماتریس پراکندگی درون گروهی است که در معیار هارتیگان تعریف شده است. بهترین تعداد گروه در مجموعه داده، معادل کمینه q است که:

$$\mathrm{Gap}(q)\geq\mathrm{Gap}(q+1)-s_{q+1}\ ,\quad q=1,2,\dots,n-2 \quad (۳۹)$$

که:

$$s_q=sd_q\sqrt{1+(1/B)} \quad (۴۰)$$

که در آن، sd_q انحراف استاندارد مقادیر $\log W_{qb}$ است[۳].

۲۶. شاخص D [۳۱]، که بر مبنای میزان بهره خوشه‌بندی[۲۵] در اینرسی درون خوشه‌ای تعریف می‌شود. اینرسی درون خوشه‌ای پارامتری است که درجه همگنی میان اعضای یک خوشه را می‌سنجد. Dindex با رابطه ۴۱ تعریف می‌گردد:

$$w(P^q)=\frac{1}{q}\sum_{k=1}^{q}\frac{1}{n_k}\sum_{x_i\in C_k}d(x_i,C_k) \quad (۴۱)$$

با در نظر گرفتن دو سطح تقسیم‌بندی، P^{k-1} متشکل از k-1 خوشه و P^k متشکل از k خوشه است. بهره خوشه‌بندی بر مبنای اینرسی درون خوشه‌ای و با رابطه ۴۲ محاسبه می‌شود:

$$\mathrm{Gain}=w(P^{q-1})-w(P^q) \quad (۴۲)$$

میزان Gain باید کمینه شود. با رسم نمودار تفاضل دوم Gain در برابر q، نقطه زانوی نمودار (جهش بزرگ مقادیر بهره) مشخص کننده تعداد بهینه گروه در داده خواهد بود[۳].

۲۷. شاخص دان[۲۶] [۴]، که بر اساس نسبت کم‌ترین تفکیک بین گروهی به بیش‌ترین تراکم درون گروهی تعریف می‌شود:

$$\mathrm{Dunn}=\frac{\min\limits_{1\leq i\leq j\leq q}d(C_i,C_j)}{\max\limits_{1\leq k\leq q}\mathrm{diam}(C_k)} \quad (۴۳)$$

که $d(C_i,C_j)$ تابع عدم شباهت بین خوشه‌های C_i و C_j است. $\mathrm{diam}(C_k)$ قطر خوشه است که می‌تواند به عنوان معیاری جهت سنجش پراکندگی و انتشار اعضای خوشه به کار رود. چنان‌چه تراکم خوشه بالا باشد، انتظار می‌رود که قطر خوشه کوچک و فاصله بین خوشه‌ها زیاد باشد، لذا شاخص Dunn باید عددی بزرگ باشد[۳].

۲۸. آماره Γ هوبرت[۲۷] [۳۲]، که از رابطه ۴۴ به دست می‌آید:

$$\Gamma(P,Q)=\frac{1}{N_t}\sum_{\substack{i=1\\i<j}}^{n-1}P_{ij}Q_{ij} \quad (۴۴)$$

همان‌گونه که ملاحظه می‌شود، Γ یک ضریب همبستگی Point-Biserial میان ماتریس‌های P و Q است. P ماتریس همسایگی مجموعه داده است. Q نیز یک ماتریس $n\times n$ است که درایه‌های (i,j) آن، معادل فاصله بین نقاط نماینده (v_{c_i},v_{c_j}) خوشه‌هایی است که اعضای x_i

و x_j به آن تعلق دارند. چنان‌چه Γ نسبت به میانگین و انحراف استاندارد ماتریس‌های P و Q نرمال شود، عددی در بازه [۰،۱] خواهد بود. مقادیر بزرگ Γ نرمال شده، نشان دهنده وجود خوشه‌هایی با درجه تراکم بالا است، لذا نقطه زانو در نمودار Γ در برابر q، معرف بهترین تعداد گروه در مجموعه داده خواهد بود[۳۳]. البته رسم نمودار تفاضل دوم Γ در برابر q می‌تواند در تشخیص جهش بزرگ Γ و تمییز آن از سایر پیک‌های آنومال مؤثر باشد[۳].

۲۹. شاخص SD [۳۳]، که بر اساس رابطه ۴۵ بیان می‌شود:

$$\mathrm{SDindex}(q) = \alpha \mathrm{Scat}(q) + \mathrm{Dis}(q) \qquad (۴۵)$$

Scat در رابطه ۴۵، بیان‌گر متوسط تراکم خوشه‌ها (فاصله درون گروهی) است که مطابق رابطه ۴۶ محاسبه می‌شود:

$$\mathrm{Scat}(q) = \frac{\frac{1}{q}\sum_{k=1}^{q}\left\|\sigma^{(k)}\right\|}{\|\sigma\|} \qquad (۴۶)$$

که σ بردار واریانس هر متغیر در مجموعه داده p متغیره و $\sigma^{(k)}$ بردار واریانس هر خوشه C_k است. مقادیر کوچک Scat، نشان از تراکم بالای خوشه‌ها دارد.

مؤلفه Dis در رابطه ۴۵، نشان‌گر تفکیک کلی خوشه‌ها (فاصله بین گروهی) است و از رابطه ۴۷ به دست می‌آید:

$$\mathrm{Dis}(q) = \frac{D_{\max}}{D_{\min}}\sum_{k=1}^{q}\left(\sum_{z=1}^{q}\|C_k - C_z\|\right)^{-1} \qquad (۴۷)$$

که $D_{\max}$ و $D_{\min}$ به ترتیب معرف بیشینه و کمینه فاصله بین گروهی است. مؤلفه α در رابطه ۴۵ یک فاکتور وزنی است که معادل $\mathrm{Dis}(q_{\max})$ است. تعداد بهینه خوشه‌ها به ازای کم‌ترین مقدار SDindex محاسبه می‌شود[۳۳].

۳۰. شاخص SDbw [۳۴]، همچون SDindex بر مبنای معیار تفکیک و تراکم خوشه‌ها و بر اساس رابطه ۴۸ تعریف می‌شود:

$$\mathrm{SDbw}(q) = \mathrm{Scat}(q) + \mathrm{Density.bw}(q) \qquad (۴۸)$$

مؤلفه Scat در رابطه ۴۸، همان مفهوم موجود در رابطه ۴۵ را دارا است. Density.bw چگالی بین گروهی است که نسبت متوسط چگالی بین گروهی را به چگالی داخل گروه‌ها می‌سنجد و از رابطه ۴۹ محاسبه می‌شود:

$$\mathrm{Density.bw}(q) = \frac{1}{q(q-1)}\sum_{i=1}^{q}\left[\sum_{\substack{j=1\\ j\neq i}}^{q}\frac{\mathrm{density}(u_{ij})}{\max[\mathrm{density}(c_i),\mathrm{density}(c_j)]}\right] \qquad (۴۹)$$

که در آن، u_{ij} نقطه میانی پاره خط واصل مراکز جرم c_i و c_j است. $\mathrm{density}(u_{ij})$ مطابق رابطه ۵۰ به دست می‌آید:

$$\mathrm{density}(u_{ij}) = \sum_{l=1}^{n_{ij}} f(x_l, u_{ij}) \qquad (۵۰)$$

که n_{ij}، تعداد اعضای متعلق به خوشه‌های C_i و C_j است. $f(x_l,u_{ij})$ تابعی باینری است، به نحوی که اگر $d(x,u_{ij})>\mathrm{Stdev}$ باشد، برابر صفر و در غیر این صورت برابر واحد است. پارامتر Stdev، بیان‌گر متوسط انحراف استاندارد خوشه‌ها است. مناسب‌ترین تعداد گروه در مجموعه داده، به ازای کمینه شاخص SDbw به دست می‌آید[۳۴].

۳- نتایج و بحث

برای محاسبه ۳۰ شاخص مطرح شده در تعیین تعداد گروه در مجموعه داده، از فرمان‌ها و توابع موجود در پکیج نرم‌افزاری NbClust [۳]، که در محیط برنامه‌نویسی R (زبان آماری S) [۳۵] تألیف شده، استفاده شده است. این پکیج، جامع‌ترین بسته نرم‌افزاری R در زمینه تخمین تعداد گروه در داده‌ها است که کلیه توابع لازم برای محاسبه شاخص‌های مورد مطالعه، با طیف وسیعی از روش‌های خوشه‌بندی و نیز گستره متنوعی از معیارهای فاصله را در اختیار گذاشته است. در این مقاله، از تکنیک خوشه‌بندی k-means که روشی متداول و کاربردی در آنالیز خوشه‌ای مجموعه داده‌های ژئوشیمیایی است، برای سنجش گروه-بندی‌های نهان در داده‌ها استفاده شده است. همچنین معیار فاصله اقلیدسی به عنوان متریک اندازه‌گیری فواصل مختلف و نیز ماتریس عدم شباهت در نظر گرفته شده است. تعداد حالات ممکن رخداد خوشه‌ها (کران بالا و پایین خوشه‌ها) در بازه [۲،۱۰] منظور شده است تا پاسخ بهینه از این بازه استخراج گردد. بدین ترتیب، اقدام به اجرای توابع شاخص‌های مختلف در سطح معنی‌دار بودن α=0.10 شده است.

۳-۱- مجموعه داده شبیه‌سازی شده

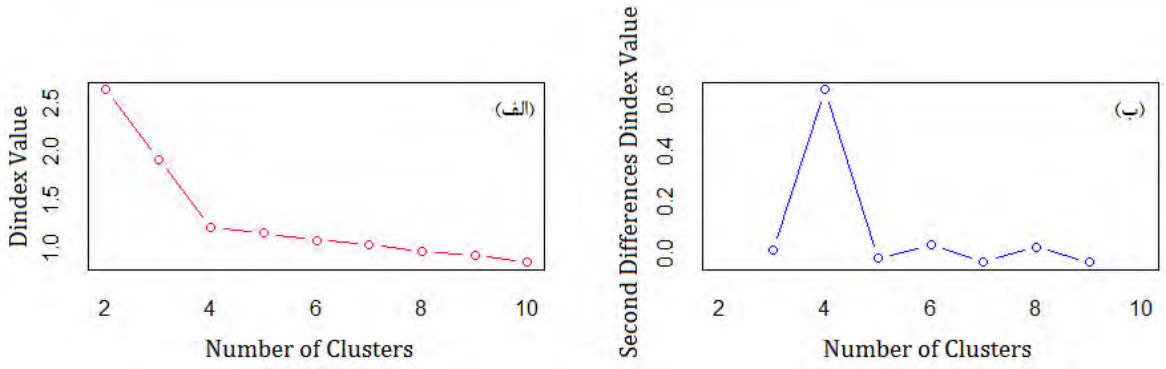
در جدول ۳، تعداد خوشه محاسبه شده توسط هر شاخص به همراه مقدار انتخابی آن در مجموعه داده شبیه-سازی شده درج شده است. همان‌گونه که ملاحظه می‌شود، اکثریت روش‌های اجرا شده (۱۹ روش) موفق به شناسایی تعداد واقعی گروه‌ها (۴ خوشه) شده‌اند. یک روش تعداد ۱

خوشه، شش روش تعداد ۲ خوشه، سه روش تعداد ۳ خوشه و یک روش تعداد ۷ خوشه را معرفی کرده‌اند. این نتایج، کارآیی روش‌های معرفی شده و نیز تصمیم‌گیری بر اساس قانون اکثریت را در مورد پاسخ‌های به دست آمده، تأیید می‌نماید. در این بین، عملکرد شاخص D و آماره هوبرت به صورت گرافیکی و بر اساس تشخیص نقطه زانوی نمودار شاخص و نقطه پیک تفاضل دوم شاخص است. نتایج این دو روش به ترتیب در شکل‌های ۳ و ۴ نشان داده شده است. با خوشه‌بندی داده‌ها به روش k-means در حالت k=4، مختصات مراکز جرم خوشه‌های تشکیل شده به صورت (۰/۹۶ ، ۱)، (۵/۹۵ ، ۱/۱۹)، (۱/۰۶ ، ۶/۰۲) و (۵/۹۶ ، ۵/۹۳) محاسبه شده است، که تقریب آشکاری از مراکز جرم حقیقی خوشه‌های شبیه‌سازی شده است.

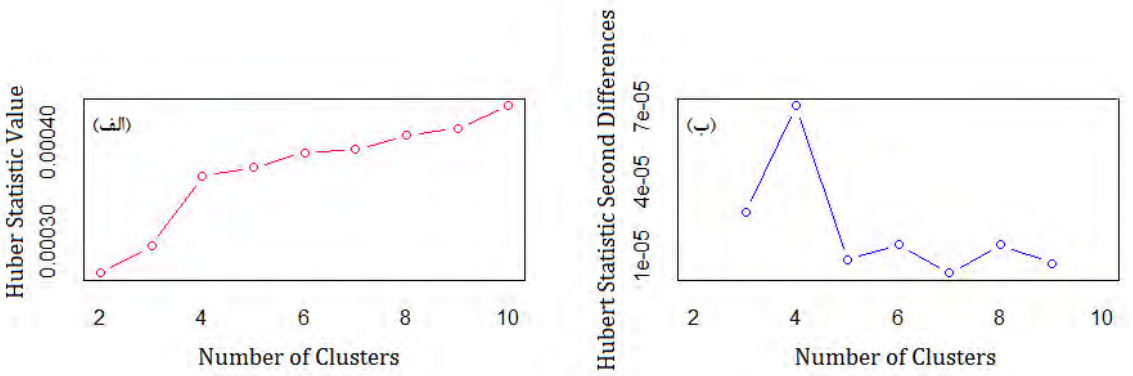
جدول ۳: تعداد بهینه گروه و مقدار مطلوب شاخص‌های اعمال شده بر روی مجموعه داده شبیه‌سازی شده

شاخص	تعداد بهینه‌ی خوشه	مقدار مطلوب شاخص	شاخص	تعداد بهینه‌ی خوشه	مقدار مطلوب شاخص
CH	۴	۸۶۲/۱۱	Marriot	۴	۴۴۶۵۲۸۸
Duda	۲	۱/۰۹	Ball	۳	۹۳۸/۸۵
Pseudo t^2	۲	-۲۶/۴۹	Trcovw	۳	۲۷۷۰۱۴۴
Cindex	۴	۰/۲۸	Tracew	۴	۱۰۶۵/۲۷
Gamma	۴	۰/۹۷	Friedman	۴	۲۰/۲۷
Beale	۲	-۰/۰۹	McClain	۲	۰/۵۶
CCC	۴	۳۵/۰۲	Rubin	۴	-۱۰/۴۴
Ptbiserial	۴	۰/۷۶	KL	۷	۷۳/۳۸
Gplus	۴	۱۷۷/۶۴	Sillhoutte	۴	۰/۶۰
DB	۴	۰/۵۶	Gap	۲	-۰/۶۵
Frey	۱	-	Dindex	۴	*
Hartigan	۴	۵۶۶/۹۵	Dunn	۲	۰/۰۸
Tau	۴	۱۴۵۳۴/۴۶	Γ	۴	*
Ratkovsky	۳	۰/۴۶	SD	۴	۰/۵۲
Scott	۴	۶۳۸/۹۵	SDbw	۴	۰/۱۵

*روش گرافیکی



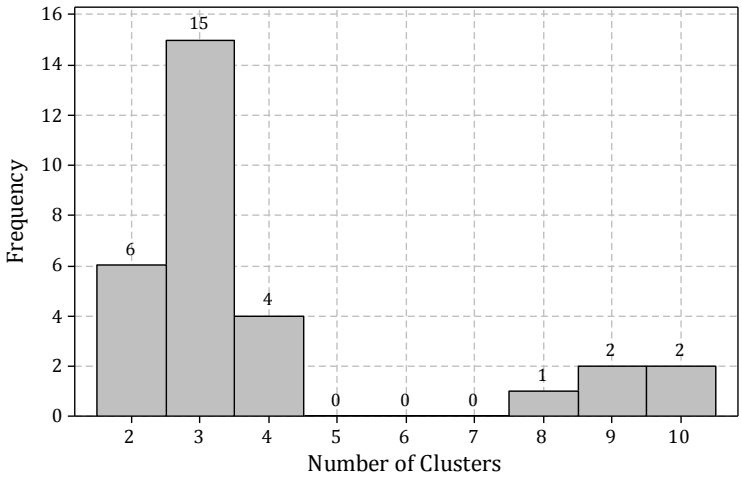
شکل ۳: الف) نمودار شاخص D و ب) نمودار تفاضل دوم شاخص D در برابر تعداد خوشه مجموعه داده شبیه‌سازی شده



شکل ۴: الف) نمودار آماره هوبرت و ب) نمودار تفاضل دوم آماره هوبرت در برابر تعداد خوشه مجموعه داده شبیه‌سازی شده

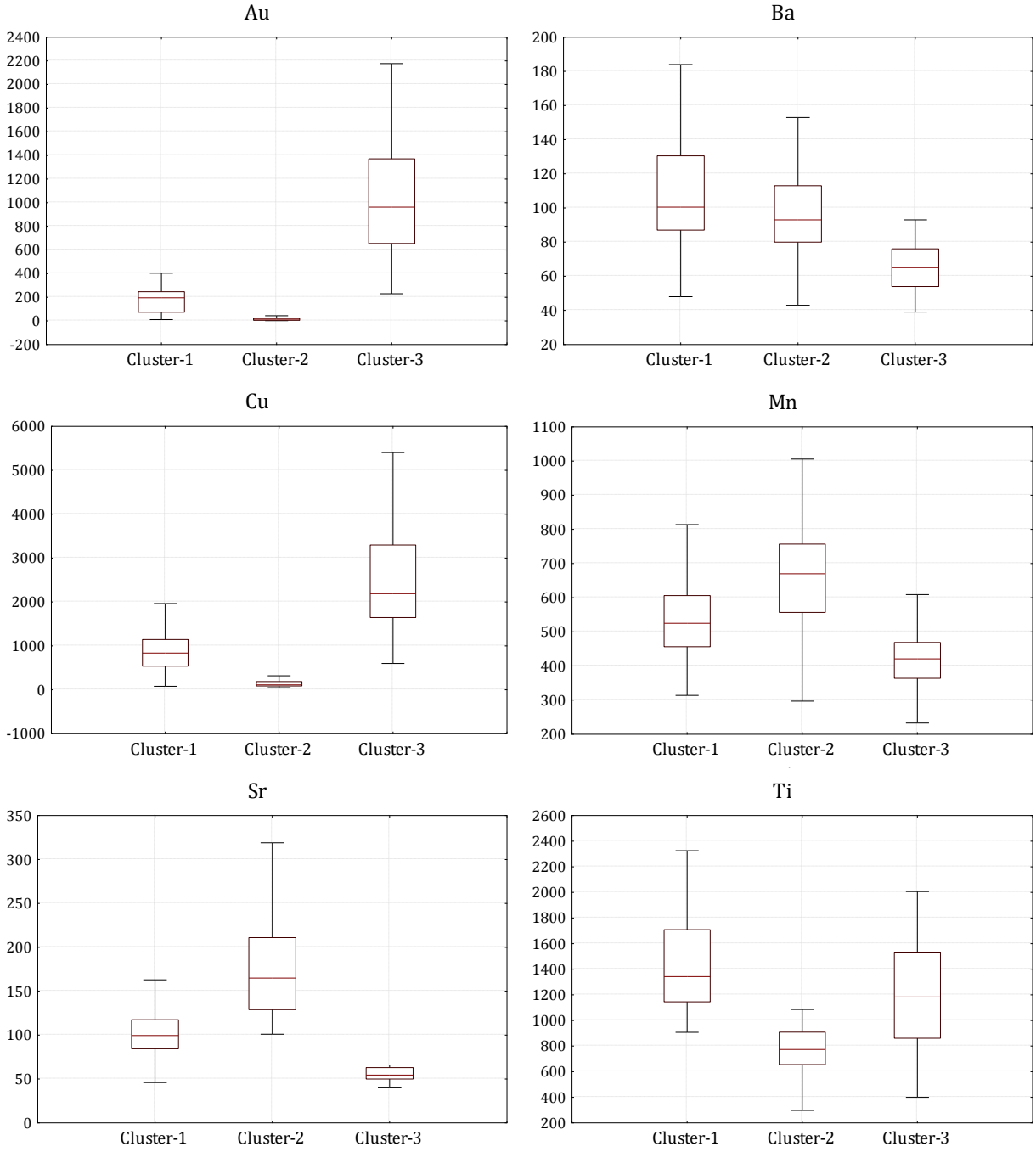
۳-۲- مجموعه داده ژئوشیمیایی

با اجرای ۳۰ شاخص معرفی شده بر روی مجموعه داده کانسار دالی شمالی، تعداد گروه محاسبه شده با هر روش مشخص شده است. شکل ۵، توزیع فرکانسی مجموعه پاسخ‌های تولید شده از اجرای شاخص‌ها را نشان می‌دهد. همان‌گونه که مشاهده می‌شود، مقدار مد توزیع که معادل تعداد گروه با بیشترین فرکانس (۱۵ شاخص) است، برابر ۳ خوشه است. تعداد ۲ خوشه با فرکانس شش، ۴ خوشه با فرکانس چهار، ۹ و ۱۰ خوشه با فرکانس دو و ۸ خوشه با فرکانس یک، در اولویت‌های بعدی قرار دارند.



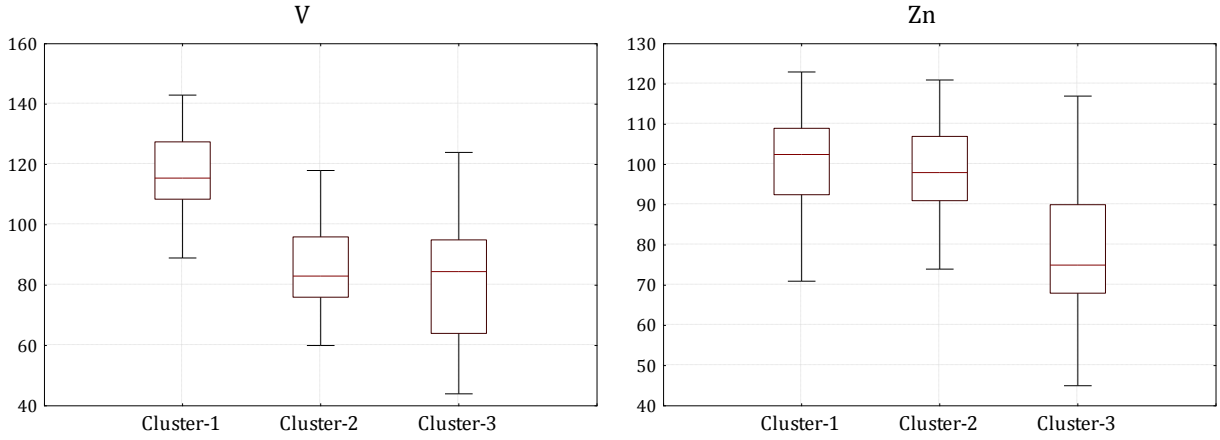
شکل ۵: نمودار توزیع فرکانسی پاسخ‌های تولید شده از اجرای شاخص‌های تعیین تعداد گروه بر روی مجموعه داده ژئوشیمیایی کانسار دالی شمالی

با توجه به این نتایج و بنا به قانون اکثریت، می‌توان تعداد ۳ گروه نهان را برای مجموعه داده ژئوشیمیایی برداشت شده از کانسار دالی شمالی متصور بود. با اجرای خوشه‌بندی k-means در حالت k=3، تمام ۱۴۹ نمونه در دسترس به سه گروه تقسیم‌بندی شده‌اند. شکل ۶، نمودار جعبه‌ای هر عنصر ژئوشیمیایی را به تفکیک زیرجوامع جدا شده نشان می‌دهد. همان‌طور که در این شکل مشاهده می‌شود، اختلاف آماری معنی‌داری میان مشخصات توزیع عناصر در هر یک از سه خوشه محاسبه شده وجود دارد.

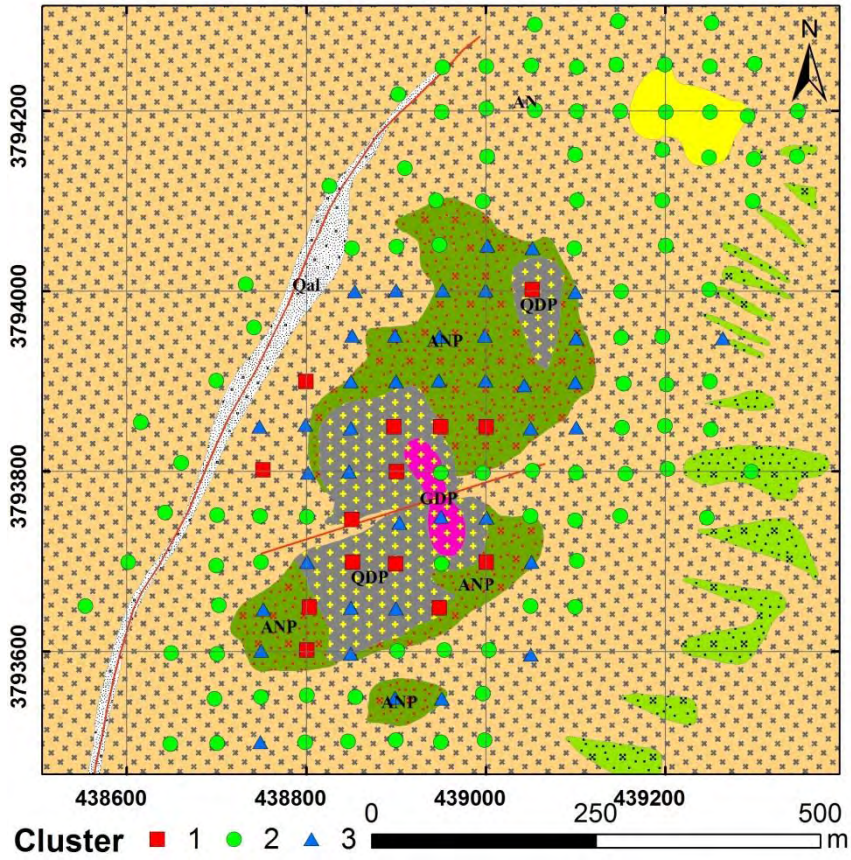


شکل ۶: نمودار جعبه‌ای توزیع عناصر ژئوشیمیایی در سه گروه تشخیص داده شده در مجموعه داده کانسار دالی شمالی.

جهت تفسیر رفتار عناصر در هر یک از گروه‌های سه‌گانه و بررسی رابطه فضایی آنها با واحدهای زمین‌شناسی و نیز زون‌های آلتراسیونی و کانه‌زایی موجود در محدوده تپه شمالی، نقشه پراکندگی نقاط نمونه‌برداری شده به تفکیک هر خوشه تهیه شده و در شکل ۷ نمایش داده شده است. همچنین در جدول ۴، تعداد نمونه‌های واقع در خوشه‌ها به تفکیک هر واحد زمین‌شناسی ارائه شده است. این جدول در حقیقت، نشان دهنده میزان همپوشانی فضایی خوشه‌های تشکیل شده با واحدهای زمین‌شناسی منطقه است. با مد نظر قرار دادن توأمان شباهت رفتاری توزیع عناصر در شکل ۶ و موقعیت فضایی و ساختاری نمونه‌های خوشه‌بندی شده در شکل ۷، می‌توان به وجود یک منطقه-بندی ژئوشیمیایی در محدوده تپه شمالی کانسار دالی پی برد.



ادامه شکل ۶: نمودار جعبه‌ای توزیع عناصر ژئوشیمیایی در سه گروه تشخیص داده شده در مجموعه داده کانسار دالی شمالی



شکل ۷: موقعیت فضایی نمونه‌های خوشه‌بندی شده و ارتباط آنها با ساختارهای زمین‌شناسی در محدوده کانسار دالی شمالی. راهنمای نقشه مطابق شکل ۲ است

بر این اساس، خوشه اول، معرف بخش هسته‌ای کانه‌زایی است، به نحوی که ۵۷/۱٪ از نمونه‌های مربوط به آن در محدوده واحد کوارتزدیوریت پورفیری جای گرفته‌اند که شامل بخش اصلی زون کانه‌زایی مس- طلا است. ۲۸/۶٪ از این گروه در مرز کنتاکت با آندزیت پورفیری قرار دارند که زون کانه‌زایی درجه دوم و ضعیف‌تر محدوده محسوب می‌شود. ۸۹/۶٪ از نمونه‌های متعلق به خوشه دوم، منطبق بر مناطق آندزیتی هستند که به طور ضعیف تا حد واسط تحت تأثیر آلتراسیون پروپیلیتیک قرار گرفته‌اند و مناطق حاشیه دور و سطح فرسایش کم‌عمق کانه‌زایی را معرفی می‌کنند. با توجه به نمودارهای جعبه‌ای شکل ۶ می‌توان استنباط نمود که این خوشه معرف عناصر

فوق‌کانساری در کانه‌زایی مس- طلای پورفیری نظیر Mn و Zn است. ۵۳/۸٪ از نمونه‌های قرار گرفته در خوشه سوم نیز منطبق بر سنگ دیواره آندزیت پورفیری هستند که مناطق مرزی و حاشیه نزدیک کانه‌زایی را معرفی می‌کنند و عموماً در زون‌های آلتراسیون کوارتز مگنتیت و پتاسیک جای گرفته‌اند. با توجه به حضور مگنتیت در این واحدها، می‌توان خوشه سوم را در ارتباط با عناصر سیدروفیل نظیر Ti و V در منطقه دانست. بدین ترتیب، نتایج حاصل از خوشه‌بندی مجموعه داده با تعداد سه گروه را می‌توان قابل تفسیر و منطبق با شواهد زمین‌شناسی، آلتراسیونی، کانه‌زایی و ژئوشیمیایی منطقه ارزیابی نمود.

جدول ۴: میزان تعلق نمونه‌های هر خوشه به واحدهای زمین‌شناسی در محدوده کانسار دالی شمالی

کلاس پیش‌بینی شده	تعداد نمونه‌های مرتبط با واحد زمین‌شناسی		
	Cluster-1	Cluster-2	Cluster-3
آندزیت	۲	۸۶	۹
آندزیت پورفیری	۴	۲	۲۱
گرانودیوریت پورفیری	۰	۱	۱
کوارتزدیوریت پورفیری	۸	۱	۸
سایر	۰	۶	۰

لازم به توضیح است که در این پژوهش از فرم ترکیبی یا بسته داده‌ها برای تعیین تعداد بهینه گروه در مجموعه داده ژئوشیمیایی محدوده دالی شمالی استفاده شده است. نظر به این که خارج کردن داده‌های ژئوشیمیایی از شکل بسته آنها، موجب کشف دانش اضافی و مستدل‌تری درباره روابط عناصر در فضای چندمتغیره می‌گردد، لذا اعمال این پیش‌پردازش می‌تواند تا حدودی روند محاسبه پارامترهای تفکیک و تراکم خوشه‌ها را تغییر داده و نتایج نهایی را تحت تأثیر قرار دهد. به عنوان یک آزمون، از روش شناخته شده نسبت لگاریتمی متمرکز ($clr^{۲۸}$) [۳۶] جهت باز کردن مجموعه داده تحت مطالعه استفاده شده است. به منظور بررسی تأثیرپذیری شاخص‌های به کار رفته از فرم باز داده‌های ژئوشیمیایی، پنج شاخص متداول و شناخته شده (دیویس- بولدین، هارتیگان، راتکوفسکی، بال و سیلووت) جهت تخمین تعداد صحیح خوشه‌ها استفاده شده است. نتایج به دست آمده حاکی از آن است که چهار شاخص از پنج شاخص (دیویس- بولدین، راتکوفسکی، بال و سیلووت) پاسخ یکسانی برای داده‌های باز و بسته ارائه می‌دهند، در حالی که این پدیده موجب تعییر پاسخ بهینه در مورد شاخص هارتیگان می‌گردد. لذا به نظر می‌رسد که پیش‌پردازش باز کردن داده‌های ترکیبی می‌تواند تا حدودی موجب تغییر در نتایج نهایی گردد. البته ماهیت عددی داده‌های باز شده به شکلی است که گاهاً و در مورد چند روش خاص (شاخص‌هایی چون CCC، اسکات، فریدمن، ماریوت، trcovw، tracew و رابین) موجب تکین‌شدگی سیستم از نظر محاسباتی می‌شود، که این مسئله از نکات منفی رویکرد مبتنی بر داده‌های باز است.

۴- نتیجه‌گیری

در این مقاله برای نخستین بار در حوزه مطالعات ژئوشیمیایی، از ۳۰ شاخص بازشناسی الگوی مبتنی بر تفکیک و تراکم خوشه‌ها برای محاسبه تعداد گروه در مجموعه داده استفاده شده و پاسخ بهینه این مسئله گسسته بر اساس بیش‌ترین فرکانس موجود در توزیع فراوانی جواب‌ها استخراج شده است. به کارگیری این روش در مورد داده‌های شبیه‌سازی شده منجر به شناسایی صحیح تعداد گروه‌ها (۴ گروه) و نیز مختصات مراکز جرم خوشه‌های مصنوعی شده است. در مورد مجموعه داده حقیقی برداشت شده از محدوده سیستم پورفیری دالی شمالی، تعداد ۳ خوشه به عنوان تعداد بهینه گروه تشخیص داده شده است. از روش خوشه‌بندی k-means برای دسته‌بندی داده‌ها بر اساس مقدار k تخمینی استفاده شده است. روش k-means یک تکنیک خوشه‌بندی افرازی است که از لحاظ سرعت، سهولت اجرا و پاسخ‌های قابل قبول به سایر رویکردهای خوشه‌بندی ارجحیت دارد و از این رو به عنوان متداول‌ترین و پرکاربردترین روش در

حوزه مطالعات ژئوشیمیایی مطرح شده است. خوشه‌های تشکیل شده بر این اساس، ضمن تفسیر معنی‌دار واحدهای زمین‌شناسی و زون‌های آلتراسیونی موجود در منطقه کانه‌زایی دالی شمالی، حاکی از وجود نوعی منطقه‌بندی ژئوشیمیایی در محدوده کانسار بوده که به نحو مناسبی مناطق داخلی، مرزی و خارجی سیستم پورفیری را دسته‌بندی کرده است. کاربرد این راهکار، واحد کوارتزدیوریت پورفیری و مرز آن با واحد آندزیتی را به عنوان مناطق امیدبخش برای حفاری اکتشافی پیشنهاد نموده است. رویکردهای سنتی موجود در زمینه تخمین تعداد گروه در داده‌ها، تا کنون مبتنی بر تجربیات کارشناسی ژئوشیمیست اکتشافی و یا استفاده از یک یا چند شاخص محاسباتی خاص بوده که با طیف وسیعی از عدم‌قطعیت مواجه هستند. در این راستا، وجود عوامل ژئوشیمیایی پنهان و یا ناپایداری عددی شاخص به کار رفته، می‌تواند منجر به حصول نتایج غیرواقعی گردد. بنابراین، رویکرد توزیع محور این پژوهش ضمن کاهش سطح عدم‌قطعیت پردازش داده‌ها، بهینه‌سازی نتایج حاصل از مطالعات ژئوشیمیایی را تضمین می‌نماید. تکنیک‌های مورد استفاده در این پژوهش در پکیج‌های نرم‌افزاری تألیف شده در محیط برنامه‌نویسی R در دسترس محققان بوده و قابل کاربرد برای انواع داده‌های ژئوشیمیایی هستند. در این راستا و به عنوان دورنمایی جهت ادامه‌ی پژوهش، پیشنهاد می‌شود که قابلیت‌های نشان داده شده در این تحقیق مبنی بر شناسایی تعداد صحیح گروه‌های ژئوشیمیایی، بر اساس داده‌های ژئوشیمیایی باز شده بررسی و میزان تفسیرپذیری نتایج حاصل با رویکرد متداول مبتنی بر داده‌های ترکیبی مقایسه شود.

مراجع

[1] Meng, H. D., Song, Y. C., Song, F. Y., and Shen, H. T. (2011). Research and application of cluster and association analysis in geochemical data processing. Comput. Geosci, 15(1), 87–98.

[2] Gazley, M. F., Collins, K. S., Roberston, J., Hines, B. R., Fisher, L. A., & McFarlane, A. (2015). Application of principal component analysis and cluster analysis to mineral exploration and mine geology. In AusIMM New Zealand Branch Annual Conference.

[3] Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. J. Stat. Softw, 61(i06), 1-36.

[4] Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. J. Cybern, 4(1), 95–104.

[5] Milligan G. W., and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159–179.

[6] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math, 20(1), 53–65.

[7] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Statistical Methods), 63(2), 411–423.

[8] Zaremotlagh, S., Hezarkhani, A., and Sadeghi, M. (2016). Detecting homogenous clusters using whole-rock chemical compositions and REE patterns: A graph-based geochemical approach. J. Geochemical Explor., 170(1), 94–106.

[9] Golestan, F. D., Riabi, S. R. G., Majlesi, M. J., Memarzadeh, M., and Harooni, H. A. (2013). "Identification and Separation of Anomal Variable Using Correspondence and Discriminant Analyses Methods at Northern–Dalli Areae." Journal of Analytical and Numerical Methods in Mining Engineering, **2**(3): 35–43 (In Persian).

[10] Golestan, F. D., Riabi, S. R. G., Hezarkhani, A., Khalookakaei, A. R., Sakaki, S. H., and Harooni, H. A. (2016). "The Structure of Exploration Project Management by Spatial Geometry Methods for Separation Anomaly Using GERT Networking - A Case Study of Cu-Au Northern-Dally Porphyry." Journal of Analytical and Numerical Methods in Mining Engineering, **6**(11): 1–10 (In Persian).

[11] Caliński, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. Commun. Stat. Methods, 3(1), 1–27.

[12] Duda, R. O., and Hart, P. E. (1973). Pattern classification and scene analysis. vol. 3, Wiley New York.

[13] Gordon, A. D. (1999). Classification. Monogr. Stat. Appl. Probab, vol. 82.

[14] Hubert, L. J., and Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. Psychol. Bull, 83(6), 1072-1080.

[15] Baker, F. B., and Hubert, L. J. (1975). Measuring the power of hierarchical cluster

analysis. J. Am. Stat. Assoc, 70(349), 31–38.

[16] Beale, E. M. L. (1969). Euclidean cluster analysis. Scientific Control Systems Limited.

[17] Sarle, W. S. (2003). SAS Technical report a-108, cubic clustering criterion. SAS Institute Inc.

[18] Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45(3), 325–342.

[19] Milligan, G. W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis. Psychometrika, 46(2), 187–199.

[20] Rohlf, F. J. (1974). Methods of comparing classifications. Annu. Rev. Ecol. Syst, 101–113.

[21] Davies D. L., and Bouldin, D. W. (1979). A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell, 2(1), 224–227.

[22] Frey, T., and Van Groenewoud, H. (1972). A cluster analysis of the D2 matrix of white spruce stands in Saskatchewan based on the maximum-minimum principle. J. Ecol, 873–886.

[23] Hartigan, J. A. (1975). Clustering algorithms (probability & mathematical statistics). John Wiley & Sons Inc.

[24] Ratkowsky, D. A., and Lance, G. N. (1978). A criterion for determining the number of groups in a classification. Aust. Comput. J, 10(3), 115–117.

[25] Scott, A. J., and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. Biometrics, 387–397.

[26] Marriott, F H. C. (1971). Practical problems in a method of cluster analysis. Biometrics, 501–514.

[27] Ball G. H., and Hall, D. J. (1965). ISODATA, a novel method of data analysis and pattern classification. DTIC Document.

[28] Friedman, H. P., and Rubin, J. (1967). On some invariant criteria for grouping data. J. Am. Stat. Assoc, 62(320), 1159–1178.

[29] McClain, J. O., and Rao, V. R. (1975). Clustisz: A program to test for the quality of clustering of a set of objects. Journal of Marketing Research. JSTOR, 456–460.

[30] Krzanowski, W. J., and Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics, 23–34.

[31] Lebart, L., Piron, A., Labert, M., Morineau, A., and Piron, M. (2000). Statistique exploratoire multidimensionnelle. Dunod.

[32] Hubert, L., and Arabie, P. (1985). Comparing partitions. J. Classif, 2(1), 193–218.

[33] Halkidi, M., Vazirgiannis,M., & Batistakis, Y. (2000). Quality scheme assessment in the clustering process. In European Conference on Principles of Data Mining and Knowledge Discovery.

[34] Halkidi, M., and Vazirgiannis, M. (2001). Clustering validity assessment: Finding the optimal partitioning of a data set. In Proceedings IEEE International Conference on Data Mining.

[35] R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[36] Aitchison, J. (1986). The statistical analysis of compositional data.

[1] Dunn
[2] Milligan and Cooper
[3] Rousseeuw
[4] Sillhouette Coefficient
[5] Tibshirani et al.
[6] Gap Statistic
[7] Separation of Clusters
[8] Compactness of Clusters
[9] Zaremotlagh et al.
[10] Calinski and Harabasz Index
[11] Gamma Index
[12] Beale Index
[13] Cubic Clustering Criterion
[14] Davies and Bouldin Index
[15] Frey Index
[16] Hartigan Criterion
[17] Ratkovsky Index
[18] Scott Index
[19] Marriot Index
[20] Ball Index
[21] Friedman Index
[22] McClain Index
[23] Rubin Index
[24] Krzanowski and Lai Index
[25] Clustering Gain
[26] Dunn Index
[27] Hubert's Γ statistic
[28] Centered Log-Ratio