نشریه علمی پژوهشی «مهندسی معدن» دوره ۱۴، شماره ۴۵، زمستان ۱۳۹۸، صفحه ۶۳ تا ۸۲
Iranian Journal of Mining Engineering (IRJME) Vol 14, No 45, 2020, pp 63-82

کاربرد الگوریتم‌های داده‌کاوی در تشخیص داده‌های ژئوشیمیایی خارج از ردیف چند متغیره

حمید گرانیان[1*]، زهرا خواجه میری[2]

۱. استادیار گروه معدن دانشگاه صنعتی بیرجند، h.geranian@birjandut.ac.ir
۲. کارشناس سازمان صنعت، معدن و تجارت استان خراسان جنوبی، zahra_khajemiri@yahoo.com

(دریافت: ۱۳۹۸/۰۲/۲۴ - پذیرش: ۱۳۹۸/۰۹/۲۵)

چکیده

تشخیص داده‌های خارج از ردیف چند متغیره به کمک الگوریتم‌های داده کاوی یکی از نکات ضروری پیش‌پردازش داده‌های اکتشافات ژئوشیمیایی محسوب می‌شود. در این مقاله چهار الگوریتم برآورد چگالی کرنل (KDE)، ضریب خارج از ردیف بودن محلی (LOF)، OPTICS-OF و SVDD که به ترتیب جزو روش‌های آماری، روش‌های مبتنی بر مجاورت، روش‌های مبتنی بر خوشه‌بندی و روش‌های مبتنی بر دسته‌بندی‌اند، معرفی شده‌اند و در ادامه کاربرد آن‌ها بر روی داده‌های ژئوشیمیایی ورقه ۱:۱۰۰،۰۰۰ روم با ماتریس داده ۴۱×۹۰۲ بررسی شده است. برای این منظور ابتدا روش ilr برای باز کردن سیستم عددی داده‌ها به کار رفته و سپس داده‌ها در بازه صفر تا یک استاندارد شده است. نتایج پیاده شده چهار الگوریتم فوق بر روی مجموعه داده‌های استاندارد شده، نشان می‌دهد که در رویکرد تشخیص نمونه‌های دارای خطا، ۱۰ نمونه که دارای بالاترین احتمال خارج از ردیف بودن‌اند و در هر چهار الگوریتم نیز یکسان‌اند را می‌توان برای بررسی بیشتر به عنوان نمونه‌های انتخابی برای نمونه‌برداری تکراری در نظر گرفت. در رویکرد تشخیص نمونه‌های غیرنرمال، از ۱۵۰ نمونه انتخابی ۷۴/۵ درصد از نمونه‌ها در هر چهار الگوریتم و ۱۶/۱ و ۹/۴ درصد نیز به ترتیب در یک و دو الگوریتم به عنوان داده خارج از ردیف شناسایی شده است. مقایسه نتایج الگوریتم‌های انتخابی با روش کلاسیک فاصله ماهالانوبیتس نشان‌دهنده برتری آن‌ها در هر دو رویکرد است. همچنین پیشنهادی می‌شود از الگوریتم‌های تشخیص داده‌های خارج از ردیف چند متغیره می‌توان برای تعیین نمونه‌برداری‌های تکراری، محاسبه ماتریس موقعیت و پراکندگی در آمار چند متغیره مقاوم پس از حذف داده‌های غیرنرمال و تعیین آنومالی‌های ژئوشیمیایی استفاده کرد.

کلمات کلیدی

داده‌های خارج از ردیف، برآورد چگالی کرنل، ضریب خارج از ردیف بودن محلی، روش OPTICS-OF، روش SVDD، ورقه روم.

* نویسنده مسئول مکاتبات.

۱- مقدمه

در مبحث آمار، داده خارج از ردیف به داده‌ای گفته می‌شود که با دیگر داده‌های هم‌گروه خود فاصله چشمگیری داشته باشد و یا به اصطلاح با دیگر داده‌ها همخوانی نداشته باشد [۹]. بنابراین به داده‌هایی که با یک تابع توزیع (به عنوان مثال تابع توزیع گوسی) تولید شده باشند، داده‌های نرمال یا داده‌های مورد انتظار و برای سایر داده یا داده‌ها واژه غیرنرمال یا آنومال به کار می‌رود [۴ و ۱۹]. در مباحث اکتشافی به داده‌ای که اختلاف معنی‌دار با سایر داده‌ها داشته باشد و ناشی از خطاهای انسانی در اندازه‌گیری، آماده‌سازی، آنالیز و ثبت و یا تعلق نمونه به جوامع آماری متفاوت باشد، داده خارج از ردیف اطلاق می‌شود [۳].

تشخیص داده‌های خارج از ردیف جزو گام‌های ضرروی در پیش‌پردازش داده‌های اکتشافی به ویژه داده‌های ژئوشیمیایی محسوب می‌شود. داده‌های ژئوشیمیایی دارای سیستم عددی بسته و بیشتر ابعاد بالا (به عنوان مثال آنالیز ۴۴ تا ۵۶ عنصره) هستند [۱۵ و ۱۶]. بنابراین استفاده از روش‌های انتقال لگاریتمی (نسبت لگاریتمی افزایشی، alr؛ نسبت لگاریتمی میان مرکز، clr و نسبت لگاریتمی ایزومتریک، ilr) برای تبدیل داده‌ها از سیستم بسته به باز اولین قدم خواهد بود [۱۷ و ۱۸]. تشخیص داده‌های خارج از ردیف چند متغیره به دلیل مخفی شده آن‌ها در میان ابعاد مختلف به روش‌های کلاسیک امکان‌پذیر نیست. بنابراین باید از روش‌های پیچیده‌تر همچون روش‌های داده‌کاوی برای این منظور استفاده کرد. از کارهای انجام شده در این زمینه می‌توان به مقالات فیلزموزر[1] و همکارانش در سال‌های ۲۰۰۵ و ۲۰۱۲ اشاره کرد که از روش‌های آماری مقاوم و نمودارهای یک و دو متغیره برای تشخیص داده‌های خارج از ردیف چند متغیره استفاده کرده‌اند [۱۵ و ۱۶]. همچنین مقالات استفاده از برآوردگرهای مقاوم در تعیین داده‌های خارج از ردیف اکتشافی [۳]، جداسازی داده‌های خارج از ردیف در داده‌های ژئوشیمی محدوده طلای اپی‌ترمال ساری‌گونای [۲] از جمله کارهای انجام شده است.

از آنجا که روش‌های داده‌کاوی کمتر در شناسایی داده‌های خارج از ردیف چند متغیره اکتشافی به کار رفته است، هدف این مقاله معرفی این روش‌ها و بررسی امکان کاربرد آن‌ها است. روش‌های معرفی شده، در سایر شاخه‌های علوم از قبیل کامپیوتر، صنایع، برق و اقتصاد استفاده شده است. در این مقاله کارایی الگوریتم‌های داده‌کاوی و ارایه پیشنهادهای جدید برای کاربرد آن‌ها در مباحث اکتشافی به ویژه بر روی داده‌های ژئوشیمیایی مورد توجه قرار می‌گیرد. برای این منظور از داده‌های ژئوشیمیایی رسوبات آبراهه‌ای یکی از ورقه‌های استان خراسان جنوبی در فاز اکتشافی ناحیه‌ای استفاده می‌شود.

۲- انواع داده‌های خارج از ردیف

نمونه یا مشاهده‌ایی که به طور غیرعادی یا اتفاقی دارای انحراف از سایر داده‌های تحت بررسی باشد، داده خارج از ردیف است. در داده‌کاوی این داده‌ها را به سه گروه داده‌های خارج از ردیف سراسری[2]، داده‌های خارج از ردیف شرطی[3] و داده‌های خارج از ردیف گروهی[4] دسته‌بندی می‌کنند [۱۹ و ۲۳]. داده خارج از ردیف سراسری به داده‌ای اطلاق می‌شود که انحراف قابل توجهی از سایر داده‌ها داشته باشد. این نوع داده، ساده‌ترین نوع داده‌های خارج از ردیف محسوب می‌شود. در مباحث اکتشافی به داده‌ای که ناشی از خطاهای انسانی و یا دستگاهی در ثبت، اندازه‌گیری، آماده‌سازی و آنالیز باشد، می‌توان داده خارج از ردیف سراسری گفت. هدف اکثر روش‌های تشخیص داده‌های خارج از ردیف نیز شناسایی این نوع داده است. نمونه‌ای که خارج از ردیف بودن آن مشروط به شرط یا بافت خاصی باشد، داده خارج از ردیف شرطی گفته می‌شود. به طور مثال سنگ‌های اولترابازیک، غنی‌شدگی نسبت به عناصری از قبیل Ni، Co، Cu و Cr دارند، بنابراین نمونه دارای عیار بالا برای این عناصر و در این نوع سنگ‌ها یک داده نرمال است، در حالی که در سایر سنگ‌ها ممکن است، یک داده آنومال باشد. در این نوع از داده‌ها، شرط و بافت خاص می‌تواند به موضوعاتی مانند زمان و مکان ارتباط داشته باشد که توسط افراد خبره تعیین می‌شود.

یک مجموعه داده‌ها هنگامی یک داده خارج از ردیف گروهی را تشکیل می‌دهند که رفتار جمعی آن‌ها به صورت چشمگیری با کل داده‌ها متفاوت باشد. به طور مثال، داده‌های ژئوشیمیایی معمولا از دو گروه داده با میانگین پایین (داده‌ها یا جامعه زمینه) و داده‌ها با میانگین بالا (داده‌ها یا جامعه آنومالی) تشکیل شده است. بنابراین مجموعه داده‌های آنومال ممکن است به عنوان داده‌های خارج از ردیف گروهی محسوب شوند. برخلاف فرآیند شناسایی داده‌های خارج از ردیف سراسری و شرطی، در تشخیص داده‌های خارج از ردیف گروهی نه تنها رفتار نمونه‌ها به صورت تکی باید بررسی شود، بلکه رفتار

1- Filzmoser

2- Global outliers
3- Conditional outliers
4- Collective outliers

گروهی آن‌ها نیز باید مورد بازبینی قرار گیرد. از این رو برای شناسایی داده‌های خارج از ردیف گروهی نیاز به داشتن دانشی در زمینه روابط بین نمونه‌ها نیز است [۱۹ و ۲۳]. به طور کلی یک مجموعه داده می‌تواند حاوی چندین نوع داده خارج از ردیف باشد. همچنین یک نمونه نیز ممکن است به بیش از یک نوع داده خارج از ردیف متعلق باشد.

۳- روش‌های تشخیص داده‌های خارج از ردیف چند متغیره

روش‌های مختلفی برای شناسایی داده‌های خارج از ردیف وجود دارد. این روش‌ها را می‌توان برحسب استفاده از برچسب نرمال یا خارج از ردیف داشتن تعدادی از نمونه‌ها توسط افراد خبره به روش‌های باناظر (دانش- محور) یا عدم شناخت اولیه از این برچسب‌ها به بی‌ناظر (داده- محور) و یا بینابین این دو دسته روش، روش‌های نیمه‌ناظر تقسیم‌بندی کرد اما مهمترین تقسیم‌بندی بر اساس استفاده از فرضیات در مورد داده‌های خارج از ردیف در مقابل داده‌های نرمال است. بر این مبنا روش‌های تشخیص داده‌های خارج از ردیف به چهار گروه روش‌های آماری، روش‌های مبتنی بر مجاورت، روش‌های مبتنی بر خوشه‌بندی و روش‌های مبتنی بر دسته‌بندی تفکیک می‌شوند [۴، ۱۹ و ۲۳]، که در ادامه کلیات این گروه‌ها توضیح داده می‌شود.

۳-۱- روش‌های آماری

در این روش‌ها فرض می‌شود که داده‌ها بیشتر نرمال‌اند و نمونه‌ها با یک مدل مولد و به صورت فرآیندی تصادفی تولید شده‌اند. در نتیجه نمونه‌های نرمال در نواحی با احتمال بالا و نمونه‌های خارج از ردیف در نواحی با احتمال پایین قرار دارند. بنابراین برای تشخیص داده‌های خارج از ردیف ابتدا باید مدل مولدی که با مجموعه داده‌ها مطابقت داشته باشد را با الگوریتم‌های یادگیری مشخص کرد و سپس نمونه‌هایی که در نواحی با احتمال پایین این مدل قرار دارند را به عنوان داده خارج از ردیف معرفی کرد. روش‌های آماری را می‌توان به دو بخش روش‌های پارامتری و روش‌های بدون پارامتر تقسیم‌بندی کرد. روش‌های مبتنی بر پارامترهای آماری، فاصله ماهالانوبیس و برآوردگرهای مقاوم جزو روش‌های پارامتری و روش هیستوگرام و برآورد چگالی کرنل جزو روش‌های بدون پارامتر محسوب می‌شوند. برای آشنایی با روش‌های پارامتری می‌توان به مقاله گرانیان و خواجه‌میری (۱۳۹۶) مراجعه کرد. در این مقاله روش برآورد چگالی کرنل[1] معرفی می‌شود.

$$D = \{x_1, x_2, \dots, x_n\} \quad (۱)$$

که $x_i \in R^d$ یک نمونه یا داده، n تعداد نمونه‌ها و d تعداد متغیرها یا بعد داده‌ها است. تقریب چگالی کرنل با استفاده از تابع چگالی احتمال مدل آماری مجموعه داده‌ها برابر است با [۲۰ و ۳۰]:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(x_i)^d}K\left(\frac{x-x_i}{h(x_i)}\right) \quad (۲)$$

که $h(x_i)$ پهنای باند بر روی نمونه‌ی x_i است و معمولا برای ساده‌سازی برابر مقدار ثابت در نظر گرفته می‌شود (یعنی $h(x_i) = h$). کمیت h به عنوان یک پارامتر هموارساز عمل می‌کند. $K(u)$ تابع کرنل انتگرال‌پذیر با مقدار واقعی و غیر منفی است که برای تمام مقادیر u دارای دو شرط زیر نیز است [۱۹ و ۲۰]:

$$\int_{-\infty}^{+\infty} k(u)du = 1 \; and \; K(-u) = K(u) \quad (۳)$$

توابع Gaussian، Laplacian و Epanetchnikov مهم‌ترین توابع کرنل‌اند که تابع گوسی چندمتغیره استاندارد با میانگین صفر و واریانس ۱ بیشترین کاربرد را به عنوان تابع کرنل دارد. بنابراین طبق رابطه ۴:

$$k(u) = \frac{1}{2\pi^{\frac{d}{2}}}\exp(-\frac{\|u\|^2}{2}) \quad (۴)$$

با ترکیب روابط ۴ و ۲ تقریب چگالی کرنل برای نمونه x_j معادل رابطه ۵ است:

$$\hat{f}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\pi^{\frac{d}{2}}h^d}\exp\left(-\frac{\|x_j - x_i\|^{\Delta}}{2h^2}\right) \quad (۵)$$

j=1,2,…,n

از آنجا که برای شناسایی داده خارج از ردیف نیاز به محاسبه و مقایسه چگالی کرنل نمونه با نمونه‌های موجود در همسایگی آن است، بنابراین رابطه ۵ را می‌توان برای k نمونه موجود در نزدیکترین همسایگی (k- NN) به صورت رابطه ۶ نوشت [۲۰]:

$$\hat{g}(x_j) = \frac{1}{k}\sum_{x_i \in k-NN(x_j)}\frac{1}{2\pi^{\frac{d}{2}}h^d}\exp\left(-\frac{d(x_j, x_i)^2}{2h^2}\right) \quad (۶)$$

که $d(x_j, x_i)$ فاصله اقلیدسی بین نمونه x_j با x_i است. در صورتی که تقریب چگالی کرنل برای نمونه x_j بالا باشد، داده نرمال و در غیر این صورت داده خارج از ردیف محسوب می‌شود.

1- Kernel Density Estimation) KDE)

قبل از برآورد چگالی کرنل یک نمونه، دو پارامتر پهنای باند h و تعداد نمونه‌های نزدیکترین همسایگی k باید توسط کاربر تعیین شود. برای محاسبه پهنای باند از رابطه‌های ۷ و ۸ استفاده می‌شود و مقدار k نیز معمولا بین ۲ تا ۱۰ درصد تعداد کل نمونه‌ها در نظر گرفته می‌شود [۲۱ و ۲۵].

$$h = \sqrt{\prod_{i=1}^{d} h_i^2} \qquad (۷)$$

$$h_i = 0.9 \times \min\left\{\sigma_i, \frac{Q_{i3} - Q_{i1}}{1.34}\right\} \times n^{-\frac{1}{5}} \qquad (۸)$$

i=1,2,…,d

به منظور سادگی محاسبات در فرمول برآورد چگالی کرنل از رابطه ۷ به جای محاسبه دترمینال ماتریس قطری پهنای باند استفاده شده است، که σ_i انحراف معیار متغیر iام و Q_{i3} و Q_{i1} نیز به ترتیب چارک سوم و اول متغیر iام است.

۳-۲- روش‌های مبتنی بر مجاورت

در این روش‌ها، داده خارج از ردیف به داده‌ای گفته می‌شود که در مجاورت آن نمونه کمی وجود داشته باشد، یا اصلا نمونه‌ای موجود نباشد. بنابراین مجاورت یک نمونه خارج از ردیف با نزدیکترین همسایگانش به صورت چشمگیری متفاوت از مجاورت یک نمونه با اکثر نمونه‌های دیگر در مجموعه داده‌ها است [۱۹]. برای تعیین میزان مجاورت می‌توان از دو معیار فاصله یا چگالی استفاده کرد. در روش مبتنی بر فاصله، نمونه‌هایی که فاصله آن‌ها از بقیه نمونه‌ها بیشتر باشد را می‌توان به عنوان داده خارج از ردیف در نظر گرفت. برای این منظور دو پارامتر شعاع همسایگی r $(r \geq 0)$ و یک حد آستانه نسبی p $(0 < p \leq 1)$ در نظر گرفته می‌شود. بنابراین نمونه x یک داده خارج از ردیف است اگر (رابطه ۹) [۶]:

$$\frac{\left\|\left\{x' \middle| dist(x, x') \leq r\right\}\right\|}{\|D\|} \leq p \qquad (۹)$$

که $dist()$ معیار فاصله است. در روش تشخیص داده‌های خارج از ردیف مبتنی بر چگالی، چگالی اطراف یک نمونه نرمال، شبیه چگالی نمونه‌های همسایه‌هایش است. در حالی که چگالی اطراف نمونه خارج از ردیف به صورت معنی‌داری با چگالی همسایگانش متفاوت است. بنابراین تعداد نمونه‌های اطراف یک داده نشان‌دهنده ضریب خارج از ردیف بودن آن است. تشخیص داده‌های خارج از ردیف چند متغیره به کمک الگوریتم ضریب خارج از ردیف بودن محلی اولین بار توسط بروئینگ و همکارانش در سال ۲۰۰۰ ارایه شده است [۱۱].

فاصله نمونه x از نمونه $p \in D$ به صورت $dist_k(x)$ تعریف می‌شود به شرطی که:

۱- حداقل k نمونه مانند $x' \in D\text{-}\{x\}$ وجود داشته باشد به طوری که $dist(x, x') \leq dist(x, p)$

۲- حداکثر k-1 نمونه مانند $x' \in D\text{-}\{x\}$ وجود داشته باشد به طوری که $dist(x, x') \leq dist(x, p)$

بنابراین عبارت $\mathrm{dist_k(x)}$ حداکثر فاصله نمونه x از k همسایه نزدیک این نمونه خواهد بود. تعداد نمونه‌های که فاصله آن‌ها از این مقدار کمتر است برابر $\mathrm{N_k(x)}$ و به صورت رابطه ۱۰نمایش داده می‌شود [۱۹]:

$$N_k(x) = \left\{x' \middle| x' \in D, dist(x, x') \leq dist_k(x)\right\} \qquad (۱۰)$$

چون ممکن است چندین نمونه فاصله یکسان از نمونه x داشته باشند، بنابراین $\mathrm{N_k(x)} \geq k$ خواهد بود. میانگین فاصله نمونه‌های متعلق به مجموعه $\mathrm{N_k(x)}$ را می‌توان به عنوان سنجه چگالی محلی برای نمونه x در نظر گرفت ولی از آنجا که بعضی از نمونه‌ها ممکن است در فاصله بسیار نزدیک نمونه قرار داشته باشند، بنابراین میانگین‌گیری باعث تغییرات زیاد در سنجه چگالی محلی خواهد شد. به عبارت دیگر پارامتر آماری میانگین نسبت به کرانه‌ها در مجموعه داده‌ها پایا یا مقاوم[1] نیست. برای رفع این مشکل می‌توان از سنجه فاصله قابل دسترسی[2] استفاده کرد که یک پارامتر برای هموارسازی است. برای دو نمونه x و x' فاصله قابل دسترسی برابر است با (رابطه ۱۱) [۱۱ و ۱۹]:

$$reachdist_k(x \leftarrow x') = \max\left\{dist_k(x), dist(x, x')\right\} \qquad (۱۱)$$

فاصله قابل دسترسی، سنجه‌ای است که خاصیت تقارن ندارد به طوری که $\mathrm{reachdist_k(x \leftarrow x')}$برابر با $reachdist_k(x \leftarrow x')$ نیست. حال می‌توان چگالی قابل دسترسی محلی[3] برای نمونه x را به صورت رابطه ۱۲ تعریف کرد [۲۲]:

$$lrd_k(x) = \frac{\|N_k(x)\|}{\sum_{x' \in N_k(x)} reachdist_k(x' \leftarrow x)} \qquad (۱۲)$$

که عبارت $\|N_k(x)\|$ برابر نرم یا تعداد نمونه‌های داخل مجموعه است. ضریب خارج از ردیف بودن محلی[4] (LOF) برای نمونه x نیز برابر رابطه ۱۳ است [۱۱ و ۱۹]:

1- Robust
2- Reachable distance
3- Local reachability Density
4- Local Outlier Factor

$$\sum_{x N_K(x)} lr(x').\sum_{x'\in N_k(x)} reachdist_k(x'\leftarrow x) \qquad (۱۳)$$

ضریب خارج از ردیف بودن محلی برابر میانگین نسبت چگالی قابل دسترسی محلی نمونه x از k همسایه نزدیک آن است. بالا بودن ضریب خارج از ردیف بودن محلی نشان‌دهنده مقدار کمتر چگالی قابل دسترسی محلی نمونه x و همچنین مقدار بیشتر چگالی‌های قابل دسترسی محلی k همسایه نزدیک نمونه x است. این نکته دلالت بر یک داده خارج از ردیف محلی دارد، چون چگالی محلی نمونه در مقایسه با چگالی محلی نمونه‌های k همسایه نزدیک به آن کوچکتر است [۱۱ و ۱۳]. مقدار ضریب خارج از ردیف بودن محلی برای داده‌های نرمال حدود ۱ و برای داده‌های خارج از ردیف بیشتر از ۱ خواهد بود. به طور کلی حد پایین و بالایی برای LOF را به صورت رابطه ۱۴ می‌توان تعریف کرد [۱۱ و ۱۳]:

$$\frac{direct_{\min}(x)}{indirect_{\max}} \le LOF \le \frac{direct_{\max}(x)}{indirect_{\min}(x)} \qquad (۱۴)$$

که در آن:

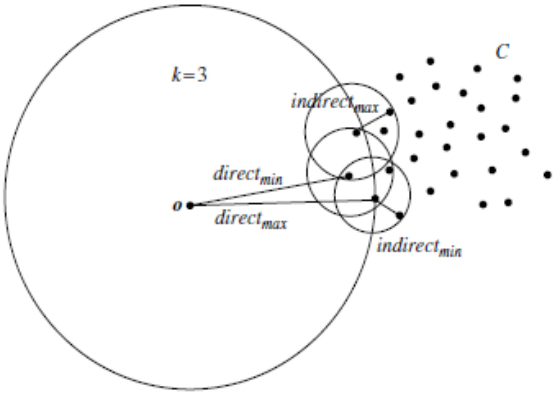
$direct_{\min}(x)$ حداقل فاصله قابل دسترس مستقیم نمونه x از k همسایه نزدیک آن

$direct_{\max}(x)$ حداکثر فاصله قابل دسترس مستقیم نمونه x از k همسایه نزدیک آن

$indirect_{\min}(x)$ حداقل فاصله قابل دسترس غیرمستقیم نمونه x از k همسایه نزدیک آن

$indirect_{\max}(x)$ حداکثر فاصله قابل دسترس غیرمستقیم نمونه x از k همسایه نزدیک آن

در شکل ۱ این پارامترها تعریف شده‌اند.



شکل ۱- تعریف فاصله قابل دسترسی مستقیم و غیرمستقیم در الگوریتم LOF [۱۹].

۳-۳- روش‌های مبتنی بر خوشه‌بندی

در روش‌های خوشه‌بندی، نمونه‌ها به چندین دسته یا خوشه تقسیم‌بندی می‌شوند، به نحوی که اشیا در هر خوشه بسیار به هم شبیه بوده و بین خوشه‌ها نیز کمترین شباهت وجود داشته باشد. بر این مبنا داده خارج از ردیف، نمونه‌ای است که به هیچ خوشه‌ای تعلق ندارد و یا متعلق به یک خوشه کوچک و دور باشد [۲۶]. روش‌های خوشه‌بندی به چهار گروه روش‌های مبتنی بر گرانیگاه، روش‌های بر پایه اتصال، روش‌های بر پایه توزیع داده‌ها و روش‌های مبتنی بر چگالی تفکیک می‌شوند [۱۹]. در این مقاله روش OPTICS که جزو روش‌های مبتنی بر چگالی است، به دلیل بیشترین کاربرد در تشخیص داده‌های خارج از ردیف و امکان مقایسه نتایج آن با روش‌های تشخیص داده خارج از ردیف مبتنی بر مجاورت انتخاب شده است.

روش نظم‌دهی نقاط برای شناسایی ساختار خوشه‌بندی[1] (OPTICS) اولین بار توسط آنکرست و همکارانش ارایه شده است [۸]. برای خوشه‌بندی داده‌ها و سپس شناسایی داده‌های خارج از ردیف در این روش از تعاریف زیر استفاده می‌شود [۸، ۱۲ و ۱۹]:

الف- شعاع همسایگی (ε): به فضای دایره‌ای شکل به مرکز نمونه x و شعاع ε در فضای ابر داده‌های اطلاق می‌شود.

ب- حد آستانه تراکم ($MinPts$): به تعداد نمونه‌هایی که در همسایگی یک نمونه قرار داشته باشد.

پ- نمونه هسته[2]: چنانچه در شعاع همسایگی ε از یک شی، بتوان حداقل تعداد $MinPts$ نمونه پیدا کرد، آن شی به عنوان شی هسته شناخته می‌شود.

ت- قابلیت دسترسی مستقیم و غیرمستقیم[3]: اگر x نمونه هسته و x' در شعاع همسایگی آن باشد، می‌گویند x' به صورت مستقیم از x قابل دسترس است اگر و تنها اگر x یک نمونه هسته و x' نیز در شعاع همسایگی ε از x باشد. همچنین اگر x و x' نمونه‌های هسته و x'' نمونه‌ی غیرهسته باشد، به طوری که x' از x به صورت مستقیم و x'' از x' نیز به صورت مستقیم قابل دسترس باشند، می‌گویند x'' از x به طور غیرمستقیم قابل دسترس است.

ث- متصل شده با چگالی[4]: x و x' نمونه‌های متصل شده با چگالی‌اند، هرگاه هر دو از نمونه دیگری مانند x'' قابل

1- Ordering points to identify the clustering structure
2- Core- object
3- Direct and indirect-reachable
4 -Density-connected

دسترس باشند.

ج- فاصله هسته‌ای[1]: به کوچکترین مقدار از ε' اطلاق می‌شود که در شعاع همسایگی ε' از نمونه x بتوان حداقل تعداد $MinPts$ نمونه پیدا کرد (رابطه ۱۵):

(۱۵)

$$core-dis\tan ce_{\varepsilon,Minpts}(x)=\begin{cases}if\ \|N_{\in}(x)\|<\min Pts\\ \in'\end{cases}\}$$

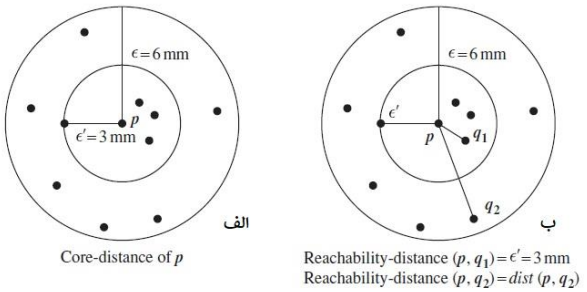
چ- فاصله قابل دسترس[2]: فاصله قابل دسترس نمونه x' از x برابر حداقل شعاعی است که باعث می‌شود، x' از x قابل دسترس باشد (رابطه ۱۶)

(۱۶)

$$reachability-dis\tan ce_{\varepsilon,MinPts}(x,x')=\begin{cases}f\ \|N_{\varepsilon}(x)\|<MinPts\\ \max\{core-dis\tan ce(x),dist(x,x')\}\end{cases}\}$$

اگر نمونه از چندین نمونه هسته به طور مستقیم قابل دسترس باشد، کوچکترین مقدار فاصله قابل دسترس ملاک عمل خواهد بود. در شکل ۲- الف و ۲- ب به ترتیب فاصله هسته‌ای و فاصله قابل دسترس نشان داده شده است.

استفاده از الگوریتم خوشه‌بندی OPTICS برای شناسایی داده‌های خارج از ردیف اولین بار توسط بروئینگ و همکارانش در سال ۱۹۹۹ مطرح شد که سپس توسط محققان دیگر این الگوریتم توسعه داده شد [۱۰، ۱۲، ۱۴ و ۲۹]. برای این منظور ابتدا با استفاده از تعاریف فوق در مرحله اول فاصله قابل دسترس هر نمونه محاسبه و سپس در مرحله بعد چگالی قابل دسترسی محلی (*lrd*) و ضریب خارج از ردیف بودن[3] (OF) نیز به کمک رابطه‌های ۱۲ و ۱۳ محاسبه می‌شود. با این تفاوت که در در فرمول‌های یاد شده پارامتر *MinPts* جایگزین پارامتر k خواهد شد. ضریب خارج از ردیف بودن نشان‌دهنده درجه خارج از ردیفی نمونه است و مطابق فرمول ۱۳ برابر میانگین نسبت چگالی قابل دسترس *MinPts*- نزدیکترین نمونه‌ها به نمونه مورد نظر است. اگر یک خوشه یکنواخت باشد ضریب خارج از ردیف بودن نمونه‌ای آن خوشه برابر ۱ خواهد بود و اگر چگالی قابل دسترس نمونه‌ای نصف چگالی قابل دسترس *MinPts*- نزدیکترین نمونه‌ها به آن باشد، ضریب خارج از ردیف بودن برابر ۲ خواهد بود. بنابراین هر چه ضریب خارج از ردیف بودن نمونه‌ای بالاتر باشد، احتمال خارج از ردیف بودن نمونه بیشتر است [۱۲ و ۱۴].



شکل ۲- تعریف فاصله‌ی هسته‌ای و فاصله‌ی قابل دسترس با فرض $\varepsilon = 6\text{mm}$ و $\begin{matrix}MinPts=5\\ \varepsilon=6mm\end{matrix}$ [۱۹] MinPts = 5.

۳-۴- روش‌های مبتنی بر دسته‌بندی

در روش‌های دسته‌بندی از نمونه‌های آموزشی که دارای برچسب نرمال یا خارج از ردیف بودن‌اند، استفاده و مدل تشخیص ساخته می‌شود. از آنجا که تعداد داده‌های خارج از ردیف بسیار کمتر از داده‌های نرمال است، معمولا مدل تک کلاسه برای تشخیص داده‌های خارج از ردیف به کار می‌رود. بنابراین مدل ساخته شده برای توصیف داده نرمال خواهد بود و هر نمونه‌ای که متعلق به کلاس نرمال نباشد، به عنوان داده خارج از ردیف محسوب خواهد شد. در این مقاله روش ماشین بردار پشتیبان معرفی می‌شود که کاربرد بیشتری در تشخیص داده خارج از ردیف دارد.

روش دسته‌بندی تک کلاسه ماشین بردار پشتیبان (OC-SVM) اولین بار توسط اسچولکف و همکارانش در سال ۱۹۹۹ ارایه شد [۲۴]. در این روش مبدا مختصات به عنوان تنها عضو کلاس غیرهدف یعنی داده خارج از ردیف در نظر گرفته می‌شود. سپس با یافتن یک ابرصفحه که حداکثر فاصله را از مبدا مختصات دارد و داده‌های نرمال را از بخش غیرداده یا داده‌های خارج از ردیف جدا می‌کند (شکل ۳- الف)، مساله قابل حل خواهد بود. بنابراین معادله بهینه‌سازی به صورت رابطه ۱۷ مطرح می‌شود [۲۴ و ۷]:

$$\min_{w,\rho,\xi}\frac{1}{2}\|w\|^2-\rho+\frac{1}{vl}\sum_{i=1}^{l}\xi_i \qquad (۱۷)$$

مشروط به قیود زیر (رابطه ۱۸):

$$\varphi(x_i)w\ge\rho-\xi_i\quad \xi_i\ge 0,\forall i, \qquad (۱۸)$$

که در آن:

$\varphi(x_i)$ تابع کرنل برای نگاشت داده‌ها به فضایی با بعد بزرگتر از بعد داده‌های اولیه

1- Core-distance
2- Reachability-distance
3- Outlier factor

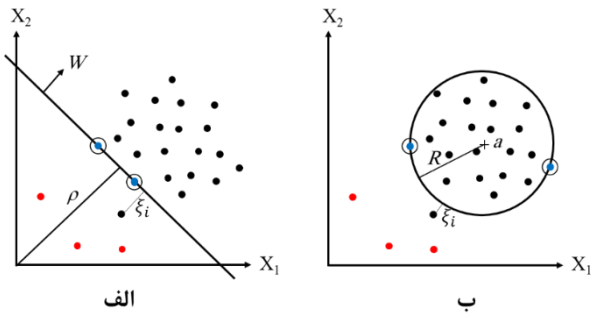
ξ_i پارامتر کمکی و برابر مقدار ضرر حاصل از نمونه‌های خارج از محدود

l تعداد نمونه‌های آموزشی

ν پارمتر تعدیل یا نرخ عدم پذیرش دسته‌بندی که کنترل کننده مصالحه بین پیچیدگی مدل و خطای آموزش و برابر $\nu \in (0,1]$ است.

w و ρ مرتبط با تابع تصمیم یا هدف و به ترتیب برابر بردارهای یکه عمود بر ابرصفحه و فاصله ابرصفحه از مبدا مختصات

برای بدست آوردن پارامترهای تابع هدف می‌توان از تابع لاگرانژین دوگانه ولف با حفظ شرایط Karush-Kuhn-Tucker (KKT) استفاده کرد [۷].



شکل ۳- ابرصفحه (الف) و ابرکره (ب) جداکننده داده‌های نرمال از داده‌های خارج از ردیف در روش بردار پشتیبان؛ نقاط سیاه داده‌های نرمال، نقاط قرمز داده‌های خارج از ردیف و نقاط آبی داده‌های بردار پشتیبان.

روش توصیف حوزه بردار پشتیبان[1] مدل دیگری از دسته‌بندی تک کلاسه ماشین بردار پشتیبان است که توسط تکس و دوین در سال ۱۹۹۹ ارایه شده است [۲۷]. در این روش به جای استفاده از یک ابرصفحه از یک ابرکره با حداقل شعاع استفاده می‌شود که داده‌های نرمال داخل ابرکره و داده‌های خارج از ردیف خارج آن قرار می‌گیرند (شکل ۳- ب). بنابراین معادله بهینه‌سازی و قیود آن به صورت رابطه ۱۹ تعریف می‌شود [۲۷ و ۳۱]:

$$\begin{aligned} &min_{R,a,\xi}\ R^2 + C\sum_{i=1}^{l}\xi_i \\ &Subject\ to\ \|\varphi(x_i) - a\|^2 \le R^2 + \xi_i \\ &\xi_i \ge 0\ ,\forall i \end{aligned} \qquad (۱۹)$$

که در آن:

C پارامتر کنترل‌کننده تعادل بین شعاع ابرکره و خطای آموزش

R و a پارامترهای تابع هدف و به ترتیب شعاع و مختصات مرکز ابرکره

در این مدل نیز پارامترهای تابع هدف با الگوریتم تابع لاگرانژین دوگانه ولف با حفظ شرایط KKD به صورت رابطه ۲۰ به دست می‌آید.

$$\begin{aligned} min_{\alpha}\ L(\alpha) = &\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j < \varphi(x_i), \varphi(x_j) \\ &> -\sum_{i=1}^{l}\alpha_i < \varphi(x_i), \varphi(x_i) \\ &> \end{aligned} \qquad (۲۰)$$

مشروط به قید (رابطه ۲۱):

$$\sum_{i=1}^{l}\alpha_i = 1\ 0 \le \alpha_i \le 1 \qquad (۲۱)$$

در صورت استفاده از توابع کرنل RBF به فرمول زیر نتایج به دست آمده از هر دو مدل یکسان است (رابطه ۲۲).

$$\varphi(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \qquad (۲۲)$$

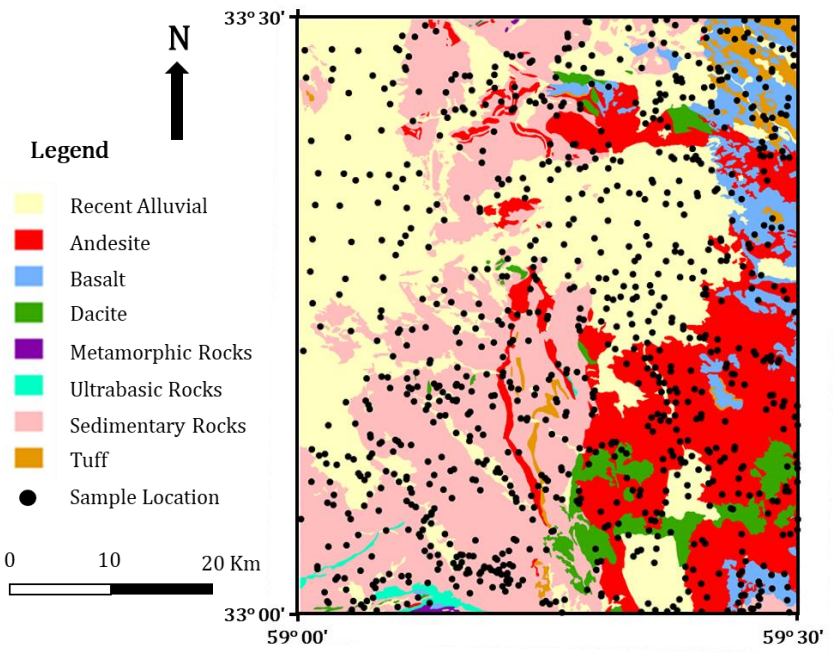
در رابطه‌های بالا، پارامترهای ν، C و σ معمولا با روش‌های سعی و خطا به دست می‌آیند.

۴- مثال موردی

۴-۱- معرفی داده‌ها

نتایج آنالیز حاصل از نمونه‌های ژئوشیمیایی برداشت شده از رسوبات آبراهه‌ای در ورقه ۱:۱۰۰،۰۰۰ روم به عنوان مثال موردی استفاده شده است. ورقه روم در استان خراسان جنوبی و در چهارگوشه قائن قرار دارد. از این ورقه ۹۰۲ نمونه برداشت شده و هر نمونه نیز برای ۴۱ عنصر به روش ICP-OES و طلا به روش Fire Assay آنالیز شده است. نمونه‌برداری توسط سازمان زمین‌شناسی و اکتشاف معدنی ایران و آنالیزها در آزمایشگاه این سازمان انجام شده است. در شکل ۳ الگوی پراکندگی نمونه نشان داده شده است. منطقه مطالعاتی را بیشتر رسوبات آبرفتی متعلق به دوره کواترنر و سنگ‌های رسوبی با سن کرتاسه تا پالئوژن پوشش داده است. جنس سنگ‌های رسوبی بیشتر از نوع کنگلومرا، ماسه‌سنگ، شیل، مارن و آهک است. سنگ‌های آذرین حد واسط آندزیت و داسیت و بازی از نوع بازالت با سن نئوژن در شرق محدوده مطالعاتی بر روی واحد رسوبی قرار گرفته است. واحد توفی نیز در بخش شمال شرقی محدوده قابل مشاهده است. سنگ‌های اولترابازیک از نوع سرپانتینیت، پریدوتیت، گابرو و دیاباز و سنگ‌های دگرگونی از جنس شیست به صورت بسیار محدود در بخش جنوب غربی محدوده مطالعاتی و با سن پرمین قدیمترین واحدهای سنگ منطقه مطالعاتی را تشکیل می‌دهند.

1- Support Vector Domain Description (SVDD)



شکل ۴- موقعیت نمونه‌های ژئوشیمیایی بر روی نقشه زمین‌شناسی ساده شده محدوده مطالعاتی.

جدول ۱- پارامترهای آماری نتایج آنالیزهای شیمیایی (واحدها برحسب ppm است).

عناصر	میانگین	انحراف معیار	حداقل	حداکثر	عناصر	میانگین	انحراف معیار	حداقل	حداکثر
Al	۷۱۱۱۶	۱۲۵۳۸	۲۰۵۵۷	۱۱۲۷۴۸	Nb	۱۲/۱۹	۲/۴۵	۴/۹۲	۲۴
As	۱۲/۸۲	۵۲/۶۲	۱/۷۰	۱۵۸۲/۱۸	Ni	۶۹/۹۰	۱۴۲/۷۰	۱	۳۷۳۱
Au	۰/۰۰۱۴	۰/۰۰۳۳	۰/۰۰۰۵	۰/۰۴	P	۶۸۸/۵۵	۱۷۴/۵	۵۱/۱۱	۱۷۵۹/۳۹
Ba	۴۲۲	۱۳۴/۶۵	۱۶۳/۸۸	۲۱۵۱/۲	Pb	۱۶/۷	۱۵/۰۹	۱/۸۵	۳۵۷/۲۲
Be	۱/۲۳	۰/۳۰	۰/۴۷	۸	Rb	۶۹/۴۴	۱۳/۷۴	۲۰	۲۵۳/۶۸
Bi	۰/۵۳	۱/۳۶	۰/۱	۴۰/۹۳	S	۳۶۴	۱۴۴۸/۶	۲۹/۳	۲۱۴۲۲/۱
Ca	۶۹۷۴۶	۱۶۳۹۱	۷۶۵۶	۱۱۸۶۱۸	Sb	۰/۹۱	۱/۸۶	۰/۰۲	۴۷/۶۱
Cd	۰/۱۹	۰/۴۷	۰/۰۵	۱۳/۹۴	Sc	۹/۶۲	۱/۵۰	۲/۵۳	۲۱/۲۴
Ce	۵۴/۷۴	۱۷/۱۱	۷/۶۳	۱۱۵	Sn	۱/۹۷	۰/۶۱	۰/۲۳	۵/۴
Co	۱۳/۱۳	۴/۸۳	۳/۷۴	۱۲۴/۰۲	Sr	۳۴۶/۲۳	۶۵/۱۵	۲۹/۸۷	۱۰۱۶/۴۷
Cr	۸۷/۷۱	۱۲۰/۵	۱۶/۶۷	۳۳۹۲/۱	Th	۸/۲۸	۱/۷۸	۲/۷۳	۱۶/۷۱
Cs	۶/۷۹	۱/۸۷	۱/۷۲	۱۶	Ti	۴۰۴۲/۱	۷۲۳	۱۳۱۳/۹	۷۴۶۱/۸
Cu	۳۲/۰۷	۱۴۹/۸۷	۳	۴۳۵۷/۵۲	Tl	۰/۶۱	۰/۲۸	۰/۱	۵/۱۹
Fe	۲۷۷۹۳	۴۴۴۰	۲۲۴۳	۵۶۹۰۱	U	۲/۳۵	۰/۴۴	۱/۱۹	۴/۹۶
K	۱۶۹۱۰	۳۰۰۹	۳۲۷۱	۲۶۱۰۲	V	۸۳/۸۷	۱۷/۶۷	۲۶/۲۸	۲۳۹/۹۹
La	۳۹/۵۶	۸/۲۵	۱۴	۶۸	W	۰/۵۸	۱/۴۱	۰/۰۲	۳۴/۶۳
Li	۳۳/۱۱	۷/۱۹	۶/۳۴	۶۷/۷۸	Y	۲۰/۲۲	۲/۶۸	۶/۲۱	۴۴/۷۶
Mg	۱۵۴۵۳	۲۲۱۸	۵۷۷۲	۳۲۲۲۱	Yb	۲/۴۶	۰/۳۲	۰/۶۹	۳/۵۲
Mn	۶۸۲/۲۳	۱۰۵/۹۱	۳۲۸/۹	۱۴۴۴	Zn	۶۳/۲۷	۴۶/۴۳	۱۱	۱۳۸۰
Mo	۰/۶۷	۱/۰۹	۰/۰۴	۳۰/۵۸	Zr	۳۶۴/۷۲	۷۴/۵۰	۶۱/۰۴	۶۸۷/۱۰
Na	۱۷۷۶۱	۳۹۵۲	۳۶۱۹	۴۰۹۸۵					

در جدول ۱ پارامترهای آماری حاصل از نتایج آنالیزهای شیمیایی پس از جایگزینی داده‌های سنسورد با ۳/۴ حد تشخیص آمده است. با توجه به ماهیت بسته بودن داده‌های حاصل از نتایج آنالیزهای ژئوشیمیایی، برای پردازش آماری نیاز به تبدیل داده‌ها از سیستم بسته به باز است. برای این منظور از روش تبدیل نسبت لگاریتمی ایزومتریک (ilr) استفاده شده است. اگر چه در این تبدیل ابعاد ماتریس داده‌ها از ۴۱ × ۹۰۲ به ۴۰ × ۹۰۲ کاهش می‌یابد ولی عدم تکینه‌شدن ماتریس واریانس- کواریانس و ساختار هندسی بهتر ماتریس داده‌ها از مزایایی این تبدیل است [۱ و ۱۶]، سپس برای استاندارد کردن ماتریس داده‌ها، بازه هر متغیره به دامنه صفر تا یک انتقال داده شده است [۳۲]. در نهایت الگوریتم‌های تشخیص داده‌های خارج از ردیف چند متغیره بر روی ماتریس داده‌های استاندارد شده پیاده شده است که در ادامه نتایج آن ارایه می‌شود.

۴-۲- تشخیص داده‌های خارج از ردیف

روش فاصله ماهالانوبیتس (یعنی فاصله هر داده از مرکز ابر داده‌ها) یکی از روش‌های کلاسیک تشخیص داده‌های خارج از ردیف محسوب می‌شود. این فاصله از رابطه ۲۳ به دست می‌آید [۳]:

$$MD(x_i) = \sqrt{(x_i - \mu)' \sum^{-1} (x_i - \mu)} \qquad (۲۳)$$

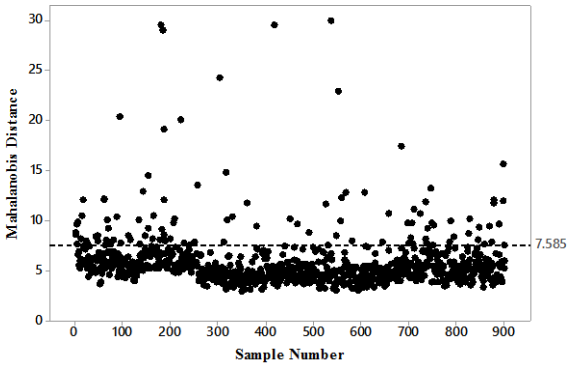
i=1,2,…,n

که در آن:

μ ماتریس میانگین

$\sum$ ماتریس واریانس-کواریانس داده‌ها

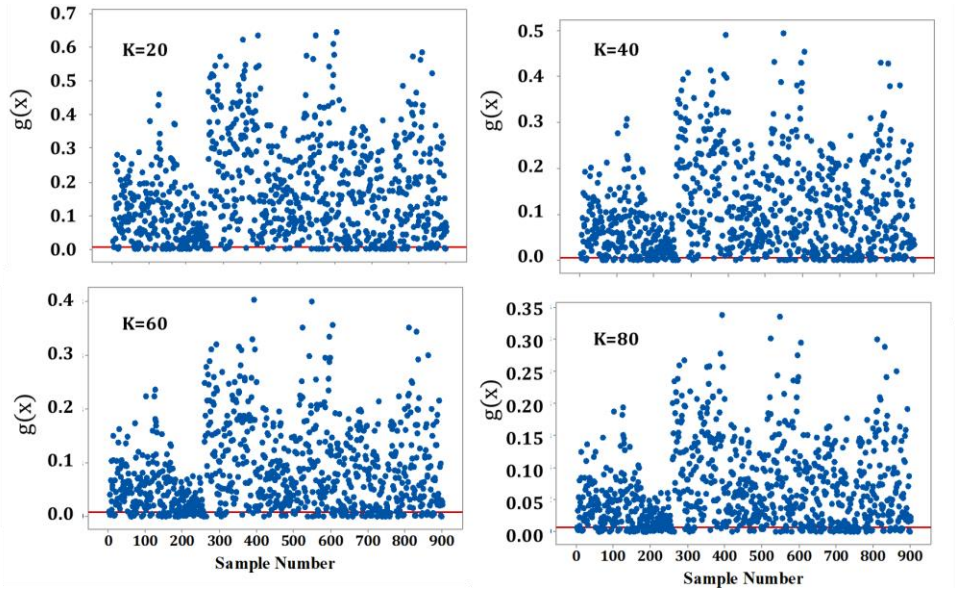
اگر فاصله ماهالانوبیتس (MD) نمونه‌ای از مقدار $\sqrt{\chi^2_{d,0.975}}$ بیشتر باشد، نمونه داده خارج از ردیف خواهد بود. مقدار زیر رادیکال برابر چندک $1-\alpha$ توزیع خی‌دو با درجه آزادی d است. در شکل ۵ نمودار فاصله ماهالانوبیتس داده‌های اولیه و خط تشخیص نمونه‌های خارج از ردیف نشان داده شده است. مطابق شکل، ۱۰۴ نمونه در بالای خط قرار دارند که داده خارج از ردیف خواهند بود.



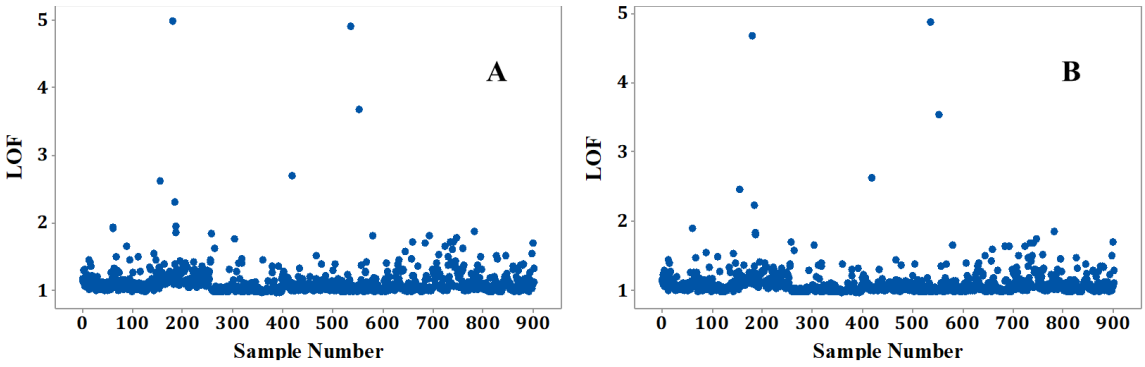
شکل ۵- فاصله ماهالانوبیتس نمونه‌های منطقه مطالعاتی.

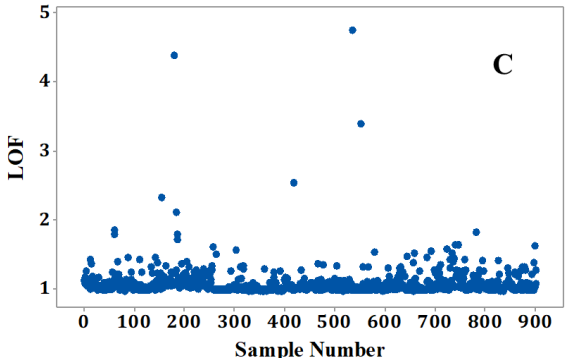
برای تشخیص داده‌های خارج از ردیف به کمک الگوریتم KDE، پارامتر h با روابط ۷ و ۸ برآورد و برابر ۰/۰۱۷۳ به دست آمده است. از آنجا که تعداد نمونه‌های نزدیکترین همسایگی ممکن است بین ۲ تا ۱۰ درصد کل نمونه‌ها باشد، مقدار KDE هر نمونه به ازای چهار عدد انتخابی k=20، k=40، k=60 و k=80 محاسبه شده است. همچنین مقدار امید ریاضی $E\left[f(\hat{x}_j)\right]$ نیز برای مجموعه داده‌های با رابطه ۵ محاسبه و برابر ۱۱ ۱۰×۶/۸۴۸۱ به دست آمده است. در شکل ۶ مقادیر KDE هر نمونه نشان داده شده است که با رابطه ۶ محاسبه شده است. نمونه‌ها با مقدار $\hat{g}(x_j)$ بالا (نمونه‌های بالای خط قرمز در شکل ۶ که برابر امید ریاضی $E\left[f(\hat{x}_j)\right]$ است) داده نرمال و نمونه‌های زیر خط داده خارج از ردیف محسوب می‌شوند. مطابق شکل ۶ تعداد داده‌های خارج از ردیف برای k=20 برابر ۷۷ نمونه و برای k=40، k=60 و k=80 نیز به ترتیب ۹۸، ۱۲۳ و ۱۳۸ نمونه است. با افزایش تعداد نمونه‌های نزدیکترین همسایگی، تعداد داده‌های خارج از ردیف افزایش می‌یابد که دلیل این نکته در الگوریتم این روش نهفته است.

برای تشخیص داده‌های خارج از ردیف به کمک الگوریتم ضریب خارج از ردیف بودن محلی، مقدار LOF هر نمونه با استفاده از رابطه ۱۳ و برای ۱۵ مقدار مختلف k (یعنی از ۲ درصد تعداد داده‌ها؛ k=20 تا ۱۰ درصد داده‌ها؛ k=90 به فواصل ۵ تایی) محاسبه شده است. انتخاب مقدار LOF برای هر داده بستگی به شدت در حساسیت انتخاب داده‌های خارج از ردیف دارد. در مجموعه داده با حساسیت بالا (حالتی که داده‌های خارج از ردیف تاثیر زیادی بر روی پردازش‌های بعدی داده‌ها داشته باشد)، متوسط و پایین (حالتی که پردازش‌های بعدی کمتر متاثر از داده‌های خارج از ردیف باشد) به ترتیب بیشترین، میانگین و کمترین مقدار LOF برای هر داده در نظر گرفته می‌شود. در شکل ۷ مقادیر LOF بر هر نمونه با حساسیت‌های مختلف نشان داده شده است. در جدول ۲ نیز تعداد نمونه‌های خارج از ردیف شناسایی شده برای حد آستانه‌های متفاوت LOF و حساسیت‌های گوناگون آمده است. نتایج نشان می‌دهد که بالا رفتن حساسیت باعث افزایش تعداد داده‌های خارج از ردیف می‌شود. در حالی که با افزایش حد آستانه، تعداد داده‌های خارج از ردیف کاهش می‌یابد. نتایج این نکته منطقی در شکل ۷ و جدول ۲ قابل مشاهده است.



شکل ۶- مقدار برآورد چگالی کرنل برای هر نمونه به ازای چهار مقدار متفاوت k (اعداد روی محور yها برحسب $a \times 10^{-11}$ است).



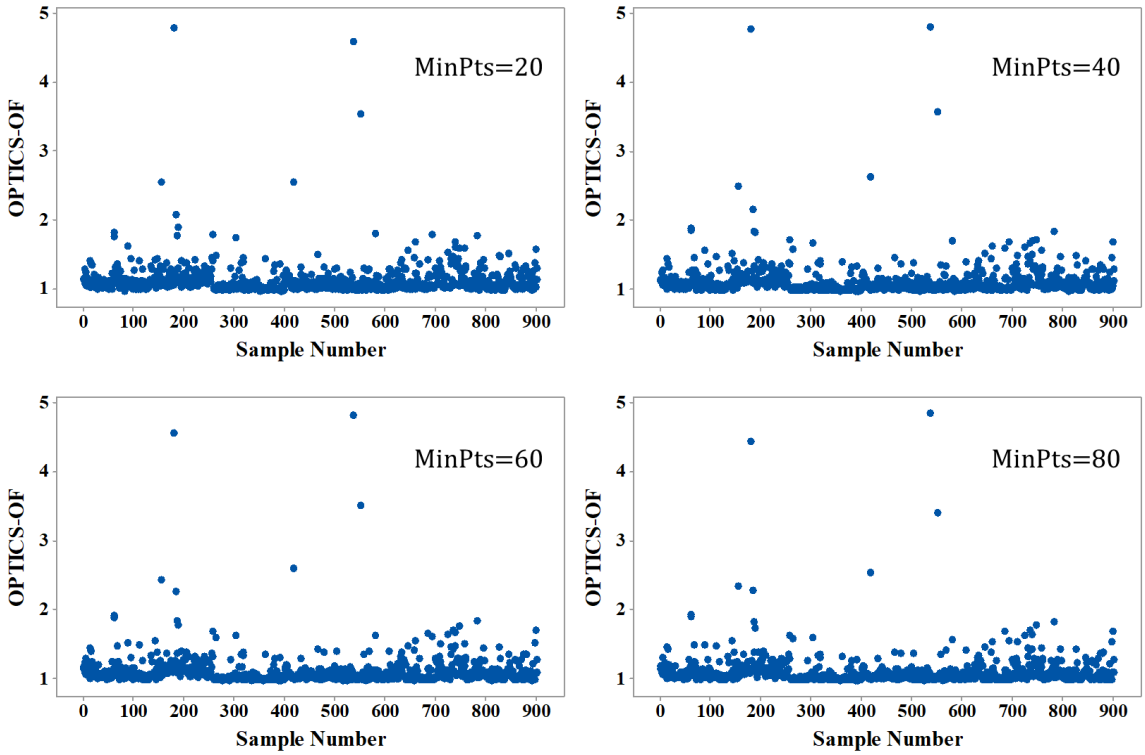


شکل ۷- نمودار مقادیر LOF برای نمونه‌ها برحسب حداکثر مقدار (A)، میانگین (B) و حداقل (C)

جدول ۲- تعداد داده‌های خارج از ردیف برای آستانه‌ها و حساسیت‌های متفاوت به روش LOF.

ضریب خارج از ردیف بودن محلی	تعداد نمونه‌های خارج از ردیف با حساسیت		
	بالا	متوسط	کم
LOF>1.1	۳۴۳	۲۷۳	۲۰۶
LOF>1.2	۱۷۰	۱۳۹	۱۱۷
LOF>1.3	۱۰۱	۷۳	۵۹
LOF>1.4	۵۷	۴۰	۳۳
LOF>1.5	۳۴	۲۶	۲۱

در مرحله سوم از الگوریتم OPTICS-OF برای تعیین تعداد نمونه‌های خارج از ردیف استفاده شده است. تعداد نمونه‌های نزدیکترین همسایگی (حد آستانه تراکم) بین ۲۰ تا ۹۰ عدد و به فواصل ۱۰تایی انتخاب شده است. مقدار ضریب خارج از ردیف بودن هر نمونه (OF) با رابطه ۱۳ و به کمک فاصله قابل دسترس هر نمونه به وسیله نمونه‌های نزدیکترین همسایگی برآورد شده است. در شکل ۸ مقادیر این ضریب برای چهار حد آستانه تراکم مشاهده می‌شود. دامنه تغییرات OF در شکل ۸ برابر دامنه تغییرات LOF در شکل ۷ است. در جدول ۳ نیز تعداد داده‌های خارج از ردیف به ازای مقادیر متفاوت انتخابی برای حداکثر مقدار ضریب خارج از ردیف بودن آمده است. نتایج نشان می‌دهد که با افزایش حد آستانه تراکم تعداد داده‌های خارج از ردیف کاهش می‌یابد. البته آهنگ کاهش با افزایش حداکثر مقدار ضریب خارج از ردیف بودن انتخابی تقلیل می‌یابد تا اینکه به یک حد ثابت برسد. تعداد داده‌های خارج از ردیف برآورده شده با الگوریتم OPTICS-OF برای مجموعه داده‌های منطقه مطالعاتی بین ۲۵ تا حداکثر ۳۰۷ متغیر است.



شکل ۸- نمودار مقادیر OF برای نمونه‌ها برحسب حد آستانه‌های مختلف.

جدول ۳- تعداد داده‌های خارج از ردیف برای حد آستانه و ضریب خارج از ردیف بودن متفاوت به روش OPTICS-OF.

ضریب خارج از ردیف بودن	حد آستانه تراکم (*MinPts*)							
	۲۰	۳۰	۴۰	۵۰	۶۰	۷۰	۸۰	۹۰
OF>1.1	۳۰۷	۲۸۷	۲۸۳	۲۷۷	۲۶۹	۲۵۳	۲۴۳	۲۳۲
OF>1.2	۱۵۷	۱۵۱	۱۴۶	۱۳۹	۱۳۷	۱۳۲	۱۲۶	۱۲۴
OF>1.3	۸۲	۸۷	۸۰	۷۶	۷۱	۶۸	۶۵	۶۳
OF>1.4	۴۶	۴۶	۴۴	۴۰	۴۰	۴۰	۴۰	۳۹
OF>1.5	۲۵	۲۷	۲۷	۲۷	۲۷	۲۷	۲۶	۲۶

جدول ۴- تعداد داده‌های خارج از ردیف برای مقادیر متفاوت σ و C به روش SVDD.

		Sigma (σ)									
		۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
C	۰/۱	۴۶	۱۲۲	۱۳۹	۱۳۹	۱۴۴	۱۴۵	۱۴۷	۱۴۹	۱۴۷	۱۵۳
	۰/۲	۸۰	۱۳۹	۱۴۶	۱۴۸	۱۵۳	۱۴۲	۱۵۲	۱۵۰	۱۵۱	۱۵۸
	۰/۳	۱۰۳	۱۵۲	۱۵۶	۱۵۴	۱۵۵	۱۵۴	۱۵۳	۱۵۳	۱۵۳	۱۵۳
	۰/۴	۱۱۹	۱۶۴	۱۶۱	۱۶۱	۱۵۸	۱۵۸	۱۶۰	۱۵۸	۱۵۸	۱۵۸
	۰/۵	۱۳۰	۱۶۸	۱۶۶	۱۶۳	۱۶۴	۱۶۵	۱۶۵	۱۶۲	۱۶۲	۱۶۲
	۰/۶	۱۴۴	۱۷۱	۱۶۸	۱۶۷	۱۶۶	۱۶۵	۱۶۵	۱۶۴	۱۶۴	۱۶۴
	۰/۷	۱۵۵	۱۷۲	۱۷۲	۱۷۲	۱۶۶	۱۶۶	۱۶۹	۱۶۵	۱۶۹	۱۶۵
	۰/۸	۱۵۸	۱۷۶	۱۷۲	۱۷۱	۱۶۷	۱۷۲	۱۷۰	۱۷۲	۱۷۲	۱۷۵
	۰/۹	۱۶۸	۱۸۱	۱۷۷	۱۷۳	۱۷۳	۱۷۳	۱۷۵	۱۷۵	۱۷۵	۱۷۵
	۱	۱۷۳	۱۸۳	۱۷۹	۱۷۸	۱۷۷	۱۷۷	۱۷۶	۱۷۵	۱۷۴	۱۷۴

در مرحله آخر از الگوریتم SVDD با تابع کرنل RBF برای تشخیص داده‌های خارج از ردیف استفاده شده است. برای این منظور مقادیر متفاوت σ در بازه ۱ تا ۱۰ و مقادیر متفاوت C در بازه ۰/۱ تا ۱ به کار رفته است. جدول ۴ تعداد نمونه‌های خارج از ردیف به دست آمده با این الگوریتم را نشان می‌دهد. داده‌های این جدول نشان می‌دهند که با افزایش مقدار C تعداد نمونه‌های خارج از ردیف افزایش می‌یابد، به نحوی که در مقادیر کوچکتر σ آهنگ افزایش بیشتر و در مقادیر بزرگتر σ آهنگ افزایش کمتری در شناسایی داده‌ی خارج از ردیف دارد. همچنین در مقادیر ثابت C نیز افزایش مقدار σ باعث افزایش تعداد نمونه‌های خارج از ردیف می‌شود. البته با افزایش مقادیر C این تاثیر کمتر می‌شود. تعداد داده‌های خارج از ردیف در این الگوریتم بین ۴۶ تا ۱۸۳ عدد برآورد شده است. از نکات جالب توجه این الگوریتم مشترک بودن شماره نمونه‌های خارج از ردیف شناسایی شده برای مقادیر متفاوت σ و C است.

۴-۳- مقایسه نتایج الگوریتم‌ها

نتایج به دست آمده نشان می‌دهد که تعداد داده‌های خارج از ردیف شناسایی شده در الگوریتم‌های مختلف بستگی به نظر کارشناس و پارامترهای انتخابی دارد. بنابراین برای مقایسه نتایج می‌تواند از دو دیدگاه مختلف زیر استفاده کرد:

الف- تشخیص نمونه‌های دارای خطا: در یک مجموعه داده حاصل از آنالیز نمونه‌های ژئوشیمیایی که دقت لازم (به عنوان مثال در سطح اعتماد ۹۹ درصد) در نمونه‌برداری، آماده‌سازی و آنالیز انجام گرفته باشد، انتظار می‌رود که کمتر از ۱ درصد داده‌های دارای خطا باشد. بنابراین در مجموعه داده‌های منطقه مطالعاتی ۱۰ نمونه‌ای که دارای بیشترین میزان احتمال خارج از ردیف بودن‌اند، برای مقایسه انتخاب می‌شود. در جدول ۵ شماره این نمونه‌ها برای الگوریتم‌های مختلف آمده است. داده‌های این جدول، نشان‌دهنده نتایج کاملا یکسان برای هر چهار الگوریتم داده‌کاوی انتخابی است (جزو یک استثنا برای نمونه شماره ۲۵۸ با الگوریتم KDE). در حالی که در روش کلاسیک MD تنها ۵ نمونه با نمونه‌های سایر الگوریتم‌ها همخوانی دارد. این نکته نشان‌دهنده برتری روش‌های داده‌کاوی بر روش‌های کلاسیک در تشخیص داده‌های خارج از ردیف است. از آنجا که در پروژه‌های ژئوشیمیایی، برداشت نمونه‌های تکراری مرسوم است، بهترین راه حل برای شناسایی خارج از ردیف بودن نمونه‌های جدول ۵، برداشت و آنالیز مجدد این نمونه‌ها و مقایسه نتایج با یکدیگر است. در صورت داشتن نتایج آنالیز متفاوت، نمونه خارج از ردیف بوده و از مجموعه داده‌ها کنار گذاشته می‌شود ولی در صورت به دست آمده نتایج مشابه، نمونه متعلق به جامعه آماری غیرنرمال است و در مجموعه داده‌ها حفظ خواهد شد.

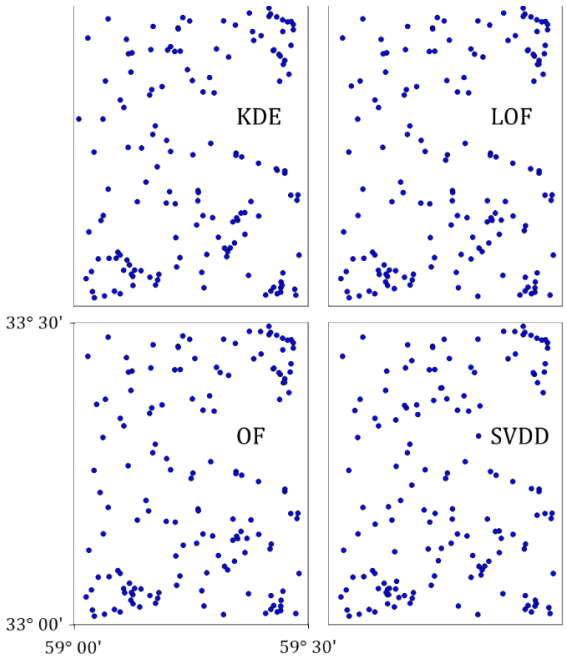
جدول ۵- شماره ۱۰ نمونه دارای بالاترین احتمال خارج از ردیف بودن با الگوریتم‌های مختلف.

Sample Number				
MD	KDE	LOF	OF	SVDD
۹۵	۶۰	۶۰	۶۰	۶۰
۱۸۰	۶۲	۶۲	۶۲	۶۲
۱۸۵	۱۵۵	۱۵۵	۱۵۵	۱۵۵
۱۸۸	۱۸۰	۱۸۰	۱۸۰	۱۸۰
۲۲۳	۱۸۵	۱۸۵	۱۸۵	۱۸۵
۳۰۴	۲۵۸	۱۸۷	۱۸۷	۱۸۷
۴۱۹	۴۱۹	۴۱۹	۴۱۹	۴۱۹
۵۳۷	۵۳۷	۵۳۷	۵۳۷	۵۳۷
۵۵۲	۵۵۲	۵۵۲	۵۵۲	۵۵۲
۶۸۵	۷۸۲	۷۸۲	۷۸۲	۷۸۲

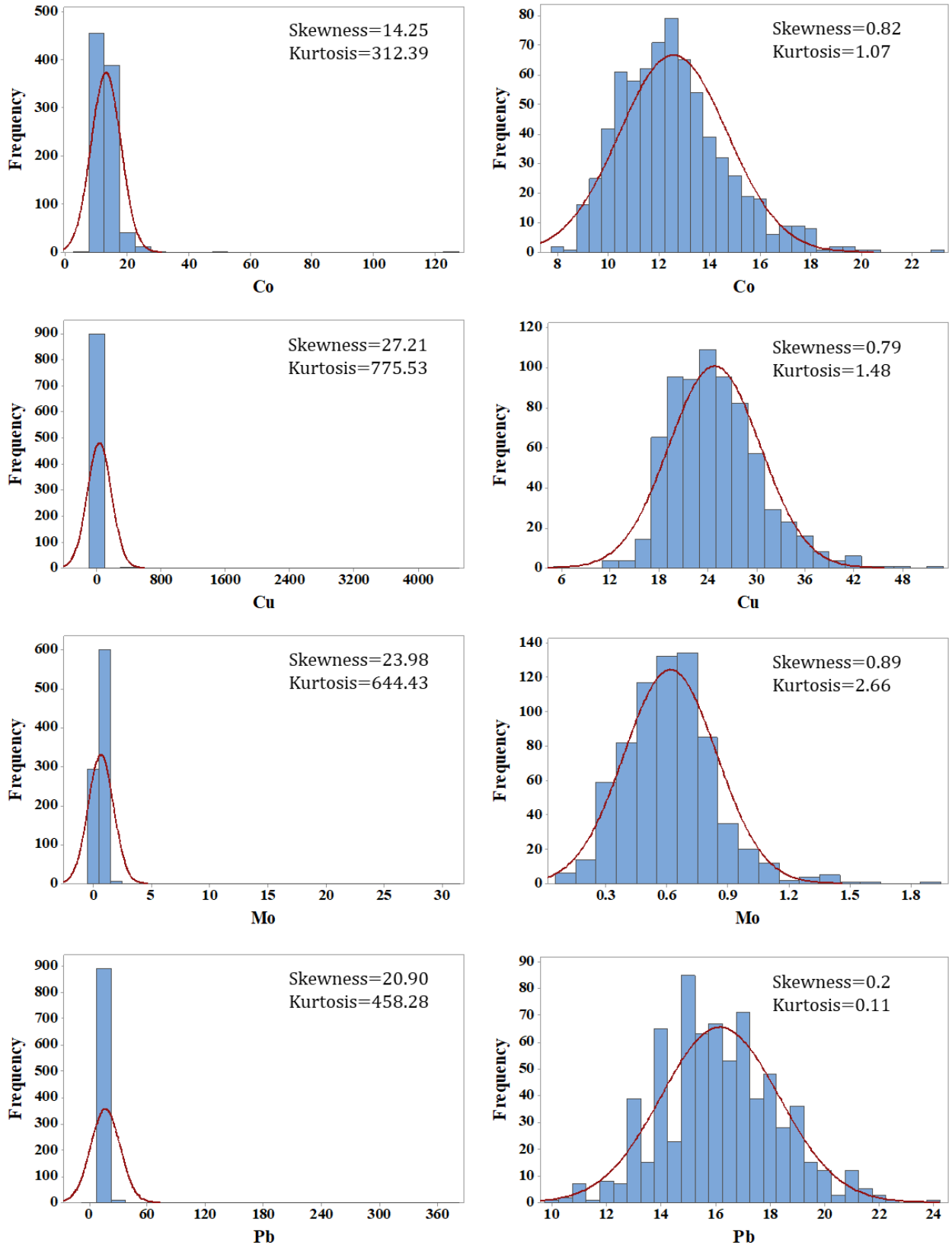
ب- تشخیص نمونه‌های متعلق به جامعه غیرنرمال: در صورتی که فرض شود در یک مجموعه داده ژئوشیمیایی ۱۰ تا ۲۰ درصد نمونه‌ها متعلق به جامعه غیرنرمال (آنومالی) باشد (این درصدها از متوسط‌گیری چند مجموعه داده ژئوشیمیایی در ورقه‌های ۱:۱۰۰،۰۰۰ اطراف ورقه روم به دست آمده است. اگر چه تعداد نمونه‌های آنومالی در ورقه‌های مختلف و بسته به نوع روش‌های تعیین آستانه، متفاوت است ولی این درصدها صرفا برای مقایسه نتایج الگوریتم‌های تشخیص داده‌های خارج از ردیف انتخاب شده است)، می‌توان فرض کرد که به طور متوسط حدود ۱۵۰ نمونه در مجموعه داده ورقه روم متعلق به جامعه غیرنرمال‌اند. بنابراین ۱۵۰ نمونه دارای بالاترین احتمال خارج از ردیف بودن در الگوریتم‌های مختلف انتخاب شده است. شکل ۹ موقعیت این نمونه‌ها را نشان می‌دهد که حاکی از مطابقت بالای الگوریتم‌ها با یکدیگر است. از این تعداد نمونه، ۷۴٫۵ درصد نمونه‌ها در همه الگوریتم‌ها به عنوان داده خارج از ردیف شناسایی شده است در حالی که ۱۶٫۱ و ۹٫۴ درصد از نمونه‌ها به ترتیب فقط با یک و دو الگوریتم به عنوان داده خارج از ردیف معرفی شده‌اند. میزان همپوشانی روش کلاسیک فاصله ماهالانوبیتس با الگوریتم‌های انتخابی بین ۵۰ تا ۵۵ درصد است.

برای ارزیابی صحت شناسایی داده‌های خارج از ردیف، می‌توان از پارامترهای آماری و نوع توزیع داده‌ها پس از کنار گذاشتن داده‌های خارج از ردیف استفاده کرد. در شکل ۱۰ هیستوگرام فراوانی چهار عنصر فلزی به عنوان نمونه آمده است. هیستوگرام فراوانی داده‌های باقیمانده نشان‌دهنده نزدیکی قابل قبول توزیع داده‌ها به توزیع نرمال است. همچنین پارامتر آماری چولگی همه عناصر (به استثنا Li) کاهش نشان می‌دهد (مقدار چولگی ۳۴ عنصر به مقدار صفر نزدیک شده و ۹ عنصر کاهش یافته است) و پارامتر کشیدگی ۲۴ عنصر نیز کاهش قابل ملاحظه‌ای یافته و به عدد ۳ نزدیک شده است.

برای بررسی برتری آنالیز بر روی مجموعه داده‌های مقاوم شده (فاقد داده خارج از ردیف) نسبت به داده‌های اولیه از روش تحلیل فاکتوری نیز استفاده شده است. تحلیل فاکتوری مقاوم مطابق الگوریتم پیشنهادی فیلزموز و همکارانش (۲۰۰۹) انجام شده است. برای انجام این تحلیل داده‌ها به روش clr از سیستم بسته به باز تبدیل شده است. همچنین برای محاسبه ضرایب از روش مولفه‌های اصلی و چرخش داده‌ها به روش Varimax بهره برده شده است. جدول‌های ۶ و ۷ به ترتیب بارهای فاکتوری عناصر برای تحلیل بر روی مجموعه داده‌های مقاوم شده و داده‌های اولیه باز شده را نشان می‌دهد (بارهای فاکتوری مثبت و منفی معنی‌دار هر عنصر به ترتیب به صورت پررنگ و پررنگ ایتالیک نمایش داده شده است) در این جدول ۱۱ فاکتور اول نشان داده شده‌اند که دارای مقادیر ویژه بزرگتر از یک‌اند.



شکل ۹- موقعیت ۱۵۰ نمونه دارای بالاترین احتمال خارج از ردیف بودن با الگوریتم‌های مختلف.

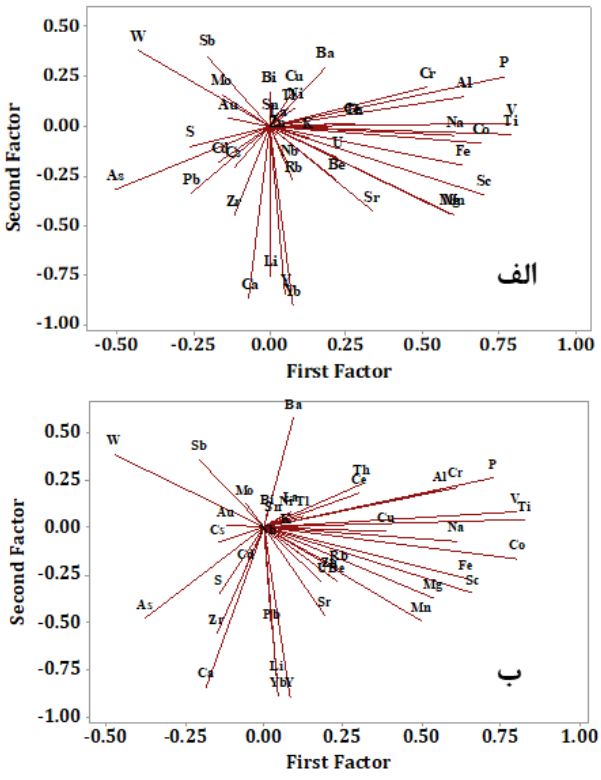


شکل ۱۰- هسیتوگرام فراوانی چهار عنصر فلزی قبل از حذف داده‌های خارج از ردیف (نمودارهای سمت راست) و پس از حذف داده‌های خارج از ردیف (نمودارهای سمت چپ) به همراه مقادیر چولگی و کشیدگی آن‌ها.

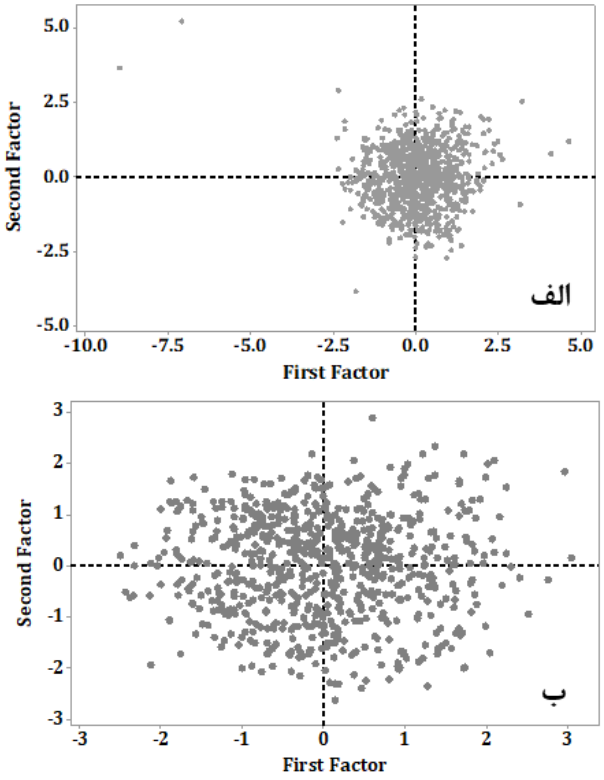
مقایسه مقادیر توجیه‌پذیری (واریانس) فاکتورها نشان می‌دهد که اگر چه تفاوت معنی‌داری بین داده‌های دو جدول نیست ولی مقادیر توجیه‌پذیری فاکتور ۲ تا ۵ در مجموعه داده‌های مقاوم (جدول ۷) کمی از داده‌های متناظرشان بیشتر است و این نکته به دلیل تمرکز تحلیل فاکتوری مقاوم بر روی فرآیندهایی است که متاثر از اکثریت داده‌ها همراه با کاهش اثرات داده‌های خارج از ردیف است. همچنین تعداد عناصر تاثیرگذار در این فاکتورها به ویژه فاکتور ۲ و ۳ که مهمترین فاکتورها محسوب می‌شوند، در مجموعه داده‌های مقاوم از داده‌های اولیه بیشتر است. وجود عناصر As، Pb و Zr در فاکتور ۲ و عناصر Cu و Zn در فاکتور ۳ در جدول ۷ نشان‌دهنده این موضوع است.

شکل ۱۱ بارهای فاکتوری عناصر را برای دو فاکتور اول نشان می‌دهد. افزایش همبستگی بین Ba با عناصر Sn و Bi، Pb با عنصر Zn، Cu با عناصر Fe، Pb و Zn و Ni با عنصر Cr و همچنین کاهش همبستگی بین عنصر Au با Mo و S و عنصر Cu با Bi از مهمترین مشخصه نمودار داده‌های مقاوم شده نسبت به نمودار داده‌های اولیه است.

در شکل ۱۲ امتیاز فاکتوری نمونه‌ها در دو فاکتور اول به صورت نمودار پراکندگی ارایه شده است. کاهش دامنه امتیاز فاکتوری نمونه‌ها در هر دو فاکتور و یک دست شدن ابر پراکندگی داده‌ها از مزیت‌های نمودار ۱۲- ب نسبت به نمودار ۱۲- الف است. ساختار تقریبا کروی شکل نمودار امتیاز فاکتوری نمونه‌ها برای داده‌های مقاوم شده (شکل ۱۲- ب) نشان‌دهنده یک جامعه‌ای شدن تقریبی داده‌های مقاوم شده دارد. بنابراین هر گونه تحلیل آماری چند متغیره یا تحلیل داده‌کاوی بر روی این داده‌ها نتیجه بهتر به همراه تفسیر راحتر نتایج را به همراه خواهد داشت.



شکل ۱۱- نمودار بارهای فاکتوری عناصر در دو فاکتور اول برای الف) داده‌های اولیه و ب) داده‌های مقاوم شده.



شکل ۱۲- نمودار امتیاز فاکتوری نمونه‌ها در دو فاکتور اول برای الف) داده‌های اولیه و ب) داده‌های مقاوم شده.

جدول ۶- بارهای فاکتوری عناصر برای تحلیل فاکتوری بر رروی داده‌های اولیه.

عناصر	F1	F2	F3	F4	F5	F6	F7	F8	F9	F0	F11
Al	۰/۶۳۵	۰/۱۴۸	۰/۳۵۶	۰/۱۸۴	-۰/۲۰۵	-۰/۳۱۷	۰/۰۷۷	-۰/۰۷۰	-۰/۰۸۲	-۰/۱۲۹	-۰/۱۴۷
As	-۰/۵۰۳	-۰/۳۱۹	-۰/۰۶۵	۰/۰۸۴	۰/۰۹۴	۰/۱۴۰	۰/۳۹۳	۰/۲۴۱	۰/۰۷۷	۰/۲۲۴	۰/۱۱۵
Au	-۰/۱۳۸	۰/۰۴۱	-۰/۰۴۰	-۰/۰۱۶	-۰/۰۲۷	۰/۰۴۲	-۰/۱۱۹	۰/۰۱۰	-۰/۲۷۳	-۰/۰۰۴	-۰/۷۲۷
Ba	۰/۱۸۳	۰/۲۹۸	-۰/۰۵۷	۰/۷۱۷	-۰/۲۰۰	-۰/۰۰۹	۰/۰۴۲	-۰/۲۴۹	-۰/۰۸۹	-۰/۱۷۸	۰/۰۰۶
Be	۰/۲۱۷	-۰/۲۶۵	۰/۸۰۹	-۰/۰۶۱	۰/۱۰۵	-۰/۲۲۸	-۰/۰۱۵	-۰/۱۲۸	-۰/۰۶۰	-۰/۱۰۱	۰/۰۰۲
Bi	-۰/۰۰۱	۰/۱۷۴	-۰/۰۲۵	-۰/۰۸۶	۰/۰۴۷	-۰/۱۴۹	۰/۷۹۷	۰/۱۷۷	۰/۰۲۲	-۰/۱۱۹	-۰/۰۲۳
Ca	-۰/۰۶۷	-۰/۸۶۲	-۰/۱۷۱	-۰/۰۶۹	-۰/۰۰۷	-۰/۲۲۸	-۰/۰۲۴	۰/۰۱۵	-۰/۱۲۳	۰/۰۳۰	۰/۰۰۴
Cd	-۰/۱۶۴	-۰/۱۷۹	۰/۰۶۰	-۰/۰۲۹	۰/۱۳۴	۰/۱۸۸	۰/۷۱۶	-۰/۱۶۳	۰/۰۵۸	۰/۱۵۴	۰/۰۶۹
Ce	۰/۲۶۸	۰/۰۱۷	-۰/۱۰۹	۰/۲۵۲	-۰/۲۴۰	۰/۲۸۲	۰/۰۳۸	-۰/۱۶۸	-۰/۵۱۲	-۰/۲۶۳	۰/۱۸۷
Co	۰/۶۹۰	-۰/۰۸۶	-۰/۱۷۶	-۰/۱۴۲	۰/۴۷۰	۰/۱۰۹	۰/۱۲۴	-۰/۰۷۵	۰/۱۵۹	-۰/۰۸۰	۰/۰۵۰
Cr	۰/۵۱۶	۰/۱۹۶	-۰/۰۶۵	۰/۱۲۶	۰/۴۳۷	۰/۱۳۲	-۰/۱۱۳	۰/۰۴۱	۰/۲۷۹	۰/۰۹۱	-۰/۱۵۷
Cs	-۰/۱۱۵	-۰/۲۰۵	۰/۰۵۶	۰/۳۷۰	۰/۰۵۹	-۰/۱۷۱	-۰/۱۱۱	۰/۱۵۲	-۰/۳۱۷	-۰/۴۸۷	-۰/۰۳۹
Cu	۰/۰۸۰	۰/۱۸۴	۰/۲۰۵	-۰/۴۰۰	۰/۵۲۱	۰/۱۱۶	۰/۲۷۹	-۰/۰۵۶	-۰/۰۴۲	-۰/۰۴۸	۰/۲۷۸
Fe	۰/۶۳۱	-۰/۱۹۸	۰/۳۰۱	-۰/۱۳۷	۰/۳۹۳	-۰/۱۸۶	۰/۰۶۴	-۰/۰۵۵	-۰/۰۶۶	-۰/۱۸۲	-۰/۰۱۶
K	۰/۱۲۵	-۰/۰۵۷	۰/۷۱۲	۰/۰۷۶	-۰/۰۲۲	-۰/۳۵۰	-۰/۰۵۳	-۰/۰۳۰	-۰/۲۵۹	-۰/۱۹۹	-۰/۰۵۴
La	۰/۰۳۰	۰/۰۰۵	۰/۲۳۳	۰/۸۳۶	۰/۰۱۷	-۰/۰۲۶	-۰/۰۳۲	۰/۱۰۵	۰/۰۳۴	۰/۰۸۶	-۰/۱۵۹
Li	۰/۰۰۳	-۰/۷۵۴	۰/۳۰۲	-۰/۱۲۵	-۰/۱۹۰	۰/۱۲۳	۰/۰۳۱	۰/۰۲۶	۰/۰۷۶	۰/۰۴۲	-۰/۱۱۹
Mg	۰/۵۸۹	-۰/۴۳۹	۰/۰۱۱	۰/۲۵۰	۰/۲۰۲	-۰/۱۳۴	-۰/۲۳۳	-۰/۱۶۰	۰/۰۱۶	۰/۱۴۸	-۰/۰۰۶
Mn	۰/۶۰۰	-۰/۴۴۳	۰/۱۰۱	-۰/۰۵۷	۰/۰۳۰	-۰/۲۵۳	۰/۰۹۷	-۰/۲۳۱	-۰/۰۱۹	-۰/۲۰۶	۰/۰۸۳
Mo	-۰/۱۵۶	۰/۱۶۱	-۰/۰۶۵	-۰/۱۸۶	۰/۰۵۷	-۰/۰۶۵	-۰/۰۸۲	-۰/۰۳۷	-۰/۲۴۵	۰/۰۲۶	۰/۷۱۲
Na	۰/۶۰۶	-۰/۰۴۵	۰/۰۳۳	۰/۱۷۵	-۰/۲۳۶	-۰/۴۲۵	-۰/۱۸۳	-۰/۱۶۴	۰/۰۲۵	۰/۱۷۲	۰/۱۴۸
Nb	۰/۰۶۲	-۰/۱۹۰	-۰/۰۸۴	۰/۴۴۰	۰/۰۱۰	-۰/۱۵۶	-۰/۳۲۱	۰/۰۲۲	۰/۰۵۶	-۰/۰۷۱	-۰/۰۳۸
Ni	۰/۰۸۴	۰/۰۹۲	-۰/۱۸۵	-۰/۰۴۴	۰/۸۱۶	-۰/۱۰۶	-۰/۰۴۴	۰/۱۴۸	۰/۰۰۳	۰/۰۳۱	۰/۰۱۷
P	۰/۷۶۸	۰/۲۵۱	۰/۱۰۶	-۰/۰۳۹	۰/۰۳۷	-۰/۱۳۱	-۰/۰۳۵	-۰/۱۳۱	-۰/۱۵۹	۰/۰۱۶	۰/۰۲۶
Pb	-۰/۲۵۴	-۰/۳۳۶	۰/۴۷۹	۰/۱۴۹	۰/۳۹۱	۰/۰۸۹	۰/۳۱۹	۰/۱۰۲	۰/۱۰۵	-۰/۰۶۲	۰/۰۰۶
Rb	۰/۰۷۷	-۰/۲۷۴	۰/۷۹۷	۰/۲۳۰	-۰/۰۶۹	۰/۱۹۰	-۰/۰۲۳	-۰/۰۱۵	-۰/۰۵۶	۰/۰۳۳	۰/۰۲۶
S	-۰/۲۵۸	-۰/۱۰۰	-۰/۲۵۸	-۰/۰۶۴	-۰/۰۰۵	-۰/۰۳۷	۰/۰۰۳	-۰/۲۲۲	-۰/۱۶۰	۰/۶۵۵	۰/۰۱۶
Sb	-۰/۲۰۶	۰/۳۵۵	-۰/۱۵۷	-۰/۱۲۶	-۰/۳۱۰	۰/۳۵۷	-۰/۱۰۹	۰/۲۱۴	۰/۳۵۴	۰/۲۵۷	۰/۰۴۰
Sc	۰/۷۰۳	-۰/۳۴۷	۰/۱۳۲	۰/۱۶۵	۰/۲۸۶	۰/۰۳۹	-۰/۱۳۱	۰/۰۴۸	۰/۱۱۶	۰/۰۳۶	-۰/۲۳۹
Sn	۰/۰۰۳	۰/۰۳۹	۰/۱۹۷	-۰/۰۳۵	۰/۰۱۶	-۰/۰۳۲	-۰/۰۶۹	۰/۱۰۰	-۰/۶۸۹	۰/۱۱۳	-۰/۰۸۱
Sr	۰/۳۳۶	-۰/۴۲۵	-۰/۲۹۲	۰/۲۲۹	-۰/۰۴۵	-۰/۴۰۲	-۰/۲۵۴	-۰/۰۹۰	-۰/۱۳۰	۰/۲۶۰	-۰/۰۴۶
Th	۰/۲۷۶	۰/۰۱۳	۰/۳۶۴	۰/۶۲۲	-۰/۱۷۵	۰/۱۳۷	۰/۰۴۲	-۰/۴۲۰	-۰/۱۱۴	-۰/۱۵۳	۰/۱۲۵
Ti	۰/۷۸۸	-۰/۰۴۳	۰/۰۰۲	۰/۲۴۶	-۰/۲۶۱	-۰/۰۸۶	-۰/۰۳۸	۰/۲۵۵	-۰/۱۱۸	-۰/۱۴۲	-۰/۰۰۴
Tl	۰/۰۶۲	۰/۰۸۵	-۰/۰۹۱	-۰/۱۰۰	۰/۱۳۹	-۰/۱۲۱	۰/۰۸۱	۰/۸۲۷	-۰/۰۶۹	-۰/۲۱۴	-۰/۰۴۹
U	۰/۲۱۹	-۰/۱۵۵	۰/۱۳۷	۰/۰۱۱	۰/۱۰۰	-۰/۶۹۰	-۰/۰۲۱	۰/۱۰۷	۰/۰۴۲	۰/۰۰۱	۰/۰۵۳
V	۰/۷۹۰	۰/۰۱۵	۰/۰۳۳	۰/۰۶۳	۰/۰۴۶	۰/۱۴۵	-۰/۰۹۶	۰/۲۹۶	۰/۰۱۵	-۰/۰۳۴	۰/۱۳۶
W	-۰/۴۲۸	۰/۳۷۸	-۰/۲۲۱	-۰/۲۷۸	-۰/۲۰۶	۰/۰۴۳	-۰/۱۵۷	-۰/۳۲۷	۰/۱۹۱	-۰/۳۴۲	۰/۰۴۵
Y	۰/۰۵۲	-۰/۸۴۴	۰/۲۶۸	۰/۱۶۹	۰/۰۱۴	-۰/۰۵۴	-۰/۰۳۱	-۰/۰۷۳	۰/۱۱۷	۰/۰۶۷	-۰/۰۹۸
Yb	۰/۰۷۴	-۰/۸۹۹	۰/۲۷۲	-۰/۰۷۸	۰/۰۱۶	-۰/۱۱۸	۰/۰۱۵	-۰/۰۵۱	-۰/۰۱۳	-۰/۰۸۵	۰/۰۲۴
Zn	۰/۰۲۳	-۰/۰۴۸	۰/۳۶۵	-۰/۰۹۱	۰/۶۲۴	-۰/۰۳۴	۰/۳۴۳	۰/۰۶۷	-۰/۰۳۴	-۰/۰۱۷	۰/۰۴۸
Zr	-۰/۱۱۵	-۰/۴۴۶	۰/۰۳۹	۰/۱۰۳	-۰/۰۶۸	-۰/۴۸۷	-۰/۰۵۴	۰/۲۸۶	-۰/۰۵۱	-۰/۱۲۱	۰/۱۶۱
Cumulative Variance	۱۴/۹	۲۶/۸	۳۴/۹	۴۱/۷	۴۸/۴	۵۳/۴	۵۸/۴	۶۲/۷	۶۶/۴	۷۰/۰	۷۳/۵

جدول ۷- بارهای فاکتوری عناصر برای تحلیل فاکتوری بر رروی داده‌های مقاوم شده.

عناصر	F1	F2	F3	F4	F5	F6	F7	F8	F9	F0	F11
Al	۰/۵۵۹	۰/۱۹۷	۰/۵۰۲	-۰/۰۴۰	-۰/۱۲۸	۰/۲۱۱	۰/۰۱۸	-۰/۱۰۳	۰/۰۷۲	-۰/۱۱۴	-۰/۲۱۸
As	-۰/۳۷۵	-۰/۴۷۵	-۰/۱۸۰	-۰/۱۶۵	-۰/۱۵۵	-۰/۰۵۲	۰/۳۲۰	۰/۲۱۵	-۰/۰۲۷	۰/۲۴۴	۰/۰۳۲
Au	-۰/۱۲۰	۰/۰۱۲	-۰/۱۴۰	-۰/۰۸۴	-۰/۱۴۴	-۰/۰۹۱	-۰/۲۱۸	-۰/۲۴۰	-۰/۴۵۲	-۰/۱۱۵	-۰/۳۵۸
Ba	۰/۰۹۴	۰/۵۷۸	۰/۰۴۵	۰/۲۶۱	-۰/۳۹۱	-۰/۰۴۴	۰/۰۸۱	-۰/۱۹۹	۰/۰۱۳	۰/۳۶۷	۰/۰۲۴
Be	۰/۲۳۴	-۰/۲۸۴	۰/۸۳۷	۰/۰۰۸	۰/۰۸۰	-۰/۱۲۳	-۰/۰۵۳	-۰/۰۰۲	۰/۰۴۴	-۰/۰۲۰	-۰/۱۱
Bi	۰/۰۱۲	۰/۰۷۸	۰/۱۳۰	-۰/۲۸۲	۰/۰۲۶	۰/۰۵۰	۰/۷۵۵	-۰/۱۲۷	۰/۰۷۴	-۰/۱۹۰	۰/۰۳۵
Ca	-۰/۱۸۱	-۰/۸۳۷	-۰/۰۵۰	-۰/۰۲۲	۰/۱۹۶	-۰/۱۸۴	۰/۱۱۱	-۰/۰۹۵	۰/۰۲۱	۰/۱۱۱	-۰/۰۸۹
Cd	-۰/۰۵۷	-۰/۲۱۰	۰/۰۰۸	۰/۲۲۲	۰/۰۳۷	-۰/۰۳۲	۰/۷۶۱	۰/۱۱۲	-۰/۰۱۰	-۰/۰۴۲	-۰/۰۴۵
Ce	۰/۳۰۲	۰/۱۸۴	-۰/۰۲۸	۰/۳۶۱	-۰/۱۲۰	۰/۳۲۶	۰/۲۰۹	-۰/۵۴۲	-۰/۰۷۳	۰/۲۰۳	۰/۰۲۱
Co	۰/۸۰۳	-۰/۱۶۲	۰/۱۱۰	۰/۰۶۳	۰/۱۹۲	-۰/۱۸۷	۰/۰۲۴	-۰/۰۷۶	۰/۰۰۱	-۰/۱۶۶	۰/۱۲۵
Cr	۰/۶۱۱	۰/۲۱۲	۰/۰۴۲	-۰/۰۵۲	-۰/۱۷۶	-۰/۰۶۶	-۰/۱۱۳	۰/۰۴۳	-۰/۳۰۷	۰/۰۷۵	۰/۰۷۰
Cs	-۰/۱۴۴	-۰/۰۸۱	۰/۱۰۴	-۰/۲۲۴	-۰/۱۴۳	-۰/۱۴۹	-۰/۱۱۴	-۰/۷۱۰	-۰/۰۰۸	۰/۰۲۸	-۰/۰۴۹
Cu	۰/۳۸۸	-۰/۰۱۹	۰/۴۶۱	۰/۰۲۶	۰/۴۱۹	-۰/۰۹۷	۰/۰۹۴	-۰/۰۰۸	۰/۱۱۳	-۰/۱۷۰	۰/۰۳۲
Fe	۰/۶۴۲	-۰/۲۶۸	۰/۴۳۹	-۰/۰۴۰	۰/۱۷۱	-۰/۲۱۴	۰/۱۲۵	-۰/۲۱۲	۰/۰۳۵	۰/۱۶۳	۰/۰۲۶
K	۰/۰۷۲	-۰/۰۲۶	۰/۸۸۱	-۰/۰۵۴	-۰/۰۱۹	-۰/۰۳۵	۰/۰۹۲	-۰/۱۶۷	۰/۰۱۴	۰/۱۵۶	-۰/۱۲۲
La	۰/۰۸۶	۰/۰۸۵	۰/۰۵۵	-۰/۰۰۸	-۰/۸۷۲	-۰/۱۸۷	-۰/۰۷۵	-۰/۱۲۱	-۰/۱۳۴	۰/۰۵۷	-۰/۰۴۴
Li	۰/۰۴۳	-۰/۸۰۲	۰/۱۲۵	۰/۰۴۳	۰/۰۳۶	۰/۲۷۶	۰/۰۷۶	۰/۰۵۱	-۰/۱۷۹	-۰/۰۴۵	۰/۰۹۱
Mg	۰/۵۴۰	-۰/۳۷۰	۰/۲۷۰	۰/۲۰۴	-۰/۰۶۸	-۰/۴۴۸	-۰/۰۶۶	۰/۰۶۲	۰/۰۱۵	۰/۲۲۰	-۰/۰۹۵
Mn	۰/۵۰۲	-۰/۴۹۳	۰/۲۲۳	۰/۱۱۷	۰/۱۶۰	۰/۰۴۲	۰/۱۰۵	-۰/۲۳۳	۰/۲۴۹	-۰/۰۴۳	۰/۰۵۸
Mo	-۰/۰۶۰	۰/۱۲۸	-۰/۰۹۳	۰/۰۰۴	۰/۱۲۶	-۰/۰۷۶	۰/۰۲۱	-۰/۰۳۸	۰/۶۸۹	-۰/۰۵۴	۰/۰۲۳
Na	۰/۶۱۰	-۰/۰۷۰	۰/۱۳۲	۰/۰۵۹	-۰/۱۲۶	۰/۱۳۷	-۰/۲۴۳	۰/۰۳۹	۰/۴۹۶	-۰/۱۱۳	-۰/۲۱۰
Nb	۰/۰۱۱	-۰/۰۸۱	۰/۰۳۰	-۰/۰۶۳	-۰/۰۶۶	-۰/۰۱۳	-۰/۱۶۶	-۰/۰۷۱	-۰/۰۴۱	۰/۸۱۶	-۰/۰۳۶
Ni	۰/۰۷۱	۰/۰۷۳	۰/۱۴۹	-۰/۲۰۲	-۰/۱۵۹	-۰/۸۵۰	۰/۰۱۷	-۰/۱۵۲	۰/۰۱۳	-۰/۰۱۶	-۰/۰۱۶
P	۰/۷۲۵	۰/۲۶۰	۰/۲۱۸	۰/۲۰۹	۰/۰۳۹	-۰/۰۶۵	۰/۰۱۵	۰/۰۱۷	۰/۱۵۲	-۰/۱۷۱	-۰/۱۲۱
Pb	۰/۰۲۶	-۰/۵۳۲	۰/۳۵۴	-۰/۱۷۱	-۰/۴۳۳	-۰/۱۳۰	۰/۰۵۴	-۰/۱۸۵	-۰/۱۲۶	-۰/۱۱۸	۰/۰۶۰
Rb	۰/۲۳۶	-۰/۲۱۷	۰/۶۷۴	۰/۱۹۳	-۰/۳۳۵	۰/۰۶۲	-۰/۰۰۵	۰/۰۵۴	-۰/۱۶۸	-۰/۰۵۷	۰/۰۰۷
S	-۰/۱۴۰	-۰/۳۴۷	-۰/۳۷۱	۰/۳۸۵	۰/۰۳۴	-۰/۰۳۱	۰/۱۸۲	۰/۰۶۳	-۰/۰۸۳	-۰/۰۷۲	-۰/۳۶۹
Sb	-۰/۲۰۸	۰/۳۵۹	-۰/۳۱۴	-۰/۰۱۵	-۰/۰۰۳	۰/۲۳۴	-۰/۰۷۹	۰/۵۲۴	-۰/۰۹۵	-۰/۰۵۲	۰/۱۷۹
Sc	۰/۶۶۱	-۰/۳۴۳	۰/۲۲۵	۰/۰۴۸	-۰/۱۵۲	-۰/۲۹۲	-۰/۱۱۱	۰/۱۰۱	-۰/۳۱۳	۰/۰۸۵	۰/۰۲۲
Sn	۰/۰۳۱	۰/۰۳۸	۰/۱۹۶	-۰/۰۷۸	-۰/۰۰۵	-۰/۰۳۷	۰/۰۱۴	-۰/۰۵۸	-۰/۰۱۶	۰/۰۴۸	-۰/۷۱۱
Sr	۰/۱۹۴	-۰/۴۶۵	-۰/۱۳۴	۰/۰۷۹	-۰/۰۴۴	-۰/۴۷۹	-۰/۰۶۷	-۰/۰۰۵	۰/۱۴۱	۰/۰۷۳	-۰/۳۳۷
Th	۰/۳۰۸	۰/۲۳۰	۰/۳۲۶	۰/۵۲۳	-۰/۴۲۶	۰/۰۸۷	۰/۱۲۲	-۰/۳۰۵	۰/۰۵۷	۰/۱۷۵	۰/۰۸۳
Ti	۰/۸۲۷	۰/۰۴۲	-۰/۰۲۹	-۰/۲۳۱	-۰/۱۸۵	۰/۱۶۱	۰/۰۶۳	-۰/۱۴۵	۰/۰۳۸	۰/۱۰۲	-۰/۰۴۶
Tl	۰/۱۲۵	۰/۰۶۳	-۰/۰۴۴	-۰/۸۳۹	-۰/۰۴۸	-۰/۱۱۴	۰/۱۰۶	-۰/۱۱۵	-۰/۰۹۸	۰/۰۴۰	-۰/۰۴۱
U	۰/۱۸۲	-۰/۲۸۳	۰/۲۲۳	-۰/۴۵۱	۰/۰۲۷	-۰/۰۳۵	-۰/۰۸۱	-۰/۰۵۵	۰/۲۷۹	۰/۲۰۱	-۰/۱۹۸
V	۰/۷۹۸	۰/۰۸۱	۰/۱۲۰	-۰/۱۶۷	-۰/۰۱۵	-۰/۰۱۵	-۰/۰۴۲	۰/۱۳۶	-۰/۰۲۷	۰/۱۲۶	۰/۰۳۴
W	-۰/۴۷۴	۰/۳۸۶	-۰/۱۹۱	۰/۱۲۹	۰/۳۴۶	۰/۱۶۲	-۰/۲۰۳	-۰/۰۳۰	۰/۰۹۱	-۰/۱۲۱	۰/۲۹۰
Y	۰/۰۸۳	-۰/۸۸۸	۰/۱۱۲	۰/۰۹۴	-۰/۱۸۵	-۰/۰۶۰	-۰/۰۵۲	-۰/۰۱۱	-۰/۰۸۲	۰/۰۵۹	-۰/۰۰۸
Yb	۰/۰۴۴	-۰/۰۸۸۵	۰/۲۷۴	۰/۰۰۵	۰/۱۵۰	-۰/۰۶۲	۰/۰۵۷	-۰/۰۸۲	۰/۰۱۸	۰/۰۵۷	۰/۰۱۷
Zn	۰/۲۱۰	-۰/۲۵۷	۰/۴۶۲	-۰/۱۱۹	-۰/۱۱۶	-۰/۳۲۳	۰/۱۲۰	-۰/۰۷۸	-۰/۰۸۱	-۰/۰۴۰	۰/۰۱۹
Zr	-۰/۱۴۹	-۰/۵۵۴	۰/۱۰۱	-۰/۳۲۷	-۰/۱۵۹	-۰/۱۱۹	-۰/۰۶۹	-۰/۱۴۳	۰/۴۰۲	۰/۰۰۹	-۰/۰۴۲
Cumulative Variance	۱۴/۸	۲۹/۲	۳۸/۷	۴۴/۱	۴۹/۴	۵۴/۳	۵۸/۵	۶۲/۶	۶۶/۶	۶۹/۹	۷۳

۵- نتیجه‌گیری

تشخیص داده‌های خارج از ردیف یکی از مراحل لازم پیش‌پردازش داده‌های اکتشافی محسوب می‌شود. شناسایی این نوع داده‌ها در مجموعه داده‌های ژئوشیمیایی که تعداد ابعاد آن‌ها زیاد است، نیاز به استفاده از الگوریتم‌های داده‌کاوی دارد. در این مقاله ضمن معرفی این الگوریتم‌ها که در چهار دسته روش‌های آماری، روش‌های مبتنی بر مجاورت، روش‌های مبتنی بر خوشه‌بندی و روش‌های مبتنی بر دسته‌بندی، تقسیم‌بندی می‌شوند، کاربرد آن‌ها را بر روی داده‌های ژئوشیمیایی ورقه روم با ابعاد ماتریس ۴۱×۹۰۲ بررسی شد. نتایج نشان داد که در رویکرد تشخیص نمونه‌های دارای خطا، می‌توان ۱۰ نمونه که در هر چهار الگوریتم دارای شماره‌های یکسان‌اند را برای بررسی بیشتر انتخاب کرد. همچنین از ۱۵۰ نمونه انتخابی در رویکرد تشخیص نمونه‌های متعلق به جامعه غیرنرمال، ۷۴٫۵ درصد نمونه‌ها با هر چهار الگوریتم انتخابی به عنوان داده‌های خارج از ردیف معرفی شده‌اند. هیستوگرام فراوانی داده‌های مقاوم شده نشان‌دهنده نزدیک شدن توزیع داده‌ها به توزیع نرمال دارد. همچنین تحلیل فاکتوری دلالت بر برتری تحلیل بر روی داده‌های مقاوم شده به دلیل نزدیک شدن توزیع داده‌ها به یک جامعه و به دست آمدن نتایج واقعی‌تر نسبت به داده‌های دارای مقادیر خارج از ردیف است. از الگوریتم‌های معرفی شده می‌توان برای مقاصد زیر استفاده کرد:

۱- معرفی یک تا دو درصد از نمونه‌ها با بالاترین احتمال خارج از ردیف بودن به عنوان نمونه‌های دارای خطا و پیشنهاد برای برداشت نمونه‌برداری تکراری از محل این نمونه‌ها در پروژه‌های اکتشافات ژئوشیمیایی

۲- محاسبه ماتریس موقعیت و پراکندگی به جای ماتریس میانگین و واریانس- کواریانس پس از حذف ۱۰ تا ۲۰ درصد از نمونه‌ها با بالاترین احتمال خارج از ردیف بودن برای محاسبات آمارهای چند متغیره (در این صورت روش‌های آماری بدون نیاز به حذف داده‌های خارج از ردیف نسبت به آن‌ها مقاوم خواهند شد. روش‌های آنالیز مولفه‌های اصلی مقاوم، آنالیز فاکتوری مقاوم، رگرسیون چند متغیره مقاوم، آنالیز تمایز مقاوم، آنالیزهای طبقه‌بندی مقاوم و آنالیزهای خوشه‌بندی مقاوم از جمله این روش‌ها است).

۳- تشخیص نمونه‌های آنومالی برای ترسیم نقشه کنتوری آنومالی‌های ژئوشیمی ممکن، احتمالی و قطعی (از این الگوریتم‌ها می‌توان به عنوان یک روش تعیین آستانه برای تفکیک جوامع آماری استفاده کرد).

تشکر و قدردانی

از سازمان زمین‌شناسی و اکتشافات معدنی کشور به دلیل استفاده از داده‌های اکتشافی منطقه مطالعاتی تشکر و قدردانی می‌شود. همچنین از همکاران محترم هیات علمی گروه مهندسی کامپیوتر دانشگاه صنعتی بیرجند، جناب آقای دکتر مصطفی سبزه‌کار و مهندس نادر ملایی به دلیل راهنمایی‌های ارزشمندشان تشکر می‌شود.

منابع

۱. روشنی‌رودسری، پریسا؛ مختاری، احمدرضا؛ طباطبائی، سید حسن؛ ۱۳۹۳؛ **"بررسی آنالیز ژئوشیمیایی عناصر در سیتم عدیی باز و بسته؛ مطالعه موردی: کانسار مس کوه‌پنج (کرمان)"**، نشریه علمی- پژوهشی روش‌های تحلیلی و عددی در مهندسی معدن، دوره دوم، شماره ۴، صفحه ۴۶ تا ۵۸.

۲. کیانپوریان، ص.؛ اسدی هارونی، ه.؛ افشاری، س.؛ فرهمندیان، م.، ۱۳۹۳؛ **"جداسازی داده‌های خارج از رده به روش تک متغیره و چند متغیره در داده‌های ژئوشیمی محدوده طلای اپی‌ترمال ساری‌گونای"**، نشریه مهندسی معدن، دوره ۹، شماره ۲۵، صفحه ۸۵ تا ۹۶.

۳. گرانیان، حمید؛ خواجه‌میری، زهرا؛ ۱۳۹۶؛ **"کاربرد برآوردگرهای مقاوم در تعیین داده‌های خارج از ردیف؛ مثال موردی: داده‌های ژئوشیمیایی منطقه شاه سلیمان علی در استان خراسان جنوبی"**، نشریه علمی- پژوهشی روش‌های تحلیلی و عددی در مهندسی معدن، شماره ۱۴، صفحه ۷۳ تا ۸۵.

4. Aggarwal, C. C., 2016. "*Outlier analysis*", Second Edition. Springer, New York, 54p.

5. Ahmed, T., 2009. "*Online Anomaly Detection using KDE*". IEEE "GLOBECOM" 2009 proceedings, p. 4244-4148.

6. Ahn, J., Lee, M.H., Lee, J.A., 2019. "*Distance-based outlier detection for high dimension, low sample size data*". Journal of Applied Statistics 46, 13-29.

7. An, W., Liang, M., Liu, H., 2014. "*An improved one-class support vector machine classifier for outlier detection*". Journal of Mechanical Engineering Science, 1-9.

8. Ankerst, M., Breunig, M.M., Kriegel, H.P.,

Sander, J., 1999. "*OPTICS: Ordering Points to Identify the Clustering Structure*", in Proceedings of the ACM SIGMOD International Conference on Management of Data, vol. 28, no. 2, pp. 49-60.

9. Maronna, R.A., Martin, R.D., Yohai, V.J., Salibian-Barrera, M., 2019. "*Robust Statistics: Theory and Methods*", John Wiley & Sons., 464 p.
10. Behera, S., Rani, R., 2016. "*Comparative analysis of density-based outlier detection techniques on breast cancer data using Hadoop and map reduce*". International Conference on Inventive Computation Technologies, India.
11. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. "*LOF: identifying density-based local outliers*". In Proc. of ACM SIGMOD International Conference on Management of Data, pages 93–104.
12. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 1999. "*OPTICS-OF: Identifying Local Outliers*". in Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery, 262-270.
13. Deng, X., Wang, L., 2018. "*Modified kernel principal component analysis using double-weighted local outlier factor and its application to nonlinear process monitoring*". ISA Transactions 72, 218-228.
14. Febriana, N.L., Sitanggang, I.S., 2017. "*Outlier Detection on Hotspot Data in Riau Province using OPTICS Algorithm*". IOP Conference Series: Earth and Environmental Science 58 (2017) 012004.
15. Filzmoser, P., Garrett, R.G., Reimann, C., 2005. "*Multivariate outlier detection in exploration geochemistry*". Computers & Geosciences 31, 579–587.
16. Filzmoser, P., Hron, K., Reimann, C., 2012. "*Interpretation of multivariate outliers for compositional data*". Computers & Geosciences 39, 77–85.
17. Filzmoser, P., Hron, K., Reimann, C., 2009. "*Principal component analysis for compositional data with outliers*". Environmetrics 20, 621–632.
18. Filzmoser, P., Hron, K., Reimann, C., Garrett, R., 2009. "*Robust factor analysis for compositional data*". Computers & Geosciences 35, 1854–1861.
19. Han, J., Kamber, M., Pei, J., 2012. "*Data Mining: Concepts and Techniques*", Morgan Kaufmann, 740 p.
20. Latecki, L.J., Lazarevic, A., Pokrajac, D., 2007. "*Outlier Detection with Kernel Density Functions*". In: Perner P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2007. Lecture Notes in Computer Science, vol 4571. Springer, Berlin, Heidelberg.
21. Lima, R., 2013. "*Outlier detection with kernel density functions in monitoring the Istat Lfs data production processes*", Electronic Journal of Applied Statistical Analysis 6(1), 118 – 129.
22. Ma, Y., Shi, H., Ma, H., Wang, M., 2013. "*Dynamic process monitoring using adaptive local outlier factor*". Chemometrics and Intelligent Laboratory Systems 127, 89–101.
23. Ranga Suri, N.N.R., Murty, N., Athithan, G., 2019. "*Outlier Detection: Techniques and Applications: A Data Mining Perspective*", Springer International Publishing, 216 p.
24. Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., 1999. "*SV estimation of a distribution's support*", in: Advances in Neural Information Processing Systems, Colorado, USA, pp.582–588.
25. Sreevani, R., Murthy, C.A., 2016. "*On bandwidth selection using minimal spanning tree for kernel density estimation*". Computational Statistics and Data Analysis 102, 67–84.
26. Sudha, P., Krithigadevi, K., 2014. "*Outlier detection using high dimensional dataset for comparison of clustering algorithms*". International Journal of Advanced Research in Computer Science & Technology 2(3), 283-288.
27. Tax, D.M.J., Duin, R.P., 1999. "*Support vector domain description*". Pattern Recognition Letters 20(11-13), 1191-1199.
28. Terrell, G.R., Scott, D.W., 1992. "*Variable kernel density estimation*". The Annals of Statistics 20(3), 1236-1265.
29. Wang, Y.F., Yu, J., Su, G.P., Qian, Y.R., 2019. "*New outlier detection method based on OPTICS*". Sustainable Cities and Society 45, 197-212.
30. Xu, Y., Xu, N., Feng, X., 2016. "*A New Outlier Detection Algorithm Based on Kernel*

Density Estimation for ITS". The IEEE International Conference on Internet of Things, Green Computing and Communications, Cyber, Physical and Social Computing and Smart Data. Chengdu, China.

31. Zheng, S., 2016. "*Smoothly approximated support vector domain description*". Pattern Recognition 49, 55–64.

32. Zhou, S., Zhou, K., Wang, J., Yang, G., Wang, S., 2017. "*Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies*", Frontiers of Earth Science 12(3), 491–505.