نشریه علمی-پژوهشی "مهندسی معدن "
Iranian Journal of Mining Engineering
(IRJME)

دوره دهم، شماره ۲۷، سال ۱۳۹۴، صفحه ۹۵ تا ۱۰۷
Vol. 10, No. 27, 2015, pp.95-107

بررسی کاربرد نگاشت‌های خودسازمانده در خوشه‌بندی داده‌های آبراهه‌ای و مقایسه آن با دندروگرام اکتشافی داده‌های ترکیبی

حمید معینی[۱]، فرهاد محمدتراب[۲*]، مجید کیخای حسین پور[۳]

۱-دانشجوی دکترای اکتشاف معدن، دانشگاه یزد، یزد، ایران،hamidmoini@stu.yazd.ac.ir

۲- استادیار گروه مهندسی اکتشاف معدن، دانشگاه یزد، یزد، ایران ، fmtorab@yazd.ac.ir

۳ -دانشجوی دکترای اکتشاف معدن، دانشگاه یزد، یزد، ایران، m.keikha@stu.yazd.ac.ir

(دریافت ۵ دی ۱۳۹۳، پذیرش ۲۶ فروردین ۱۳۹۴)

چکیده

روش نگاشت خودسازمانده (SOM[۱]) یکی از روش‌های خوشه‌بندی است که می‌تواند بدون نظارت و با کمک شبکه عصبی، فضاهای چندبعدی و پیچیده داده‌ها را به یک فضای دوبعدی تبدیل کند. به دلیل بسته بودن و خاصیت ترکیبی ذاتی داده‌های ژئوشیمیایی، قبل از هر تحلیلی بایستی با تبدیل‌های خاصی باز شوند. یکی از مهم‌ترین تبدیل‌هایی که امروزه روی این نوع داده‌ها برای بازکردن آن‌ها انجام می‌شود، خانواده تبدیل‌های نسبت لگاریتمی است که در دو دهه اخیر توسط دانشمندان علوم آمار ارائه شده است. در حال حاضر روش‌های تحلیل آماری و داده‌کاوی خاص این نوع داده‌ها مثل استانداردسازی، آمار توصیفی، کاهش بعد، خوشه‌بندی، رگرسیون، زمین‌آمار و غیره از جنبه‌های مختلف در دنیا در حال بررسی است. در این پژوهش ضمن معرفی روش SOM به‌عنوان یکی از مهم‌ترین روش‌های خوشه‌بندی مبتنی بر هوش مصنوعی، کاربرد آن در تحلیل داده های ژئوشیمیایی بررسی شده است. به‌عنوان مطالعه موردی، داده‌های ژئوشیمیایی رسوبات آبراهه‌ای برگه ۱:۱۰۰،۰۰۰ خوسف که خاصیت بسته یا ترکیبی دارند، نخست با کمک تبدیل نسبت لگاریتمی مرکزی، باز شده و دندروگرام خوشه‌بندی مربوطه ترسیم شد. در مرحله بعد داده‌ها یکبار به‌صورت بازشده و بار دیگر به‌صورت خام ولی استاندارد شده با روشهای ترکیبی، به شبکه SOM وارد شده و خروجی آن هربار به‌صورت نگاشتهای صفحات وزنی نورونها برای هر متغیر ترسیم شد. الگوی توزیع وزنی در نگاشت‌های بیان‌شده برای متغیرهایی که در یک خوشه قرار می‌گیرند، بسیار به هم شبیه است. مقایسه نتایج به کارگیری این روش با نتایج دندروگرام به‌دست آمده، نشان‌دهنده انطباق قابل قبول این دو روش درصورت استفاده از داده‌های خام استاندارد شده دارد. نقطه ضعف این روش این است که تشخیص الگوهای مشابه در خروجی به عهده ناظر است و اگر تعداد متغیرها زیاد باشد نمی‌توان تمام الگوهای مشابه را به راحتی تشخیص داد.

کلمات کلیدی

شبکه عصبی، نگاشت خودسازمانده، ژئوشیمی، داده های ترکیبی، دندروگرام، خوسف

*نویسنده مسئول مکاتبات

۱- مقدمه

یکی از اهداف اصلی مطالعات ژئوشیمیایی، بررسی نحوه پراکندگی، تمرکز و روند تغییرات ناحیه‌ای عناصر است. امروزه در تحلیل داده های ژئوشیمیایی، روش‌های مختلفی به منظور گروه‌بندی متغیرها با هدف تقلیل ابعاد ماتریس داده‌ها گسترش یافته‌اند. از جمله این روش‌ها می‌توان به روش‌های چند متغیره تحلیل فاکتوری و تحلیل مؤلفه‌های اصلی و نیز تجزیه و تحلیل خوشـه‌ای اشاره کرد. به‌طور خاص از روش‌های خوشه‌بندی می‌توان به ارتباط بین گروهی[۲]، ارتباط درون‌گروهی[۳]، نزدیک-ترین همسایگی[۴]، دورترین همسایگی[۵]، اشاره کرد که به‌طور گسترده و موفقیت‌آمیزی در علوم زمین به‌کار گرفته شده‌اند [۱،۲].

به علت پیچیدگی آماری داده‌های ژئوشیمیایی و ویژگی ترکیبی بودن داده‌ها[۶]، به‌کارگیری سیستم‌های مبتنی بر هوش مصنوعی در طبقه‌بندی و شناسایی روندهای غالب چند عنصری را می‌توان به‌عنوان یک گزینه جدید بررسی کرد. روش نگاشت خودسازمانده یا به‌طور مختصر SOM یک نوع مدل شبکه عصبی است که در پیاده‌سازی و طرح‌ریزی مشخصه‌های غیرخطی از فضای چندبعدی به فضای یک یا دو بعدی استفاده می‌شود[۳]. این روش در زمینه‌های مختلفی از قبیل آنالیز تصاویر[۴،۵]، حرکت بادها[۶،۷] و تفسیر امواج لرزه‌ای[۸،۹] به‌طور موفقیت آمیزی استفاده شده است.

در این پژوهش برای نخستین بار از الگوریتم SOM به منظور گروه‌بندی عناصر رسوبات آبراهه‌ای در برگه ۱:۱۰۰،۰۰۰ خوسف استفاده شده است.

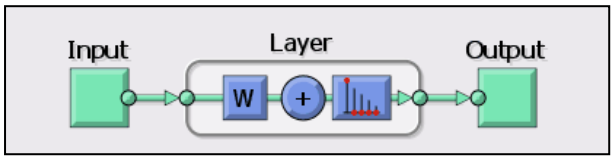
۲-روش

۲-۱- معرفی شبکه خودسازمانده

معماری‌های شبکه و تحلیل سیگنال‌های استفاده شده در مدل‌سازی سیستم‌های عصبی را می‌توان به سه گروه تقسیم کرد که هر کدام متکی بر فلسفه متفاوتی هستند[۳]. شبکه-های پیشخور[۷] مجموعه‌ای از سیگنال‌های ورودی را به مجموعه‌ای از سیگنال‌های خروجی تبدیل می‌کند. تبدیل ورودی-خروجی موردنظر اغلب با تنظیم نظارت شده و خارجی پارامترهای سیستم تعیین می‌شود. در شبکه‌های پسخور[۸]، اطلاعات ورودی، وضعیت فعالیت آغازین سیستم را تعریف می-کند، و در ادامه، وضعیت نهایی، مجانب‌وار با خروجی محاسبات تعیین می‌شود. در شبکه‌های دسته سوم، سلول‌های مجاور در یک شبکه عصبی، در فعالیت‌هایشان به وسیله تقابل‌های جانبی دوطرفه، رقابت می‌کنند و منطبق بر شناساگرهای خاص الگوهای متفاوت یک سیگنال، رشد می‌کنند. از این دسته به عنوان یادگیری رقابتی، نظارت نشده یا خودسازمانده نام برده می‌شود[۳].

نگاشت‌ها یا نقشه‌های خودسازمانده، نوعی از شبکه عصبی مصنوعی بر پایه یادگیری نظارت نشده[۹] است که نمونه‌های تحت آموزش را در یک فضای کم بعد و تفکیک شده نمایش می‌دهد. این شبکه از یک تابع همسایگی استفاده می‌کند تا خواص توپولوژی فضای ورودی[۱۰] را حفظ کند. روال قرار دادن یک بردار در یک نقشه عبارت از یافتن گره[۱۱] با نزدیکترین بردار وزن به بردار فضای داده[۱۲] است[۱۰].

شبکه SOM شامل نورون‌هایی است که در یک شبکه کم بعد[۱۳] منظم (دو یا سه بعدی) قرار گرفته‌اند. تعداد نورون‌ها ممکن است از چند ده تا چند هزار تغییر کند. به هر نورون یک بردار d بعدی با وزن m اختصاص می‌یابد که d همان بعد بردارهای ورودی است. نورون‌ها به نورون‌های مجاور خود با یک رابطه همسایگی متصل‌اند که توپولوژی یا ساختار نقشه را تحت تاثیر قرار می‌دهد. توپولوژی‌های متداول، استفاده از شبکه‌های مربعی، شش‌ضلعی، مثلثی یا بی‌قاعده است[۱۱]. شکل ۱، معماری شبکه SOM شامل ماتریس ورودی[۱۴]، نگاشت خروجی[۱۵] و لایه[۱۶] شامل نورون‌ها و بردار وزن‌های (W) نورون‌ها را نشان می‌دهد.

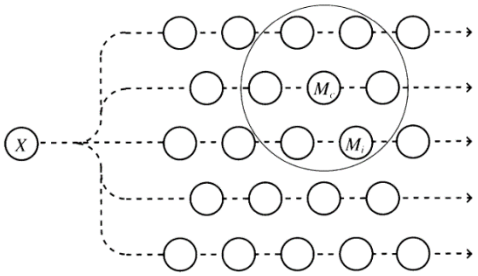


شکل۱: معماری شبکهSOM

الگوریتم آموزش SOM شبیه به الگوریتم کمی‌سازی بردار (VQ[۱۷]) یا همان k-means است، با این تفاوت مهم که علاوه بر مناسب‌ترین بردار وزنی، همسایگان توپولوژیکی آن روی نقشه نیز به روز رسانی می‌شوند. بدین معنا که ناحیه مجاور بردار مذکور در شبکه، گسترش تشابهی می‌یابد. نتیجه نهایی این است که نورون‌ها در شبکه منظم شده و نورون‌های مجاور، بردارهای وزن مشابه می‌گیرند[۱۱].

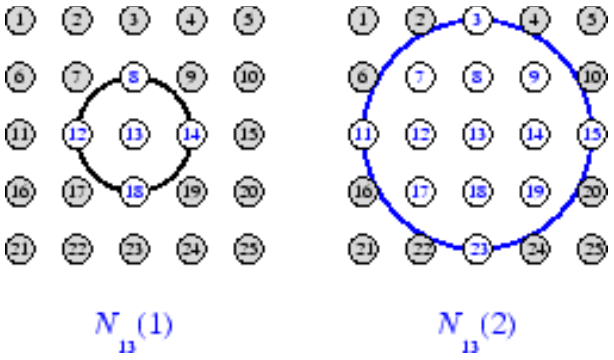
شکل ۲ مدلی از نگاشت خودسازمانده را نشان می‌دهد. برای سادگی فرض می‌شود عناصر (به‌عنوان مثال، تک نورون‌ها یا

گروه‌هایی از نورون‌های در تعامل با یکدیگر) تشکیل یک آرایه صفحه‌ای منظم بدهند و هر عنصر معرف مجموعه‌ای از مقادیر عددی M_i بنام مدل باشد. این مقادیر می‌توانند متناظر با بعضی پارامترهای سیستم عصبی باشند. فرض می‌شود که هر مدل با پیغام‌هایی که از عنصر دریافت می‌کند اصلاح می‌شود. ورودی X که یک پیغام عصبی و مجموعه‌ای از مقادیر سیگنال موازی است در مجموعه مدل‌های M_i منتشر می‌شود. در تئوری مغز انسان[۱۸] اصطلاح متداولی بنام "رقابت" بین عناصر وجود دارد بدین‌صورت که اگر عناصر توسط یک ورودی مشترک تحریک شوند، عنصری که پارامترهای آن بیشترین تطابق را با این ورودی داشته باشد از همه بیشتر تحریک می‌شود. این عنصر را "برنده"[۱۹]می‌نامند. از این پس، M_c (عنصر برنده) که بیشترین شباهت را با X دارد باعث می‌شود تمام مدل‌هایی که در مجاورت Mc (دایره بزرگتر) قرار دارند نیز تشابه خود را با X اصلاح کنند[۱۱].



شکل۲: مدلی از یک نگاشت خودسازمانده

همزمان که مدل‌های در همسایگی نورون یا عنصر برنده شروع به شبیه شدن به پیغام X می‌کنند، هرچه بیشتر نیز سعی می‌کنند به همدیگر شبیه شوند و این یعنی تمام مدل‌های در همسایگی M_c هموار[۲۰]می‌شوند. پیغام‌های مختلف در زمان‌های مختلف، قسمت‌های مختلفی از مجموعه مدل‌ها را متاثر می‌کنند و در نتیجه مدل‌های M_i پس از مراحل متعدد آموزش، مقادیری متناسب و سرشکن شده در کل آرایه گرفته‌اند درست همانند پیغام اولیه X در "فضای سیگنال". این سه مرحله یعنی انتشار ورودی، انتخاب نورون برنده و انطباق مدل‌ها در همسایگی فضایی نورون برنده، اصول اساسی یک فرایند SOM را تشکیل می‌دهند. مرحله اخیر به نوعی همان رگرسیون غیرپارامتری است[۱۲]. شکل ۳ مفهوم همسایگی با شعاع ۱ و شعاع ۲ را در یک شبکه با ۲۵ نورون نشان می‌دهد.



شکل۳: مفهوم همسایگی دریک شبکه با ۲۵ نورون

یکی از جالب‌ترین ویژگی‌های شبکه‌های خودسازمانده، عدم نیاز به نظارت در هنگام آموزش است. در این شبکه‌ها، داده‌های خروجی با آنچه باید ایجاد شود، بر اساس یک الگوریتم مقایسه می‌شوند و اگر نتیجه دلخواه حاصل نشود، وزن گره‌ها آن‌قدر تغییر می‌کند تا هدف نهایی به‌دست آید.

به‌طور خلاصه، الگوریتم آموزشی شبکه‌های عصبی خودسازمانده به صورت زیر است:

۱- محاسبه فاصله بین الگو (X) و تمام نورون‌های عصبی

$$d_{ij} = \|x_k - w_{ij}\| \qquad (۱)$$

۲- انتخاب نزدیک‌ترین نورون به عنوان نورون برنده

$$w_{ij}: d_{ij} = \min(d_{mn}) \qquad (۲)$$

۳- به روز رسانی هر نورون با توجه به تابع همسایگی

$$w_{ij} = w_{ij} + \alpha_h(w_{winner}, w_{ij})\|x_k - w_{ij}\| \qquad (۳)$$

مقدار ضریب آلفا سبب کاهش تاثیر وزن‌های غیر مشابه می‌شود.

۴- این روند تا زمانی که یک معیار توقف خاص به‌دست آید، تکرار می‌شود. اغلب معیار توقف، تعداد مشخصی از تکرار است. برای تثبیت همگرایی و ثبات نقشه، نرخ یادگیری و شعاع همسایگی در هر تکرار کاهش می‌یابد. بنابراین، همگرایی به سمت صفر میل خواهد کرد. فاصله اندازه‌گیری بین بردارها، فاصله اقلیدسی است ولی از سایر فواصل مثل ماهالانوبیس یا منهتن نیز می‌توان استفاده کرد[۱۳].

یکی از کاربردهای SOM در به تصویر کشیدن و شناسایی ارتباط بین پارامترهای ورودی است. ماتریس وزن‌های همسایه[۲۱] یا همان ماتریس فاصله یکپارچه[۲۲] یا U-matrix،

نشان‌دهنده فاصله بین نورون‌های نقشه است که از سیاه (فاصله صفر) تا قرمز (حداکثر فاصله) در تغییر است.

۲-۲. ماهیت ترکیبی داده‌های ژئوشیمیایی

با توجه به فاصله مورد استفاده در الگوریتم SOM که اقلیدسی است، این سوال پیش می‌آید که آیا کاربرد آن در برخی شاخه‌های داده‌کاوی مانند علوم زمین و به‌خصوص تحلیل داده‌های ژئوشیمیایی نیز قابل گسترش است؟ ماهیت ترکیبی یا بسته داده‌های ژئوشیمی، امروزه موضوع مهمی است که بایستی قبل از هر تحلیل مدنظر قرار گیرد. داده‌های بسته یا ترکیبی، مجموعه‌ای از داده‌ها است که متغیرهای آن مستقل از یکدیگر نبوده و به صورت درصد یا قسمت در میلیون یا جزئی از یک کل بیان می‌شوند[۱۴]. در تعریف کلاسیک، هر ردیف از داده‌ها یک مشاهده نام دارد که مجموع متغیرهای آنالیز شده آن مشاهده، یک عدد ثابت (مثل ۱، ۱۰۰یا 10^6) می‌شود. اگرچه با توجه به دو خاصیت عمده این داده‌ها یعنی ثبات مقیاس و یکپارچگی زیرمجموعه‌های آن، لزوما نیازی به برقراری شرط مجموع ثابت نیست[۱۵].

داده‌های ترکیبی دارای خواصی هستند که کاربرد روش‌های آماری استاندارد برای تحلیل آن‌ها را مشکل می‌کند. فضای اقلیدسی برای داده‌های ترکیبی مناسب نیست و محدودیت حاصل جمع ثابت این داده‌ها دلالت بر هندسه خاصی دارد که در اصطلاح، هندسه آیچیسون در محیط سیمپلکس[۲۳] نامیده می‌شود[۱۶]. به‌منظور بکارگیری روش‌های آماری استاندارد، بایستی بر روی این داده‌ها تبدیلات مناسبی صورت گیرد. از جمله این تبدیلات، خانواده تبدیلات لگاریتم نسبیتی[۲۴] است که نخستین بار توسط آیچیسون[۲۵](۱۹۸۶) ارائه شده است. فضای نمونه داده‌های بسته یا سیمپلکس برای یک ترکیب D-جزئی $X = (x_1, \dots, x_D)$ یا D-بعدی طبق رابطه (۴) تعریف می‌شود:

$$S^D = \left\{ X = (x_1, \dots, x_D) \,\middle|\, x_i > 0, i = 1,2, \dots, D; \sum_{i=1}^{D} x_i = k \right\} \quad (۴)$$

با استفاده از سه نوع تبدیل لگاریتم نسبیتی alr[۲۶]، clr[۲۷] و ilr[۲۸] می‌توان این داده‌ها را به فضای اقلیدسی انتقال داد. به عنوان مثال، تبدیل ilr که منجر به یک مشاهده چندمتغیره درفضای $D\text{-}1$ بعدی می‌گردد، مطابق رابطه (۵) تعریف می‌شود:

$$ilr(x) = (z_1, \dots, z_{D-1}) = \sqrt{\frac{D-i}{D-i+1}} ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad for\ i = 1, \dots, D-1 \quad (۵)$$

فاصله داده‌ها دراین هندسه، فاصله آیچیسون است که برای دوترکیب $X=(x_1,\dots,x_D)$ و $Y=(y_1,\dots,y_D)$ از رابطه (۶) محاسبه می‌شود:

$$d_A(X,Y) = \sqrt{\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\left(ln\frac{x_i}{x_j} - ln\frac{y_i}{y_j}\right)^2} \quad (۶)$$

خاصیت ایزومتری تبدیلilr یعنی به ازای دوترکیب X وY رابطه (۷) بین دو فضای آیچیسون و اقلیدسی آن‌ها برقرار است:

$$d_A(X,Y) = d_E\big(ilr(X), ilr(Y)\big) \quad (۷)$$

برای تفسیر صحیح نمودارهایی که از بردارهای تک‌متغیره استفاده می‌کنند (مثل هیستوگرام، باکس پلات و غیره که روابط اقلیدسی بر آن‌ها حاکم است) رابطه دیگری نیز وجود دارد که توسط پروفسور فیلزموزر برای تبدیل تک متغیره داده-ها، پیشنهاد شده است[۱۶] که به ازای هر متغیر x_i مطابق رابطه (۸) محاسبه می‌شود:

$$z_i = \sqrt{\frac{D-1}{D}} ln \frac{x_i}{\sqrt[D-1]{\prod_{j=2}^{D} x_j}} \quad (۸)$$

میانگین در این نوع داده‌ها مطابق رابطه (۹) محاسبه می‌شود:

$$\bar{X} = clr^{-1}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} clr(x_i)\right) \quad (۹)$$

معیارهای مختلفی برای واریانس در داده‌های ترکیبی تعریف شده است. واریانس متریک یا واریانس کل یا واریانس عام یکی از آن‌هاست که متوسط مربع فاصله از مرکز داده‌ها با درجه آزادی تصحیح شده واریانس داده‌های معمولی است و از رابطه (۱۰) به‌دست می‌آید:

$$var(X) = \frac{1}{n-1}\sum\nolimits_{i=1}^{n} d_A^2(x_i, \bar{X}) \quad (۱۰)$$

بنابراین استانداردکردن داده‌های ترکیبی نیز با روش‌های معمول آماری، بسیار متفاوت است. نخست اینکه داده‌های ترکیبی یک مقیاس مشترک بدون بعد دارند و در نتیجه با فرآیند معمول استاندارد کردن سبب از دست رفتن اطلاعات مهمی خواهد شد که از آن جمله می توان به تغییرپذیری داده-

ها اشاره کرد. دوم اینکه میانگین‌گیری معمولی، مقادیر منفی تولید می‌کند که تفسیر نتایج را با اشتباه همراه خواهد کرد. در نتیجه، استانداردکردن این داده‌ها با تقسیم بر میانگین ترکیبی به توان عکس مجذور واریانس ترکیبی از رابطه (۱۱) به‌دست می‌آید:

$$Z = \frac{1}{\sqrt{var(X)}} \odot (X \ominus \bar{X}) \quad (11)$$

که در این رابطه، $\odot$ عملگر توان و $\ominus$ عملگر معکوس در فضای سیمپلکس هستند[۱۷].

۲-۳. دندروگرام داده‌های ترکیبی

از آنالیز خوشه‌ای می‌توان برای دسته‌بندی نمونه‌ها و به‌دست آوردن ایده‌ای از کیفیت تمرکز متغیرها در داده‌های با ابعاد زیاد استفاده کرد. به‌دلیل طبیعت پیچیده داده‌های ژئوشیمیایی ناحیه‌ای (نبود توزیع نرمال یا لاگ نرمال، چولگی زیاد، وجود توزیع چند مده و بسته بودن)، نتایج تحلیل خوشه‌ای بسیار وابسته به آماده‌سازی داده‌ها (انتخاب تبدیل صحیح) و الگوریتم خوشه‌بندی است[۱۸]. یکی از راه‌های آسان مطالعه داده‌هایی که فضای نمونه آن اقلیدسی است، نمایش آن‌ها در مختصات، نسبت به یک مبنای متعامد است. در هر فضای اقلیدسی تعداد بی‌نهایت مبنای متعامد موجود است که فضای سیمپلکس یکی از آن‌هاست. با روش‌های متفاوتی می‌توان چنین مبنایی را ساخت که روش SVD[۲۹] یا تجزیه مقادیر منفرد یکی از آن‌هاست [۱۹]. در مورد مسائلی که با ترکیب اجزاء سروکار دارند (مانند داده‌های ژئوشیمیایی) جستجو برای روش‌های کاهش بعد متناسب با زیر مجموعه‌های ترکیبی به استراتژی جدیدی منجر شده است که ماتریس موازنه‌ها[۳۰] نامیده می‌شود. این ماتریس‌ها بر اساس یک تفکیک D-جزئی دو دوئی ترتیبی یا (SBP[۳۱])، داده‌های ترکیبی را به گروه‌هایی غیر همپوشان تبدیل می‌کنند. مختصاتی که در نهایت به‌دست می‌آید، تفسیر ترکیب‌ها را آسان می‌کند و منجر به تجزیه واریانس کل به واریانس‌های جزئی می‌شود که می‌توان آن‌ها را ناشی از تغییرپذیری درون‌گروهی یا بین‌گروهی دانست[۲۰]. با استفاده از این خاصیت داده‌های ترکیبی، ابزاری به نام دندروگرام داده-های ترکیبی توسعه یافته است.

دندروگرام داده‌های ترکیبی، نمایش یک SBP همراه با خلاصه اطلاعات آماری موازنه هاست. اجزای مختلف آن عبارتست از:
۱- SBP که به‌صورت اتصالاتی از نوع دندروگرام بین اجزا نشان داده می‌شود. این قسمت از لحاظ فرم شبیه دندروگرام عادی است که خطوط افقی، گروه‌های اجزا را در ماتریس SBP توصیف می‌کنند. طول این خطوط هیچ اطلاعات کمی را نشان نمی‌دهد.

۲- مکان میانگین یک بالانس که در محل برخورد قسمت عمودی و افقی و در یک باکس پلات تعیین می‌شود.

۳- تجزیه پراش کل نمونه و میزان تغییرپذیری هر بالانس که با طول خطوط عمودی نشان داده می‌شود. جمع همه خطوط عمودی معرف پراش کل نمونه است. یک خط عمودی کوتاه به معنی پراش کم در آن نمونه و در نتیجه سهم جزئی در پراش کل است[۲۰].

دندروگرام داده‌های ترکیبی می‌تواند اطلاعات نمونه را حتی زمانی که بردارهای ترکیبی اجزای زیاد دارند، خلاصه کند. اگر همبستگی و ارتباط بین اجزاء، از قبل تا حدودی معلوم است در اینصورت ساخت ماتریس بالانس می‌تواند توسط شخص انجام شود تا تفسیر نتایج بهبود یابد[۲۱].

۳-معرفی منطقه مورد مطالعه

منطقه مورد مطالعه در استان خراسان جنوبی و به فاصله تقریبی ۳۵ کیلومتری جنوب‌غرب شهرستان بیرجند در محدوده طول‌هـای جغرافیایی ″۰۰ ′۳۰ °۵۸ تا ″۰۰ ′۰۰ °۵۹ و عرض‌هـای جغرافیایی ″۰۰ ′۳۰ °۳۲ تا ″۰۰ ′۰۰ °۳۳ واقع شده است. در این منطقه رخنمون‌های گسترده سنگ‌های رسوبی از کرتاسه پسین تا اوایل دوران سوم و نیز سنگ‌های آتشفشانی سنوزوئیک دیده می‌شود. شکل ۴ نقشه زمین‌شناسی ساده شده منطقه مورد مطالعه را نشان می‌دهد. (نقشه زمین-شناسی ۱:۱۰۰,۰۰۰ خوسف، سازمان زمین‌شناسی و اکتشافات معدنی کشور).

از نظر پتانسیل معدنی، اندیس‌هایی از مس، سرب و روی و خاک های صنعتی نظیر بنتونیت در محدوده مورد مطالعه وجود دارند. در قسمت‌های عمده‌ای از منطقه نیز لایه نازکی از خاک‌های رسی آهن‌دار، لایه ضخیم‌تری از ژیپس را پوشانده است. همچنین اثراتی از کانی سازی مس به شکل مالاکیت به‌صورت سطحی و مرتبط با دایک‌های آندزیتی سیلیسی در منطقه مشاهده شده است.

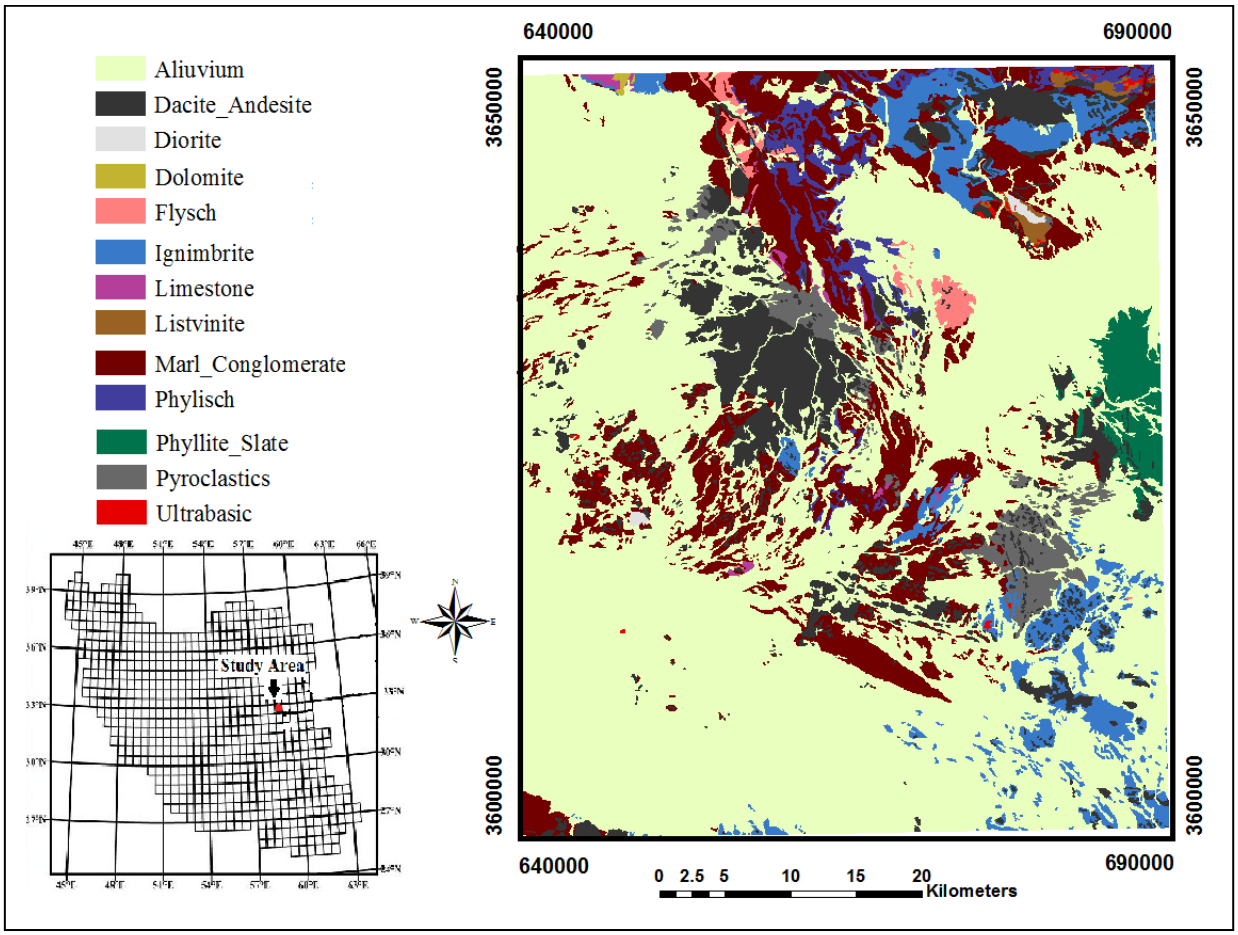
در محدوده ورقه مذکور، تعداد ۶۷۰ نمونه از رسوبات آبراهه‌ای منطقه برداشت و مورد آنالیز قرار گرفته است که در این پژوهش از نتیجه آنالیز این داده ها به‌عنوان مطالعه موردی استفاده شده است.

۴-بحث و بررسی

پس از اصلاح و آماده‌سازی داده‌ها (جایگزینی مقادیر BDL[۳۲] و مفقود با الگوریتم‌های مناسب برای این نوع داده‌های ترکیبی)، ماتریسی شامل ۶۷۰ نمونه به‌دست آمد. به‌دلیل ماهیت ترکیبی داده‌ها این سوال مطرح بود که آیا می توان از SOM با وجود معیار فاصله اقلیدسی آن در آموزش نورون‌ها، در خوشه‌بندی این نوع داده‌ها هم استفاده کرد؟

با بررسی‌های انجام شده روی این مسئله توسط محققین [۲۲،۲۳] ثابت شده است که ماهیت ترکیبی داده‌ها در تعریف وزن‌های نورون‌ها با وجود استفاده از معیار فاصله اقلیدسی، تغییر نمی‌کند و می‌توان از داده‌ها به‌صورت خام ولی استاندارد شده در شبکه عصبی SOM استفاده کرد.

کورتز (۲۰۰۸) نشان داد که بردارهای مدل تحت شرایط اقلیدسی انطباق بهتری بر بردارهای ورودی نشان می‌دهند و توزیع همگن‌تری می‌یابند زیرا فاصله اقلیدسی بین همسایه‌ها ثابت است.



شکل۴: نقشه زمین‌شناسی ساده‌شده منطقه مورد مطالعه (بر گرفته از برگه زمین شناسی۱:۱۰۰،۰۰۰ خوسف، سازمان زمین شناسی)

به علاوه، ناهمگنی توزیع نقاط تحت هندسه آیچیسون نشان می‌دهد که آموزش نورون‌ها برای جانمایی آن‌ها در فضای کلی باید بیشتر از موقعی باشد که از فاصله اقلیدسی تبعیت می‌کنند [۲۲].

بنابراین این داده‌ها با استفاده از نرم‌افزار R[۲۴]، یکبار با روش تحلیل ترکیبی، استاندارد شده و بار دیگر همان داده‌های خام، تحت تبدیل تک‌متغیره قرار گرفت. بدین ترتیب دو نوع ماتریس داده برای ورود به شبکه، آماده شد.

با استفاده از کدنویسی در نرم‌افزار MATLAB که امکانات مناسبی از شبکه‌های عصبی را در اختیار می‌گذارد، توپولوژی شبکه نگاشت به صورت شش‌ضلعی با ابعاد ۸ نورون در ۸ نورون (در مجموع ۶۴) یعنی تقریبا ۱۰ درصد تعداد داده‌ها بنا بر پیشنهاد کوهونن جهت کارآیی بهتر شبکه [۳]، طراحی شد.

تابع فاصله همان پیش فرض اقلیدسی و تعداد تکرارها در الگوریتم آموزش شبکه برای فضای ورودی، ۱۰۰ در نظر گرفته شد. نرخ یادگیری تابع آموزش، ۰/۰۱ و شعاع همسایگی اولیه برابر ۴ قرار داده شد.

نقشه ماتریس U یا فواصل وزنی همسایه برای شبکه ۸ نورونی مربوط به تبدیل‌های تک‌متغیره داده‌های ترکیبی (چپ) و ماتریس استانداردشده داده‌های ترکیبی (راست) رسوبات آبراهه‌ای خوسف در شکل (۵) نشان داده شده است.

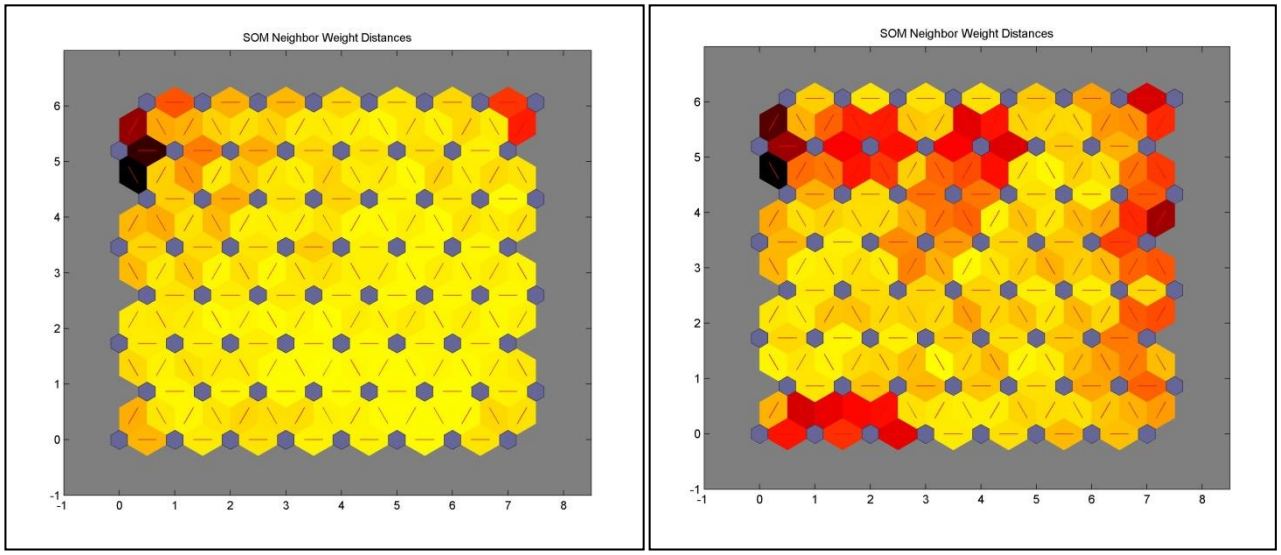
شش ضلعی‌های آبی رنگ معرف نورون‌ها هستند. خطوط قرمز اتصال بین آن‌ها و رنگ‌های هر منطقه، میزان فاصله را نشان می‌دهند. رنگ‌های تیره‌تر نشان دهنده فواصل بیشتر (دورترین فاصله = سیاه) و رنگ‌های روشن‌تر فواصل کمتر (نزدیکترین فاصله = قرمز) است. تجمع هر گروه از رنگ‌های روشن، می‌تواند معرف یک خوشه باشد که با مرزهای تیره‌رنگ از هم جدا شده‌اند.

در شکل ۵ (سمت راست) یک خوشه بزرگ تمام صفحه را گرفته است (به رنگ زرد) و با فواصل تیره کم رنگی در گوشه بالای چپ نقشه تفکیک شده است. این نشان‌دهنده اینست که عمده تراکم در دو یا سه خوشه متمرکز است. گرچه نمی‌توان تعداد دقیق آن را مشخص کرد.

در صورتی‌که در شکل سمت چپ، جدا شدگی‌ها واضح‌تر است. این امر می‌تواند به‌دلیل باز شدن تک متغیره داده‌های ترکیبی باشد که بر روابط بین متغیرها تاثیرگذار است.

شکل ۶ تعداد نمونه‌های متمرکز در هر نورون نقشه SOM داده‌های تبدیل تک متغیره ترکیبی و استاندارد شده ترکیبی رسوبات آبراهه‌ای منطقه خوسف را نشان می‌دهد.

هرچه این نقشه توزیع یکنواخت‌تری داشته باشد، خوشه‌بندی نتیجه بهتری داشته است[۲۵].



شکل۵: نقشه ماتریس U مربوط به تبدیل‌های تک‌متغیره داده‌های ترکیبی (چپ) وماتریس استانداردشده داده‌های ترکیبی (راست)

در هر دو نقشه، سه یا چهار ناحیه می‌توان یافت که تمرکز نمونه‌ها در آن بیشتر است. گرچه مکان این تجمع‌ها با هم اختلاف اندکی دارد.

به ازای هر متغیر در بردار ورودی، یک صفحه وزن وجود دارد که هر ورودی را به هر نورون متصل می‌کند. رنگ‌های تیره‌تر معرف وزن بزرگ‌تر و رنگ‌های روشن‌تر معرف وزن کوچک‌تر است. اگر الگوهای اتصال دو ورودی خیلی مشابه باشند می‌توان فرض کرد که همبستگی بالایی بین آن‌ها وجود دارد[۲۵].

بر اساس نقشه‌های صفحات وزن داده‌های استاندارد شده ترکیبی در شکل (۷)، متغیرهای همبسته در خوشه‌های این تحقیق را می‌توان به‌صورت زیر در نظر گرفت:

۱- Zn, Cr, Ni, Bi, Sc, As, Cd, Co, V, Hg,Fe_2O_3,MnO, Cu

۲- W, Mo

۳- Sn, TiO_2 .

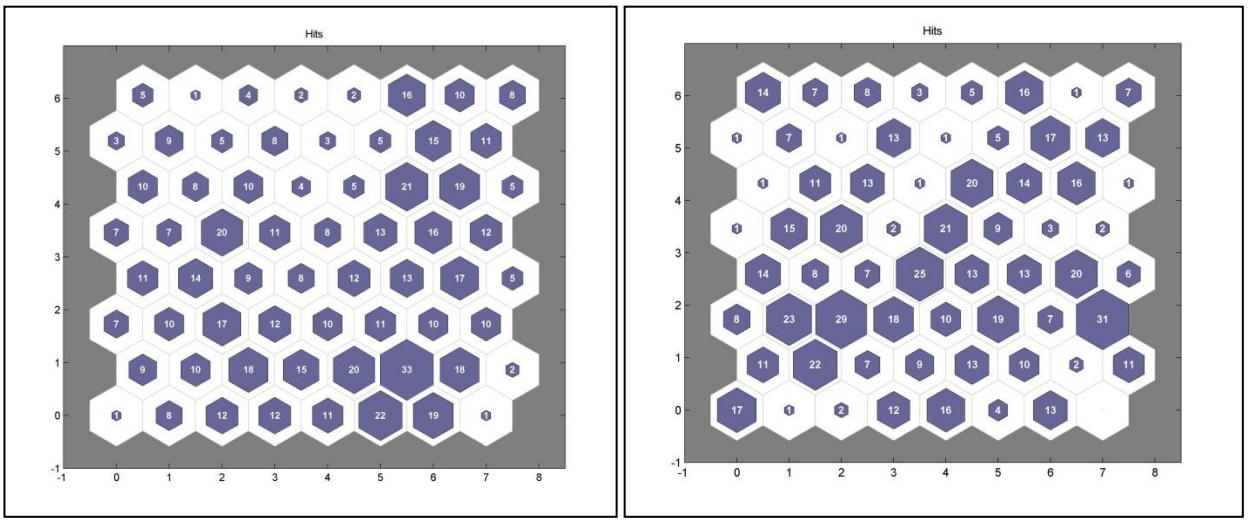
نقشه‌های صفحات وزن داده‌های تبدیل تک متغیره ترکیبی رسوبات آبراهه‌ای خوسف در شکل ۸ نشان داده شده است. الگوهای اتصال در این مورد بسیار متفاوت‌ترند و در بعضی موارد نظیر این حالت، یافتن الگوهای وزنی مشابه دشوار می‌شود. گرچه باز هم می‌توان فاکتور یا فاکتورهای اصلی اثرگذار بر داده‌ها را تشخیص داد. خوشه‌های قابل تشخیص متغیرها عبارتند از:

۱- Cr, Co, TiO_2

۲- As, Cd, Sr

۳- Zn, Ni, Hg, Fe_2O_3

۴- Sb, Ba, Cu.

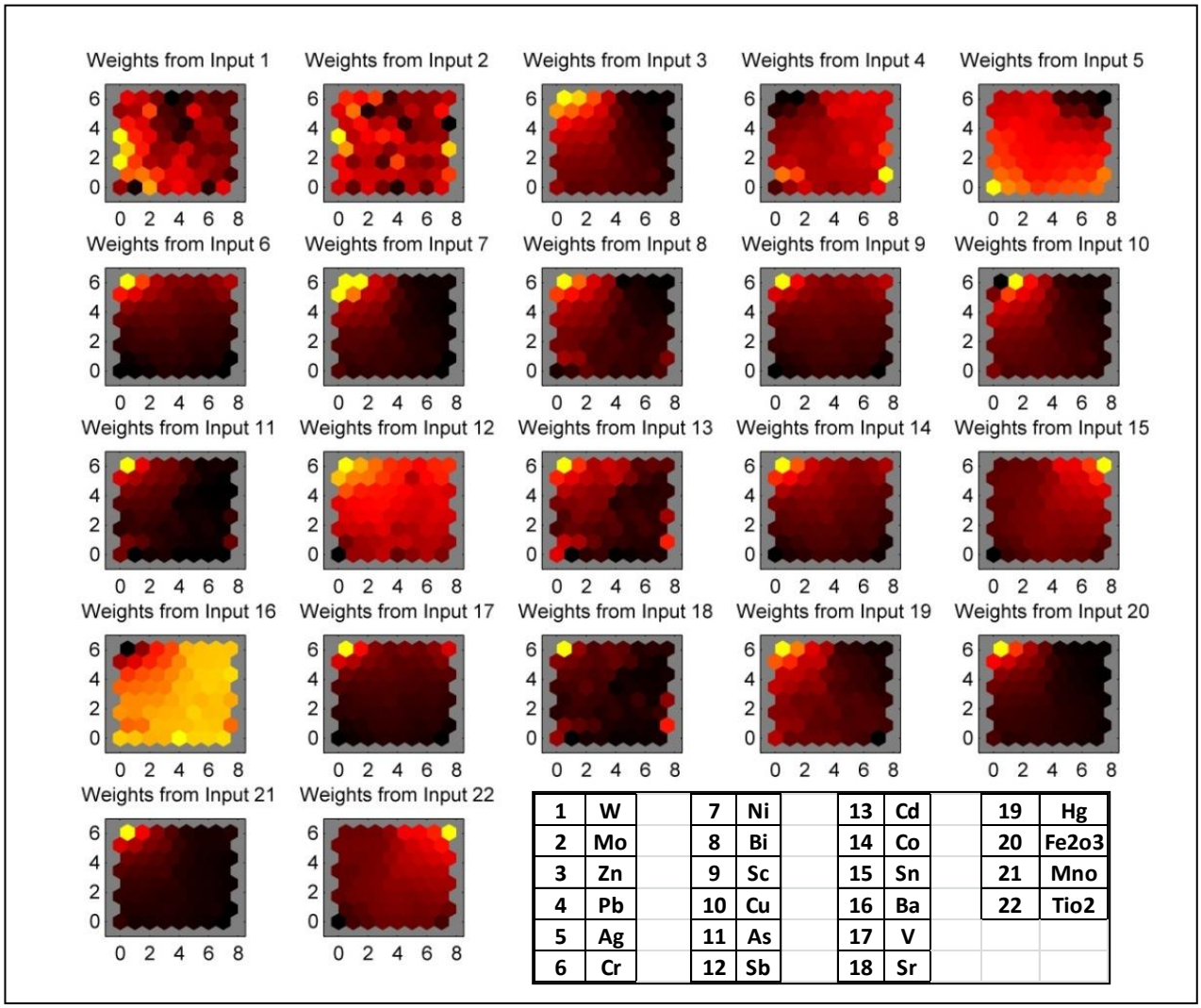


شکل۶: تعداد نمونه‌های متمرکز در هر نورون نقشه SOM مربوط به داده‌های تبدیل تک‌متغیره ترکیبی (چپ) واستانداردشده ترکیبی (راست)

در مرحله بعد، با کمک نرم افزار R تحلیل خوشه‌ای روی داده-های ترکیبی انجام شد. این کار همانند یک آنالیز خوشه‌ای ساده روی داده‌های غیر ترکیبی است با این تفاوت که اختلاف فواصل و پراش‌ها و میانگین‌های مجاز در فضای سیمپلکس در ساختار دندروگرام منظور می‌شود.

از معیارهای مختلفی برای فاصله در خوشه‌بندی می‌توان استفاده کرد مثل ماهالانوبیس[۳۳]، کانبرا[۳۴]، گاور[۳۵]، بری[۳۶]، کورد[۳۷]، موریزیتا[۳۸]، هورن[۳۹]، همبستگی و .... اما به پیشنهاد بوگارت بهتر است از فواصل منهتن[۴۰] استفاده شود که به‌طور کامل با فضای سیمپلکس سازگار است. ضمن اینکه خوشه‌بندی همیشه بر اساس فاصله اقلیدسی انجام نمی‌شود و بنابراین استفاده از فاصله آیچیسون در تحلیل ترکیبی خوشه‌بندی یک "باید" نیست. به همین دلیل نتایج در این حالت با تحلیل داده‌های ترکیبی فرقی ندارد و یکسان است[۲۶،۲۷].

برای الگوریتم خوشه‌بندی، از روش وارد[۴۱] که یک تحلیل سلسله مراتبی است استفاده شد. این روش که به کمترین واریانس نیز مشهور است، حالت خاصی از روش تابع هدف است که معیار انتخاب جفت خوشه‌ها برای ترکیب در هر مرحله، بر اساس مقدار بهینه یک تابع هدف است. این تابع هدف می‌تواند هر تابع مدنظر کاربر باشد. اغلب متداولترین تابع هدف مورد استفاده، تابع مجموع مربعات خطا[۴۲] است. این تابع سبب کمینه‌کردن پراش کلی بین‌خوشه‌ای می‌شود[۲۸].



1	W	7	Ni	13	Cd	19	Hg
2	Mo	8	Bi	14	Co	20	Fe2o3
3	Zn	9	Sc	15	Sn	21	Mno
4	Pb	10	Cu	16	Ba	22	Tio2
5	Ag	11	As	17	V		
6	Cr	12	Sb	18	Sr		

شکل۷: نقشه‌های صفحات وزن داده‌های استانداردشده ترکیبی ژئوشیمی رسوبات آبراهه‌ای برگه خوسف

نتایج دندروگرام به‌دست‌آمده با تبدیل داده‌ها و استانداردسازی ترکیبی در شکل ۹ نشان داده شده‌است.

همان‌طور که در نمودار ملاحظه می‌شود، میزان پراش کل و پراش‌های جزء در محور عمودی و میانگین اجزاء در محل اتصال (نمودارهای جعبه‌ای) نشان داده شده است. به‌طور کلی سه گروه اصلی عناصر قابل تفکیک است:

۱- Fe_2O_3, TiO_2, V, Sr, Ba, MnO

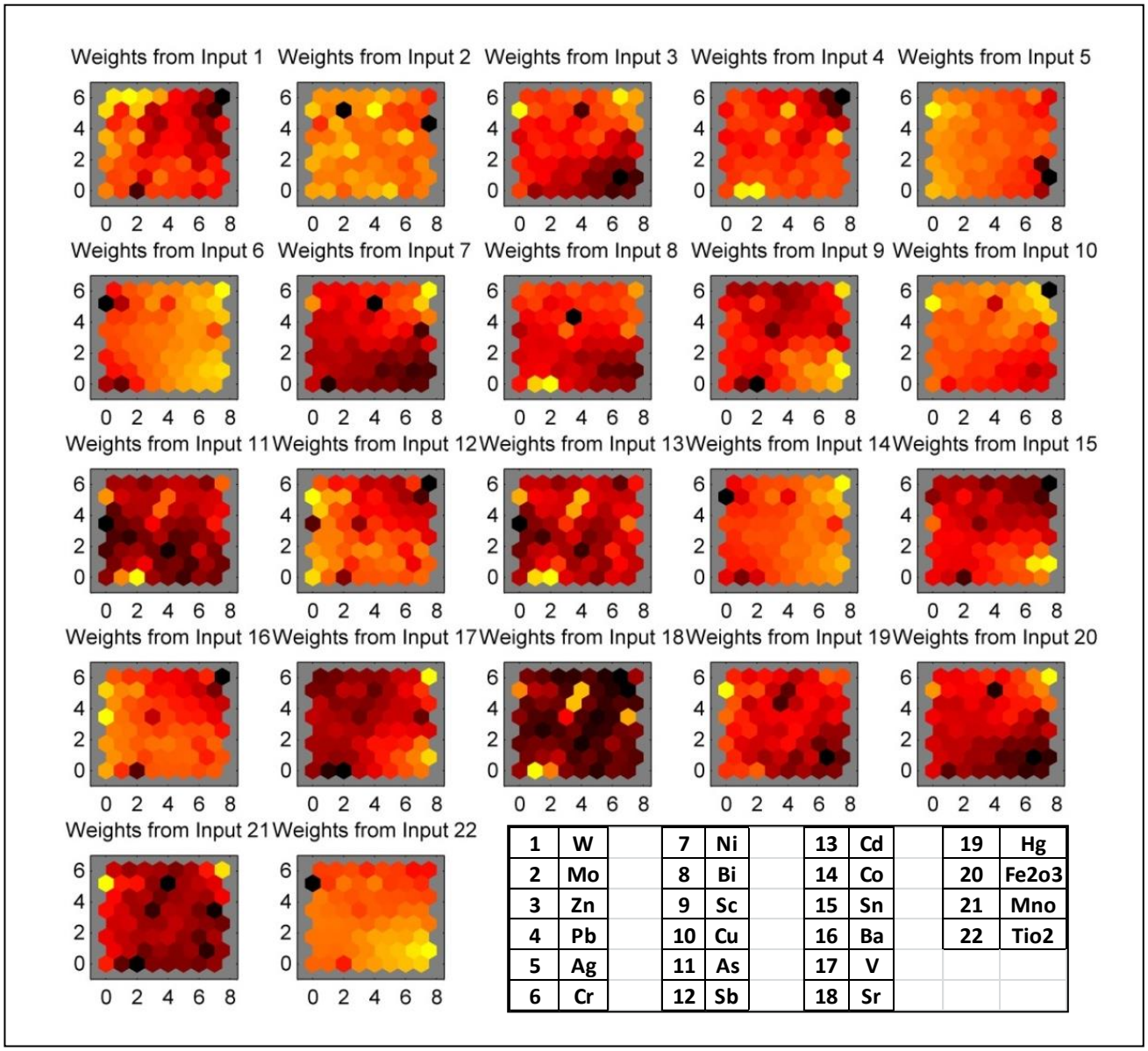
۲- Cr, Co, Cu, Zn, Ni, Pb, Sc, Sn, As, Sb (با کمترین پراش نسبت به بقیه گروه‌ها)

۳- Hg, Cd, W, Mo, Ag, Bi (با بیشترین پراش)

۵- نتیجه‌گیری

مقایسه نتایج نگاشت‌ها نشان می‌دهد که کاربرد SOM روی داده‌های خام استاندارد، نتایج قابل قبول‌تری می‌دهد و انطباق بهتری با دندروگرام دارد.

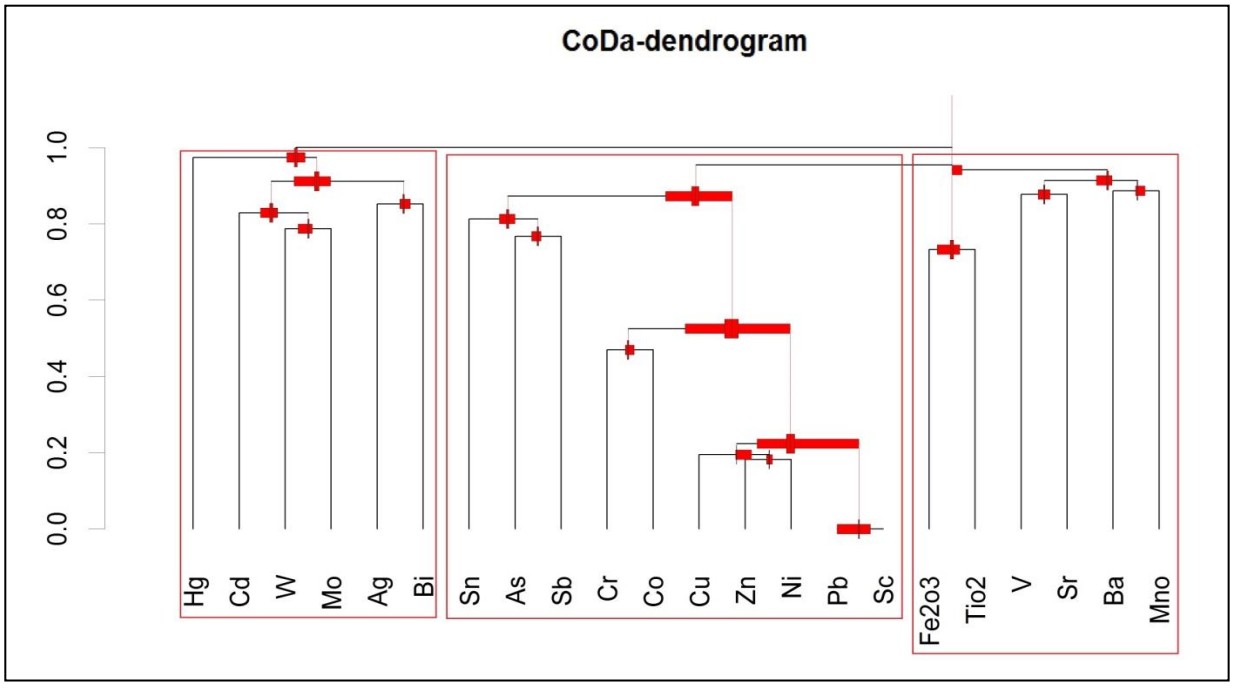
مهمترین برتری بکارگیری روش SOM نسبت به روش‌های دیگر کاهش بعد، اول عدم نیاز به فرض توزیع نرمال داده-هاست که لازمه تفسیر صحیح آمار کلاسیک است و دوم عدم نیاز به بازکردن داده‌های ترکیبی است که اغلب حساسیت و دشواری خاص خود را دارد و ابزارهای محاسباتی پیچیده نیاز دارد. گرچه به‌دلیل طبیعت ترکیبی داده‌های ژئوشیمیایی، باز هم نیاز به داده‌های استاندارد شده به روش تحلیل ترکیبی است. اشکال عمده آن نیز تولید بعضی الگوهای پیچیده در



1	W		7	Ni		13	Cd		19	Hg
2	Mo		8	Bi		14	Co		20	Fe2o3
3	Zn		9	Sc		15	Sn		21	Mno
4	Pb		10	Cu		16	Ba		22	Tio2
5	Ag		11	As		17	V			
6	Cr		12	Sb		18	Sr			

شکل۸: نقشه‌های صفحات وزن داده‌های تبدیل تک‌متغیره ترکیبی ژئوشیمی رسوبات آبراهه‌ای برگه خوسف

خروجی شبکه است که تفسیر آن را دشوار می‌کند. البته می‌توان با تنظیم پارامترهای شبکه عصبی، یا تلفیق با روش‌های دیگر مثل منطق فازی و طراحی الگوریتم‌های تشخیص الگو به نتایج بهتری دست یافت و فاکتورهای مؤثر بر ژئوشیمی یک منطقه را نشان داد. نکته دیگر در نوع داده مورد استفاده است که می‌توان این روش را با داده‌های لیتوژئوشیمیایی نیز مورد آزمایش قرار داد. در این‌صورت، احتمالاً به‌دلیل ارتباط منطقی‌تر نمونه‌ها با ترکیب سنگ‌های منطقه، خوشه‌بندی دارای نتایج گویاتری خواهد بود.



شکل۹: دندروگرام داده‌های ترکیبی بازشده با تبدیل *clr* در فضای سیمپلکس (خوشه‌های پیشنهادی در مستطیل‌های مجزا دیده می‌شوند)

مراجع

[1] Templ, M.; Filzmoser, P.; Reimann, C.; 2008; *"Cluster analysis applied to regional geochemical data: Problems and possibilities*". Applied Geochemistry, 23(8), 2198-2213.

[2] Liu, L.; Zhou, J.Z.; An, X.L.; Li, Y.H.; Liu, Q.; 2008; *"Improved fuzzy clustering method based on entropy coefficient and its application"*. In: Sun, F., Zhang, J., Tan, Y., Cao, J., Yu, W. (eds.) ISNN 2008,, Part II. LNCS, vol. 5264, Springer, Heidelberg, pp. 11–20.

[3] Kohonen, T.; 1990; "*The self-organizing map"*. Proceedings of the IEEE, 78(9), 1464-1480.

[4] Ji C. Y; 2000; *"Land-use classification of remotely sensed data using Kohonen self-organizing feature map neuralnetworks",* Photogrammetric Engineering and Remote Sensing 66, pp. 1451–1460.

[5] Vilmann T.; Merenyi E.; Hammer B.; 2003; *"Neural maps in remote sensing image analysis",* Neural Networks 16, pp. 389-403.

[6] Cassano E. N.; Lynch A. H.; Cassano J. J.; Koslow M. R.; 2006; *"Classifiation of synoptic patterns in the*

western Arctic associated with extreme events at Barrow", Alaska, USA. Climate Research 30, pp. 83-97.

[7] Fayos J.; Fayos C.; 2007; *"Wind data mining by Kohonen neural networks"*, PLoS ONE 2, pp. 210.

[8] Strecker U.; Uden R.; 2002; *"Data mining of 3D poststack seismic attribute volumes using Kohonen self-organizing maps*", The Leading Edge 21, pp. 1032-1037.

[9] Cole ou T.; Poupon M.; Azbe K.; 2003; *"Unsupervised seismic faciesclassifiation: a review and comparison oftechniques and implementation*", The Leading Edge 22, pp. 942–953.

[10] Li, L.; Wang, Y.; 2014; *"What drives the aerosol distribution in Guangdong-the most developed province in Southern China?"*, Scientific reports, 4.

[11] Vesanto, J.; 1999; *"SOM-based data visualization methods"*, Intelligent data analysis, 3(2), 111-126.

[12] Kohonen T.; 2001; *"Self-Organizing Maps",* Springer series in Information Sciences, New York, Springer-Verlag,Vol. 30, pp. 501.

[13] Kohonen T.; Kaski S.; Lappalainen H.; 1997; "*Self-organized formation of various invariantfeaturefiters in the adaptive-subspace SOM"*, Neural Computation 9, pp. 1321-1344.

[14] Filzmoser, P.; Hron, K.; 2009; *"Correlation analysis for compositional data"*, Mathematical Geosciences, 41(8), 905-919.

[15] Buccianti, A.; Pawlowsky-Glahn, V.; 2005; *"New perspectives on water chemistry and compositional data analysis".* Mathematical Geology, 37(7), 703-727.

[16] Filzmoser, P.; Hron, K.; Reimann, C.; 2009; *"Univariate statistical analysis of environmental (compositional) data: problems and possibilities",* Science of the Total Environment, 407(23), 6100-6108.

[17] Van den Boogaart, K. G.; Tolosana-Delgado, R.; 2013; "*Analyzing compositional data with R",* Berlin: Springer.

[18] Pawlowsky-Glahn, V.; Egozcue, J. J.; 2011; *"Exploring compositional data with the coda-dendrogram*", Austrian Journal of Statistics, 40(1-2), pp. 103-113.

[19] Egozcue, J. J.; Pawlowsky-Glahn, V.; 2006; *"Exploring compositional data with the CoDa-dendrogram".* In E. Pirard, A. Dassargues, and H. B. Havenith (Eds.), Proceedings of IAMG'06 – The XI Annual Conference of the International Associationfor Mathematical Geology. Liege: University of Li` ege, Belgium, CD-ROM.

[20] Egozcue, J. J.; Pawlowsky-Glahn, V.; 2005; *"CoDa-dendrogram: a new exploratorytool",* In G. Mateu-Figueras and C. Barcelo-Vidal (Eds.), Compositional Data Analysis Workshop - CoDaWork'05, Proceedings; Girona: Universitat de Girona. (http://ima.udg.es/Activitats/CoDaWork05/)

[21] Thio-Henestrosa, S.; Egozcue, J. J., Kov´acs, V. P.-G. O.; Kov´acs, G.; 2008; *"Balance-´ dendrogram. a new routine of CoDaPack*", Computer and Geosciences, 34, 1682-1696.

[22] Cortés, J. A.; Palma, J. L.; 2008; *"Using Self Organizing Map with geochemical compositional data",* In AGU Fall Meeting Abstracts (Vol. 1, p. 2159).

[23] Cracknell, M. J.; Reading, A. M.; 2014; *"Unsupervised clustering of continental-scale geophysical and geochemical data using Self-Organising Maps*", In 3rd Australian Regolith Geoscientists Association Conference, pp. 20-24.

[24] R Core Team; 2014; R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (http://www.R-project.org/).

[25] Mathworks; 2012, Matlab Neural Netwrks Toolbox Tutorial, (http://www. Mathworks.com).

[26] van den Boogaart, K.G.; R. Tolosana-Delgado; 2008*;" "compositions": a unified R package to analyze Compositional Data",* Computers & Geosciences, 34 (4), pp. 320-338

[27] van den Boogaart, G.; Tolosana-Delgado, R.; 2005; *"A compositional data analysis package for R providing multiple approaches",* In Compositional Data Analysis Workshop–CoDaWork'05, Proceedings. Universitat de Girona, ISBN ,pp. 84-8458.

[28] Gordon, A. D.; 1999; *"Classification",* 2nd Edition, Chapman and Hall, Boca Raton.

[۱] Self-Organizing Maps

[۲] Between-groups linkage

[۳] Within-groups linkage

[۴] Nearest Neighbor

[۵] Furthest Neighbor
[۶] Compositional data
[۷] Feedforward
[۸] Feedbackward
[۹] Unsupervised learning
[۱۰] Input space
[۱۱] Node
[۱۲] Data space
[۱۳] Low-dimensional grid
[۱۴] Input
[۱۵] Output
[۱۶] Layer
[۱۷] Vector Quantization
[۱۸] Brain theory
[۱۹] Winner
[۲۰] Smooth
[۲۱] Neighbor weights
[۲۲] Unified Distance
[۲۴] Simplex
[۲۴] Log-ratio transformation
[۲۶] John Aitchison
[۲۷] Additive logratio
[۲۸] Centered logratio
[۲۹] Isometric logratio
[۲۹] Singular Value Decomposition
[۳۰] Balances
[۳۱] Sequential Binary Partition
[۳۲] Below Detection Limit
[۳۳] Mahalanobis
[۳۵] Canberra
[۳۶] Gower
[۳۷] Bray-Curtis
[۳۸] Chord
[۳۹] Morisita
[۴۰] Horn
[۴۱] Manhattan
[۴۲] Ward
[۴۳] Error Sum of Squares