تخمین مقادیر آنومال به کمک ترکیب مناسبی از روش جدایش فواصل ماهالانوبیس و سه روش پرکاربرد داده‌کاوی؛ مطالعه موردی: پرکام

سید سعید قنادپور [۱]، اردشیر هزارخانی* [۲]، ترانه رودپیما [۳]

۱- دانشجوی دکتری اکتشاف معدن، دانشکده مهندسی معدن و متالورژی، دانشگاه صنعتی امیرکبیر تهران

۲- استاد، دانشکده مهندسی معدن و متالورژی، دانشگاه صنعتی امیرکبیر تهران

۳- دانشجوی کارشناسی ارشد اکتشاف معدن، دانشکده مهندسی معدن و متالورژی، دانشگاه صنعتی امیرکبیر تهران

(دریافت: دی ۱۳۹۴، پذیرش: تیر ۱۳۹۶)

چکیده

در مطالعه پیش رو به منظور کاهش خطا و ریسک در راستای صرف هزینه، زمان، انرژی و نیز دستیابی به پیشگوئی‌هایی به مراتب ارزنده‌تر، به بررسی ترکیب روش‌های داده‌کاوی و جدایش آنومالی پرداخته می‌شود. اهمیت تشخیص مقادیر آنومال از زمینه بر هیچ یک پوشیده نیست، به این منظور روش‌های متعددی ابداع گشته است که از آن جمله می‌توان به روش جدایش فواصل ماهالانوبیس اشاره کرد که روشی مؤثر و چند متغیره در جدایش مقادیر آنومال از زمینه محسوب می‌شود. از طرفی، پیش‌بینی ابزاری قدرتمند در فرآیند برنامه‌ریزی در هر فعالیتی هست، پس به کارگیری روش‌های داده‌کاوی در جهت یافتن الگو و روابط نهفته در دل داده‌ها، نیاز ما را در این زمینه مرتفع می‌سازد. لذا در مطالعه حاضر، به بررسی عملکرد ترکیب روش جدایش فوق با سه روش داده‌کاوی K-نزدیک‌ترین همسایه، طبقه‌بند ساده بیز و درخت تصمیم‌گیری پرداخته می‌شود. به‌این‌ترتیب که پس از جدایش مقادیر آنومال مس و مولیبدن در مورد ۳۷۷ نمونه حاصله از عملیات نمونه‌برداری سطحی در محدوده پرکام به کمک روش فواصل ماهالانوبیس، به منظور پیش‌بینی این مقادیر برای هر نمونه تصادفی، سه روش داده‌کاوی مذکور، مورد استفاده قرار می‌گیرند. در نهایت نیز جهت بررسی شبکه‌های طراحی شده، نمونه‌های آموزشی به عنوان داده‌های تست در اختیار شبکه‌های مذکور قرار گرفته‌اند. نتایج حاصله نشان می‌دهند که روش درخت تصمیم‌گیری به مراتب قوی‌تر ظاهر شده، زیرا در شبکه طراحی شده توسط این روش، تنها دو نمونه از بین ۳۷۷ نمونه، اشتباهاً شناسایی شده‌اند که نشان دهنده دقت بالای شبکه طراحی شده است. یعنی مقدار خطای *Resubstitution* گزارش شده برای این شبکه برابر با ۰/۰۰۵۳ هست. لازم به ذکر است که تعداد نمونه‌های به اشتباه پیش‌بینی شده برای دو روش *KNN* و بیز به ترتیب برابر با ۹ و ۲۳ و به تبع، مقدار خطای محاسبه شده برای آنها نیز به ترتیب برابر با ۰/۰۲۳۹ و ۰/۰۶۱ گزارش شده‌اند. به این ترتیب با توجه به میزان خطای به مراتب قابل‌قبول‌تر برای شبکه طراحی شده توسط ترکیب روش درخت تصمیم‌گیری و فواصل ماهالانوبیس، ترکیب مذکور به عنوان روشی قابل اطمینان و سودمند جهت رسیدن به صحیح‌ترین پیشگوئی‌ها به تصمیم گیران این صنعت معرفی شده است.

کلید واژه‌ها

تخمین، مقادیر آنومال، جدایش، فواصل ماهالانوبیس، داده‌کاوی.

* عهده دار مکاتبات: *ardehez@aut.ac.ir*

۱- مقدمه

اهمیت تشخیص مقادیر آنومال در عملیات اکتشافی معدن، امری غیرقابل انکار هست، چرا که این مقادیر، که از بقیه داده‌ها قابل تفکیک هستند، مناطق امیدبخش در راستای رسیدن به کانی‌سازی اقتصادی هست. به این منظور جدایش ناهنجاری ها از مقادیر عادی و زمینه در دو شکل ساختاری و غیرساختاری صورت می گیرد. روش‌های آماری مختلفی برای جداسازی و تشخیص محدوده‌های آنومال از زمینه توسعه یافته و توسط محققین ارائه شده است [۱-۴]. روش جدایش بر اساس فواصل ماهالانوبیس از جمله روش‌های غیرساختاری محسوب می‌شود. مسئله حائز اهمیت در این روش، چند متغیره بودن آن هست، به این معنا که تعیین مقادیر آنومال زمانی شکل می گیرد که تمامی متغیرهای مورد نظرمان آن ها را تائید کرده باشند. از جمله مطالعات صورت گرفته در این زمینه می‌توان به استفاده از روش ماهالانوبیس برای تفکیک رخساره های نفتی در یکی از میادین هیدروکربوری ایران اشاره کرد که با احتمال موفقیت بیش از ۸۵ درصد کارایی این روش اثبات گردید [۵]، اجرای روش مذکور بر اساس الگوریتم فازی خوشه بندی برای تقسیم بندی تصویر و نمایش تأثیرگذاری این روش در مقایسه با روش‌های قدیمی بر پایه‌ی فاصله اقلیدسی [۶]، تشخیص گسل های اولیه برای مدارهای آنالوگ، بسیار مهم و در عین حال دشوار هست که پارامترهای آماری و بردارها در این خصوص از طریق فاصله ماهالانوبیس بهینه سازی شده اند [۷] و حتی کاربرد این تکنیک در تشخیص ناهنجاری برای *IGBT* ها *(Insulated-gate bipolar transistor)* مورد تائید قرار گرفته است [۸]؛ به این ترتیب این روش برخلاف سادگی از کارایی و اهمیت بالایی برخوردار هست که در مطالعه حاضر با یک کاربرد حائز اهمیت دیگر آن آشنا خواهیم شد.

از سویی با توجه به حضور طبیعت نامحدود و ناپیدا در علم معدنکاری، نیاز به پیش‌بینی و تخمین به شدت احساس می‌شود. پیش‌بینی امری ارزنده در فرایند هر کاری محسوب می‌شود، پس تشخیص مناطق آنومال، به ما در راستای رسیدن به الگوریتم پایداری جهت برآورد ارزش نمونه‌های تصادفی و جامعیت بخشیدن نتایج به مناطق تحت بررسی، یاری می رساند. در نتیجه همراهی روش مذکور با روش‌های داده‌کاوی جهت رسیدن به دستاوردی به مراتب کاراتر، الزامی خواهد بود. بدین منظور روش‌های داده‌کاوی متعددی ابداع شده اند، اما مسئله مهمی که می‌بایست مورد توجه قرار گیرد این است که پیشگوئی‌ها زمانی ارزشمند هستند که دارای بیشترین دقت باشند. در مطالعه حاضر جهت رسیدن به ترکیبی با چنین نتایج، به بررسی عملکرد روش جدایش فوق با هر یک از روش‌های داده‌کاوی K-نزدیک‌ترین همسایه، طبقه‌بند ساده بیز و درخت تصمیم‌گیری پرداخته شده است.

از روش‌های معروف خوشه‌بندی در داده‌کاوی، روش K-نزدیک‌ترین همسایه هست، که به طور کلی در حالت شناسایی، الگوریتم *KNN (K-Nearest Neighbor)* روشی برای کلاسه بندی هدف مورد نظر، بر اساس نزدیک‌ترین نمونه‌های تعلیم در فضای مشخص هست. کلاسه‌بندی توسط *KNN* زمانی توسعه یافت که نیاز به تجزیه و تحلیل های دقیق و مشخص احساس گردید. در واقع در شرایطی که تخمین‌های پارامترهای قابل اعتماد تراکم های احتمالی، ناشناخته اند و حتی برای مشخص کردن دشوارند، نیاز به این روش به مراتب بیشتر احساس می‌شود.

یکی از مهم ترین ابزارها برای پیاده‌سازی تکنیک های مختلف داده‌کاوی استدلال بیزی است. اهمیت استدلال بیز در داده‌کاوی را می‌توان به دلیل زیر نسبت داد. الگوریتم‌های یادگیری بیزی به طور صریح بر روی احتمالات فرض‌های مختلف کار می‌کنند، مانند *Naive Bayes Classifier* که از جمله کاراترین و عملی‌ترین الگوریتم‌های ممکن برای برخی مسائل یادگیری هست.

از دیگر روش‌های داده‌کاوی در راستای کلاسه‌بندی، الگوریتم درخت تصمیم هست که جزو مشهورترین الگوریتم های یادگیری استقرائی است و به صورت موفقیت آمیزی در کاربردهای مختلف بکار گرفته شده است. کلاسه‌بندی داده‌ها یک فرآیند دو مرحله‌ای است، در مرحله اول یک مدل ساخته می‌شود که مجموعه‌ای از کلاس‌های داده‌ای یا مفاهیم را مشخص می‌کند. این مرحله را مرحله یادگیری گوییم که در آن یک الگوریتم کلاسه‌بندی یک مدل را با تحلیل یک مجموعه آموزشی که مجموعه‌ای از تاپل‌های پایگاه است می سازد و برچسب کلاس‌های مربوط به این تاپل‌ها را مشخص می‌کند. یک تاپل X با یک بردار

صفت $X=(X_1, X_2,...,X_n)$ نمایش داده می‌شود. فرض می‌شود که هر تاپل به یک کلاس از پیش تعریف شده متعلق است و کلاس با یک صفت که به آن صفت برچسب کلاس می‌گوییم، مشخص می‌شود. مجموعه آموزشی به صورت تصادفی از پایگاه انتخاب می‌شود [۹]. در مرحله دوم، یادگیری از طریق یک تابع $y=f(x)$ انجام می‌شود که می‌تواند برچسب کلاس هر تاپل X از پایگاه را پیش‌بینی کند. این تابع به صورت قواعد کلاسه‌بندی، درخت‌های تصمیم‌گیری یا فرمول‌های ریاضی است [۹].

لذا در پژوهش پیش رو به منظور جدایش مقادیر آنومال از زمینه توسط روش فواصل ماهالانوبیس تحت دو متغیر عیار مس و عیار مولیبدن در محدوده پرکام، ابتدا می بایست نوع جوامع آماری مورد مطالعه را مشخص نموده و سپس اقدام به محاسبه میانگین و واریانس و در نهایت جدایش مقادیر آنومال از زمینه نماییم (به عنوان مثال می‌توان به مطالعه [۱، ۱۰] که به تعیین نوع توزیع جوامع و محاسبه پارامترهای آماری پرداخته‌اند اشاره نمود). بدین منظور می‌توان از نرم‌افزارهای پیشنهاد شده در مطالعات [۱۱] و [۱۲] جهت شناسایی نوع توزیع و نرمال‌سازی استفاده نمود. سپس به کمک روش فواصل ماهالانوبیس به جدایش مقادیر آنومالی پرداخته شده است، به این ترتیب پس از آن روش‌های داده‌کاوی K- نزدیک‌ترین همسایه، طبقه‌بند ساده بیز و درخت تصمیم‌گیری توسط داده‌های ۴ پارامتر عیار مس، عیار مولیبدن، طول و عرض جغرافیایی هر نمونه و نتایج حاصل از فواصل ماهالانوبیس به دست آمده شان، تحت آموزش قرار گرفته و در نتیجه معادلات پیشگویانه برای نمونه‌های تصادفی و احتمالی دیگر جهت تعیین آنومال بودن یا نبودنشان از هر یک، به دست می‌آید. در نهایت نیز با بررسی میزان خطای وابسته به شبکه طراحی شده در هر مورد، ترکیب با کمترین خطا و در نتیجه دارای عملکردی فوق‌العاده جهت دستیابی به توابع پیشگویانه قابل اعتماد را به تصمیم‌گیران این صنعت پیشنهاد و عرضه خواهیم کرد.

۲- معرفی محدوده پرکام

سیستم پورفیری پرکام (سارا) با مختصات جغرافیایی ″۵۴ ′۸ °۵۵ طول شرقی و ″۲۴ ′۲۶ °۳۰ عرض شمالی در ۵۰ کیلومتری شمال شهربابک و در کمربند فلززایی کرمان قرار دارد. کمربند کرمان که بخش جنوبی ایالت فلززایی ارومیه - دختر (سهند - بزمان) را تشکیل می‌دهد غنی‌ترین کمربند مس در ایران به شمار می رود. در این کمربند با طول حدود ۴۵۰ کیلومتر و پهنای حدود ۸۰ کیلومتر بیش از ۲۰۰ کانسار و نشانه معدنی شناخته شده است که تعدادی از آنها از جمله ذخیره پرکام از نوع پورفیری است [۱۳].

ذخیره پورفیری پرکام واقع در ۲ کیلومتری معدن مس پورفیری میدوک در نقشه‌های زمین‌شناسی ۱:۲۵۰۰۰۰ انار و ۱:۱۰۰۰۰۰ شهربابک قرار دارد. این محدوده بخشی از زون ارومیه - دختر (سهند - بزمان) است (شکل ۱). واحدهای سنگی در برگیرنده در محدوده مورد مطالعه عمدتاً شامل کمپلکس رسوبی - آتشفشانی ائوسن هست که میزبان توده‌های ساب ولکانیک و کانی سازی پورفیری است. در این ناحیه مجموعه سنگ های آتشفشانی، توف و پیروکلاستیک دارای ترکیب آندزیت بازالت و آندزیت هستند که گاه لایه بندی دارند. توده های نفوذی ساب ولکانیک منطقه شامل دیوریت و میکروکوارتزدیوریت پورفیری هست (شکل ۲) که با سیستم دگرسانی و کانی سازی ارتباط نشان می دهند [۱۴]. این واحدهای سنگی توسط دایک هایی از نوع دیوریت قطع شده‌اند. دگرسانی در محدوده پرکام نسبتاً شدید است و حاوی پتاسیک، فیلیک، آرژیلیک و پروپلیتیک هست.

۳- مواد و روش‌ها

نمونه‌برداری در محدوده سیستم پورفیری پرکام در یک شبکه منظم صورت گرفته است. شبکه نمونه‌برداری به صورت مربعی با فاصله ۱۰۰ متر هست. تعداد نمونه‌ها ۳۷۷ عدد بوده، آنالیز توسط دستگاه *ICP-MS* انجام شده و مقدار عیار ۴۵ عنصر در طی این فرآیند گزارش شده است.

۳-۱- روش جدایش مقادیر آنومال بر اساس فواصل ماهالانوبیس

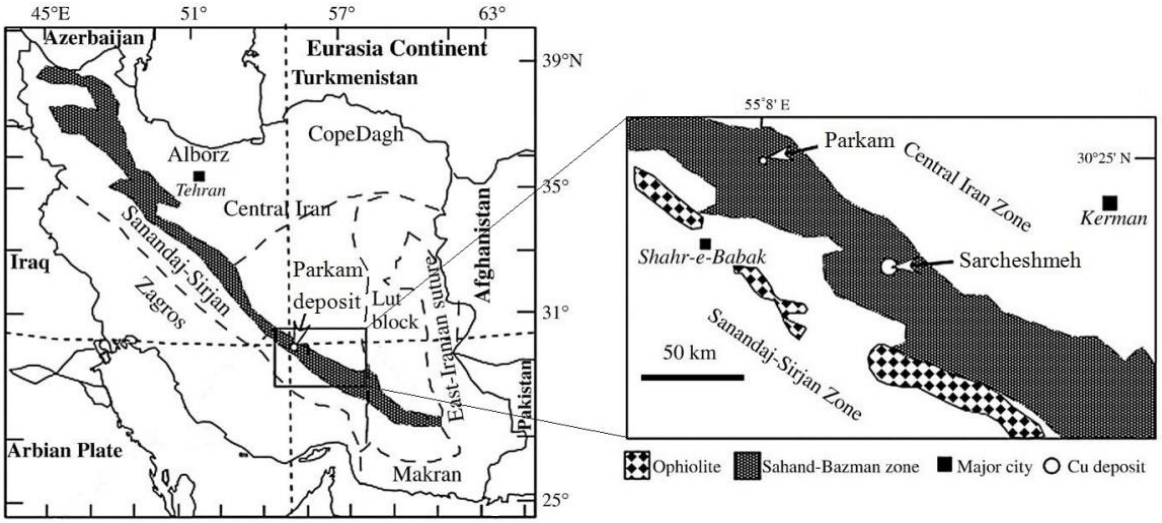
این روش یک روش چند متغیره است که در آن بر اساس فواصل ماهالانوبیس، آنومالی‌ها مشخص می‌شوند. در فضای n بعدی، فاصله ماهالانوبیس (D^2) مشابه مقادیر استاندارد شده تابع Z برای یک متغیر هست. در واقع فرمول محاسبه فواصل ماهالانوبیس شبیه استانداردکردن داده‌های

تک متغیره است. رابطه محاسبه فاصله های ماهالانوبیس هر نمونه به صورت زیر بیان می‌شود [۱۷]:

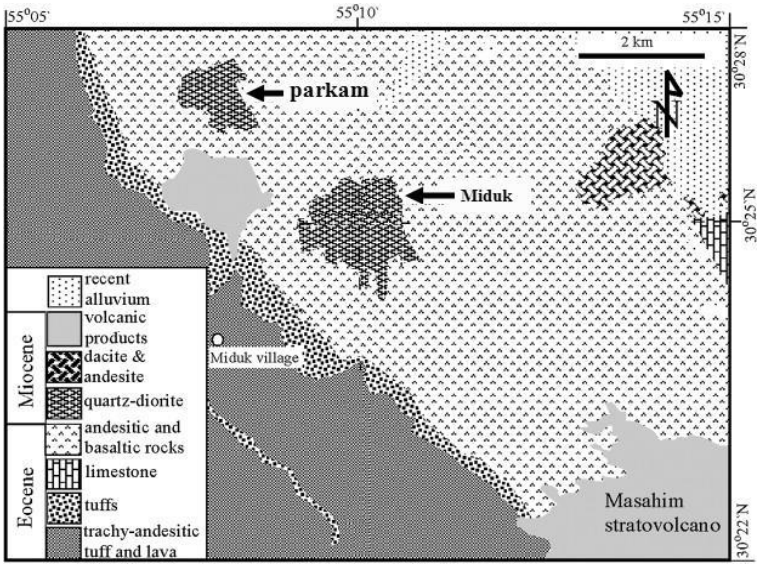
$$D^2 = ([x]-[\bar{x}])[S]^{-1}([x]-[\bar{x}])^T \quad (۱)$$

که در آن $[x]$ بردار داده‌ها برای نمونه مورد نظر، $[\bar{x}]$ بردار میانگین کل داده‌ها و $[S]^{-1}$ ماتریس کواریانس است. همانگونه که مشاهده می‌شود فاصله مورد نظر معادل کسر بردار میانگین داده‌ها از بردار هر یک از داده‌ها تقسیم بر ماتریس $[S]^{-1}$ هست.

توزیع فواصل ماهالانوبیس برای یک جامعه نرمال چند متغیره از یک توزیع کای - اسکور با درجه آزادی معادل تعداد متغیرها پیروی می‌کند. برای تشخیص مقادیر آنومال لازم است نمودار فواصل ماهالانوبیس بر حسب مقادیر χ^2 آنها رسم شود. در ادامه به پردازش اولیه داده‌ها و سپس به جدایش مقادیر آنومال از زمینه به کمک روش مذکور پرداخته می‌شود [۱۸].



شکل ۱: نقشه زون‌های لیتوتکتونیکی ایران که موقعیت سیستم پورفیری پرکام و کانسار سرچشمه در آن بزرگنمایی شده است. بخش هاشور خورده نوار سهند - بزمان را نشان می‌دهد [۱۵].



شکل ۲: قسمتی از نقشه زمین‌شناسی ناحیه‌ای شهربابک [۱۶].

۳-۲- K- نزدیک‌ترین همسایه

K- نزدیک‌ترین همسایه *(KNN)* یک الگوریتم یادگیری هست که در روش باز شناسی الگو طی چندین دهه مطالعه شده است [۱۹]. *KNN* به عنوان یکی از کاراترین روش‌ها شناخته شده است و مطالعات زیادی *KNN* را روی اسناد آزمایشی *Reuters* بکار برده‌اند، که این مطالعات پیشنهاد می‌کنند که *KNN* با *(Support Vector Machine)*

SVM، از روش‌هایی مانند تقریب خطی کوچک‌ترین مربعات، *Naïve Bayes* و شبکه‌های عصبی بهتر عمل می‌نماید ([۱۹]، [۲۰]). این روش هنوز یک روش کارا و ساده برای دسته‌بندی متن هست. ایده *KNN* طبق زیر است:

یک داده آموزشی برای دسته‌بندی وجود دارد، الگوریتم Kهمسایه نزدیک در میان داده‌های آموزشی پیش دسته‌بندی شده، بر اساس یک معیار شباهت پیدا کرده و دسته‌های این K همسایه نزدیک برای پیش‌بینی دسته داده آزمایشی به وسیله امتیاز دهی داده‌های هر دسته منتخب، استفاده می‌شود. اگر بیشتر از یک همسایه به دسته‌های مشابه تعلق داشته باشد، مجموع امتیاز آن‌ها به عنوان وزن آن دسته استفاده می‌شود و دسته با بالاترین امتیاز به داده مورد آزمایش انتساب می‌یابد، که اگر از یک مقدار آستانه تجاوز کند، بیشتر از یک دسته می‌تواند به سند آزمایشی انتساب یابد. یک مشکل در روش *KNN*، تعیین مقدار K هست و برای تعیین آن باید یک سری آزمایش‌ها با مقادیر مختلف K انجام شود، تا بهترین مقدار برای K را تعیین کند. عیب دیگر *KNN* پیچیدگی زمانی محاسباتی مورد نیاز برای پیمایش همه داده‌های آموزشی هست [۲۱]. طبقه‌بند *KNN* به دلیل قابلیت درک بالا و عدم نیاز به ایجاد فرضیه روی داده‌ها، روشی ساده و پرکاربرد محسوب می‌شود [۲۲]. در مطالعه پیش رو با در نظر گرفتن مقادیر مختلف K و محاسبه خطاهای مورد نظر و استفاده از برنامه‌نویسی الگوریتم آن در نرم‌افزار *MATLAB* سعی می‌شود دو مشکل فوق‌الذکر تا حد مطلوبی برطرف گردد.

۳-۳- طبقه‌بندی ساده بیزی

یک روش بسیار کاربردی یادگیری بیز روش یادگیرنده ساده بیزی هست که عموماً روش طبقه‌بندی ساده بیز نامیده می‌شود. در برخی زمینه‌ها نشان داده شده است که کارایی آن قابل قیاس با کارایی روش‌هایی مانند شبکه عصبی و درخت تصمیم هست. طبقه‌بندی ساده بیزی برای مسائلی که هر نمونه x در آن توسط مجموعه‌ای از مقادیر صفات و تابع هدف $f(x)$ از مجموعه‌ای مانند v انتخاب می‌گردد، کاربرد دارد. روش بیزی برای طبقه‌بندی نمونه جدید این است که محتمل‌ترین طبقه یا مقدار هدف v_{MAP} را با داشتن مقادیر صفات $\{a_1,a_2,...,a_n\}$ که توصیف کننده نمونه جدید است شناسایی کند،

$$v_{MAP} = \arg\max P(v_j \mid a_1,a_2,...,a_n), v_j \in V \qquad (۲)$$

با استفاده از قضیه بیز می‌توان عبارت بالا را به صورت زیر بازنویسی کرد،

$$v_{MAP} = \arg\max \frac{P(a_1,a_2,...,a_n \mid v_j)}{P(a_1,a_2,...,a_n)}, v_j \in V$$
$$v_{MAP} = \arg\max P(a_1,a_2,...,a_n \mid v_j)P(v_j), \quad v_j \in V \qquad (۳)$$

حال با استفاده از داده‌های آموزشی سعی می‌کنیم دو جمله معادله بالا را تخمین بزنیم . محاسبه از روی داده‌های آموزشی به این صورت که میزان تکرار v_j در داده‌ها چقدر است، آسان هست. اما محاسبه جملات مختلف $P(a_1,a_2,...,a_n \mid v_j)$ به این صورت قابل قبول نخواهد بود مگر اینکه حجم بسیار زیادی از داده‌های آموزشی در اختیار داشته باشیم. مشکل اینجاست که تعداد این جملات برابر تعداد نمونه‌های ممکن ضرب در تعداد مقادیر تابع هدف هست. بنابراین باید هر نمونه را چندین بار مشاهده کنیم تا تخمین مناسبی از آن به دست آید.

فرض روش طبقه‌بندی ساده بیز بر اساس این ساده سازی است که مقادیر صفات با داشتن مقادیر تابع هدف از یکدیگر مستقل شرطی باشند. به عبارت دیگر، این فرض بیانگر این است که به شرط مشاهده خروجی تابع هدف، احتمال مشاهده صفات $a_1,a_2,...,a_n$ برابر ضرب احتمالات هر صفت به طور جداگانه باشد. اگر این را جایگزین معادله بالا کنیم روش طبقه‌بندی ساده بیز را نتیجه می‌دهد،

$$v_{NB} = \arg\max P(v_j)\prod_i P(a_i \mid v_j), v_j \in V \qquad (۴)$$

که v_{NB} خروجی طبقه‌بندی ساده بیز برای تابع هدف هست. توجه کنید که تعداد جملات $P(a_i \mid v_j)$ که در این روش باید محاسبه شوند برابر تعداد صفات ضرب در تعداد دسته‌های خروجی برای تابع هدف هست که این مقدار از تعداد جملات $P(a_1,a_2,...,a_n \mid v_j)$ بسیار کمتر است.

نتیجه اینکه یادگیری ساده بیزی سعی در تخمین مقادیر مختلف $P(v_j)$ و $P(a_i \mid v_j)$ با استفاده از میزان تکرار آنها در داده‌های آموزشی دارد. این مجموعه تخمین‌ها متناظر با فرض یاد گرفته شده است.

۳-۴- روش درخت تصمیم‌گیری

درخت تصمیم، روشی معروف برای دسته‌بندی است که نتایج آن در یک فلوچارت، شبیه ساختار درخت ارائه شده است که هر گره نشانگر یک تست بر روی ارزش مشخصه و هر شاخه، خروجی هر تست را نمایش می‌دهد، برگ‌های درخت نیز نمایانگر کلاس‌هاست. به طور عادی، پیچیدگی یک درخت تصمیم با افزایش تعداد مشخصه‌ها افزایش می یابد. اگرچه در بعضی از شرایط دیده شده است که تنها تعداد کمی از مشخصه‌ها می‌توانند کلاسی را که هر شیء به آن تعلق دارد، تعیین کنند و بقیه مشخصه‌ها کم یا بی‌تأثیرند [۲۳].

در ساخت درخت‌های تصمیم معمولاً داده‌ها را به دو دسته تقسیم می‌کنند:

- داده‌های آموزشی که برای ساخت مدل مورد استفاده قرار می‌گیرند؛
- داده‌های تست که برای تست و ارزیابی مدل ساخته شده کاربرد دارند.

کیفیت داده‌های آموزشی اغلب نقش مهمی در تعیین کیفیت درخت تصمیم دارد. در صورتی که آموزش سیستم زیاد شود یعنی داده‌هایی که برای آموزش و ساخت مدل به کار می‌رود درصد زیادی از داده‌ها باشد، دچار حالتی به نام آموزش بیش از حد مدل خواهیم شد که به دلیل وجود موارد غیرعادی در داده‌های آموزشی خطا تولید می‌کند [۲۴].

۴- پردازش داده‌ها و نتایج

در این قسمت ابتدا نتایج حاصل از آنالیزهای شیمیایی را آماده پردازش نموده و سپس اقدام به جدایش مقادیر آنومال مس و مولیبدن به کمک روش فواصل ماهالانوبیس می‌شود. در نهایت نیز به کمک روش‌های داده‌کاوی، به پیش‌بینی نمونه‌های آنومال در منطقه مورد مطالعه، پرداخته می‌شود.

با استفاده از برنامه‌های پیشنهادی در مطالعات [۱۰ و ۱۱] به تعیین نوع جوامع، نرمال‌سازی و در نهایت محاسبه پارامترهای اولیه آماری پرداخته شده است. پس از استفاده از برنامه های مذکور مشخص گردید که مس و مولیبدن دارای توزیع لاگ نرمال بوده و مشخصات آماری آنها در جدول ۲ آورده شده است (محاسبات در سطح اعتماد ۹۵ درصد انجام شده است). جدول ۱ نیز مشخصات عناصر قبل از لگاریتم‌گیری را نشان می‌دهد.

جدول ۱: مشخصات عناصر قبل از لگاریتم‌گیری.

عنصر	میانگین عیار (ppm)	واریانس عیار (ppm)^2	بیش‌ترین عیار (ppm)	شاخص W
مس	۲۸۹	۶۵۲۷۴۰	۹۴۷۰	۸۳/۸۴
مولیبدن	۵	۷۲/۴	۸۴/۸	۳۸/۲۲

جدول ۲: مشخصات عناصر در توزیع لگاریتمی.

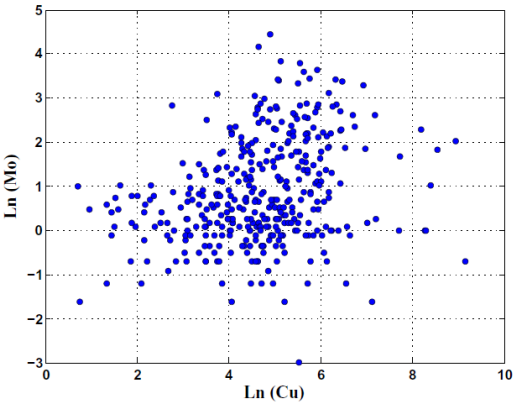
عنصر	نوع توزیع لگاریتمی	ثابت افزودنی	میانگین لگاریتمی	میانگین واقعی	واریانس لگاریتمی	واریانس واقعی	شاخص W
مس	دو متغیره	۰	۴/۶۷	۲۶۷	۱/۸	۳۷۵۴۷۰	۱/۰۲
مولیبدن	دو متغیره	۰	۰/۸۷	۴/۸	۱/۴	۶۷/۹	۰/۸۲

۴-۱- جدایش مقادیر آنومال بر اساس فواصل ماهالانوبیس

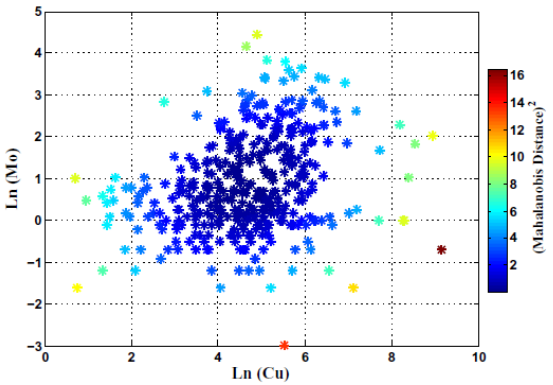
پس از پردازش اولیه مقادیر مس و مولیبدن، در این قسمت اقدام به جدایش مقادیر آنومال می‌شود. در شکل ۳ نمایی از پراکندگی ۳۷۷ نمونه مورد بررسی را بر اساس دو متغیر (متغیر اول عیار مس و متغیر دوم عیار مولیبدن) در مقیاس لگاریتمی مشاهده می‌کنیم.

سپس داده‌های لگاریتم گرفته شده به عنوان ورودی در اختیار برنامه رایانه‌ای که به کمک نرم‌افزار *MATLAB* نوشته شده است، قرار گرفتند. در برنامه مذکور مختصات نقاط نمونه‌برداری و عیار نمونه‌های مورد نظر به صورت ماتریس‌های تک ستونی به برنامه معرفی شده و سپس برنامه مذکور عیار عناصر مورد نظر را در ماتریسی چند ستونی آماده کرده و به کمک رابطه ۱ مقدار D^2 را برای هر نمونه به صورت ماتریسی تک ستونی محاسبه می‌کند. با توجه به تعداد زیاد نمونه‌ها و همچنین محدودیت در تعداد صفحات مقاله، مقادیر D^2 به شکل جدول گزارش نشده و به جای آن

از نموداری جهت نمایش پراکندگی و مقادیر D^2 برای هر نمونه استفاده شده است که در شکل ۴ قابل مشاهده است.



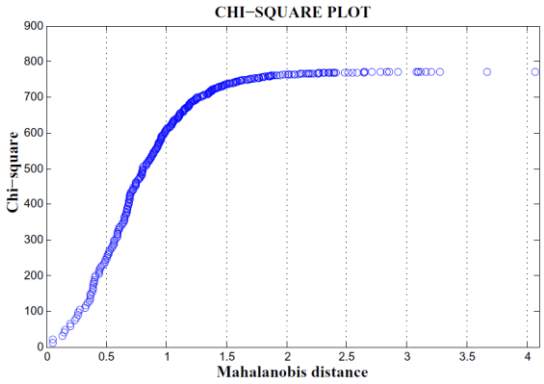
شکل ۳: ترسیمی از پراکندگی مس و مولیبدن در مقیاس لگاریتمی



شکل ۴: نمایش مقادیر فواصل ماهالانوبیس برای نمونه‌ها

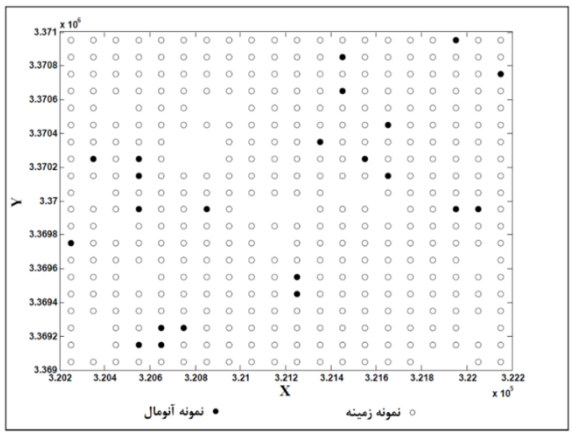
پس از محاسبه ماتریس مقادیر ماهالانوبیس، اقدام به محاسبه ماتریس مقادیر احتمال تجمعی نمونه‌ها با توجه به ترتیب صعودی فواصل ماهالانوبیس گردید. حال با در دست داشتن ماتریس مقادیر احتمال تجمعی برای نمونه‌ها و همچنین دو متغیره بودن نوع محاسبات، ماتریس مقادیر χ^2 نمونه‌های مورد نظر تعیین گردید. سپس با در دست داشتن دو ماتریس از مقادیر χ^2 و مقادیر D^2 اقدام به رسم نمودار کای - اسکور شد که نتیجه حاصل از آن را می‌توان در شکل ۵ مشاهده نمود. شایان ذکر است تمام عملیات محاسباتی به کمک نرم‌افزار رایانه‌ای نوشته شده توسط نویسندگان در نرم‌افزار *MATLAB* صورت پذیرفته است.

همان‌گونه که در شکل نیز مشاهده می‌شود، در فاصله ۲/۴ یک انفصال ضعیف و در فاصله ۲/۹ یک جدایش قوی‌تر مشاهده می‌شود. لذا می‌توان نمونه‌هایی با فاصله ماهالانوبیس فراتر از ۲/۴ را به عنوان نمونه‌های آنومال ممکن و نمونه‌هایی با فاصله ماهالانوبیس فراتر از ۲/۹ را به عنوان نمونه‌های آنومال احتمالی معرفی کرد.



شکل ۵: نمودار کای - اسکور مس و مولیبدن محدوده پرکام

نمونه‌های آنومال تعیین شده از نظر دو متغیر مس و مولیبدن توسط روش فوق‌الذکر در شکل ۶ به نمایش در آورده شده است. با توجه به محدودیت تعداد صفحات مقاله، تنها مشخصات نمونه‌های آنومال در جدول ۳ گزارش شده است.



شکل ۶: نمایش نمونه‌های آنومال جدا شده به کمک روش فواصل ماهالانوبیس

حال جهت اعمال روش‌های داده‌کاوی بر روی نتایج حاصله، مقدار ۱ به عنوان وزن نمونه‌های زمینه و مقدار ۲ به عنوان وزن نمونه‌های آنومال برای هر نمونه در نظر گرفته و برنامه ترکیب روش‌های مختلف به گونه‌ای طراحی می‌شود که علاوه بر دریافت مقادیر طول و عرض جغرافیایی و عیار دو عنصر مس و مولیبدن به عنوان ورودی، مقادیر وزن هر نمونه را نیز به منظور فرآیند پیش‌بینی دریافت و سپس الگوریتم روش مذکور را اجرا نماید.

جدول ۳: مشخصات نمونه‌های آنومال تعیین شده توسط روش فواصل ماهالانوبیس.

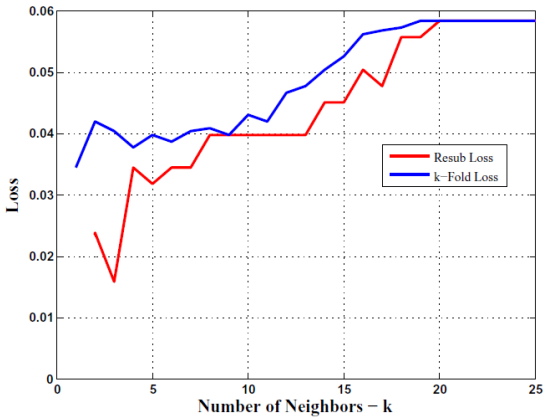
طول جغرافیایی	عرض جغرافیایی	فاصله ماهالانوبیس	احتمال تجمعی	کای – اسکور
۳۲۰۵۵۰	۳۳۶۹۱۵۰	۲/۷۸۲۳۳۸	۰/۰۵۴۳۷۷	۱۵۹/۹۵۹۲
۳۲۰۶۵۰	۳۳۶۹۱۵۰	۲/۵۱۰۰۱۹	۰/۰۵۷۰۲۹	۱۶۵/۷۵۵۴
۳۲۰۶۵۰	۳۳۶۹۲۵۰	۳/۰۹۴۴۷۲	۰/۱۰۷۴۲۷	۲۶۲/۱۳۳۸
۳۲۰۷۵۰	۳۳۶۹۲۵۰	۲/۴۰۹۱۴	۰/۱۱۰۰۸	۲۶۶/۶۱۵۵
۳۲۱۲۵۰	۳۳۶۹۴۵۰	۳/۶۶۹۳۳۳	۰/۲۲۶۷۹	۴۳۱/۶۸۵۳
۳۲۱۲۵۰	۳۳۶۹۵۵۰	۲/۵۵۰۷۵۸	۰/۲۷۷۱۸۸	۴۸۴/۵۷۳
۳۲۰۲۵۰	۳۳۶۹۷۵۰	۲/۶۴۳۲۲۶	۰/۳۵۴۱۱۱	۵۵۴/۱۸۱۵
۳۲۰۵۵۰	۳۳۶۹۹۵۰	۲/۹۲۷۲۶۳	۰/۴۶۲۸۶۵	۶۳۱/۵۸۲۳
۳۲۰۸۵۰	۳۳۶۹۹۵۰	۲/۴۷۹۵۶۲	۰/۴۷۰۸۲۲	۶۳۶/۲۷۹۴
۳۲۱۹۵۰	۳۳۶۹۹۵۰	۲/۷۱۰۵۲۳	۰/۴۸۹۳۹	۶۴۶/۷۲۶۹
۳۲۲۰۵۰	۳۳۶۹۹۵۰	۳/۲۰۵۴۹	۰/۴۹۲۰۴۲	۶۴۸/۱۶۰۸
۳۲۰۵۵۰	۳۳۷۰۱۵۰	۴/۰۶۴۸۶۳	۰/۵۵۰۳۹۸	۶۷۷/۱۴۱۷
۳۲۱۶۵۰	۳۳۷۰۱۵۰	۲/۸۵۶۰۹۷	۰/۵۷۹۵۷۶	۶۹۰/۰۵۳۸
۳۲۰۳۵۰	۳۳۷۰۲۵۰	۲/۶۵۰۵۹	۰/۵۹۸۱۴۳	۶۹۷/۷۲۵۷
۳۲۰۵۵۰	۳۳۷۰۲۵۰	۳/۰۹۶۷۴۹	۰/۶۰۳۴۴۸	۶۹۹/۸۳۹۹
۳۲۱۵۵۰	۳۳۷۰۲۵۰	۳/۱۶۱۲۴۴	۰/۶۲۴۶۶۸	۷۰۷/۹۵۰۷
۳۲۱۳۵۰	۳۳۷۰۳۵۰	۲/۵۸۰۷۰۲	۰/۶۶۷۱۰۹	۷۲۲/۵۱۱۹
۳۲۱۶۵۰	۳۳۷۰۴۵۰	۳/۱۱۸۲۶۶	۰/۷۲۸۱۱۷	۷۳۹/۵۶۵۵
۳۲۱۴۵۰	۳۳۷۰۶۵۰	۳/۲۷۱۶۲۹	۰/۸۲۰۹۵۵	۷۵۷/۵۶۰۳
۳۲۲۱۵۰	۳۳۷۰۷۵۰	۲/۸۳۱۵۸۵	۰/۸۹۲۵۷۳	۷۶۶/۰۰۳۲
۳۲۱۴۵۰	۳۳۷۰۸۵۰	۲/۶۴۵۱۶	۰/۹۲۷۰۵۶	۷۶۸/۴۳۵۲
۳۲۱۹۵۰	۳۳۷۰۹۵۰	۳/۰۷۹۷۹۴	۰/۹۹۳۳۶۹	۷۷۰/۴۱۸۵

۴-۲- پیش‌بینی مقادیر آنومال با استفاده از الگوریتم *KNN*

به منظور استفاده از روش *KNN* بر روی داده‌ها و همچنین برای تعیین بهترین مقدار K، الگوریتم روش *KNN* به شکل برنامه ای در نرم‌افزار متلب نوشته شده است که برای مقادیر K از مقدار ۱ تا۲۰، برنامه به تعداد ۲۰ بار اجرا می‌شود و در هر بار از اجرای آن، مقدار خطا اندازه‌گیری می‌شود. همچنین در مورد هر مقدار K برنامه ۵ بار الگوریتم را اجرا نموده و مقدار میانگین خطا در هر پنج مرحله را به عنوان خطا به K ی مورد نظر اختصاص می‌دهد. قابل ذکر است که برای محاسبه خطا از دو معیار ارزیابی استفاده شده است: الف) میزان خطای *Resubstitution* و ب) میزان خطا بر اساس اعتبار سنجی متقاطع (در ابتدا مدل *Cross-Validation* شبکه مورد نظر ساخته شده و سپس با استفاده از روش *K-Fold* میزان خطا محاسبه می‌شود).

با اجرای برنامه مذکور نمودار خطا نسبت به تعداد K تهیه شده و در شکل ۷ قابل مشاهده است. با توجه به شکل مشاهده می‌شود که اولاً خطای *Resubstitution* از خطای *K-Fold* کمتر است و ثانیاً بهترین مقدار برای K=۳ در نظر گرفته شده است (در مورد خطای *Resubstitution*). زیرا کمترین خطا در کلاس‌بندی با مقدار ۳ رخ داده است. لذا با مقدار K=۳ شبکه مورد نظر ساخته شده است. شبکه مذکور به گونه‌ای طراحی شده است که با توجه به چهار مقدار طول و عرض جغرافیایی و عیار دو عنصر مس و مولیبدن برای هر نمونه و همچنین وزنی که از روش فاصله ماهالانوبیس به آنها اختصاص داده شده، فرآیند پیش‌بینی مقادیر آنومال را بر روی مقادیر جدید ورودی انجام می‌دهد.

در واقع عملکرد این برنامه به گونه‌ای است که در ابتدا موارد زیر را به عنوان ورودی دریافت می‌نماید:



شکل ۷: نمودار مقدار خطای میانگین حاصل از ۵ بار اجرای الگوریتم *KNN* برای مقادیر مختلف *K*

الف- مجموعه داده‌های یادگیری: ماتریسی پنج ستونی از مقادیر عیار مس و مولیبدن نمونه‌ها به همراه طول و عرض جغرافیایی و وزن ارائه شده به آنها.

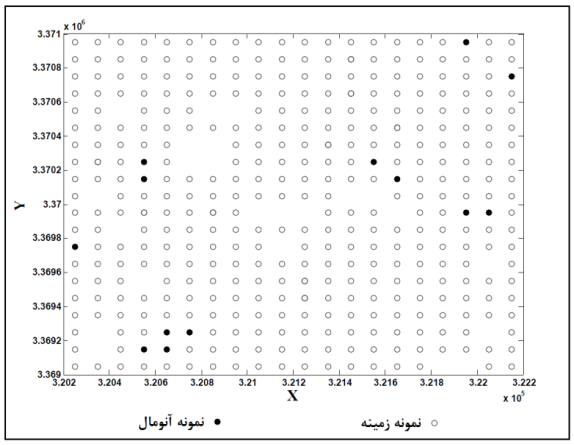
لازم به ذکر است که مجموعه داده‌های یادگیری خود به سه دسته داده‌های آموزش، اعتبارسنجی و آزمایش تقسیم‌بندی می‌شوند. معمولاً در تقسیم‌بندی مجموعه داده‌های مسئله، جهت دست‌یابی به پاسخی قابل‌قبول، باید تعداد داده‌های مورد استفاده در مجموعه مربوط به داده‌های فاز آموزش، بیشتر باشد. همچنین باید تا حد امکان تقسیم‌بندی داده‌ها به صورت تصادفی باشد. در این راستا عقیده اکثر متخصصان بر این است که ۷۰ درصد داده‌ها برای آموزش شبکه، ۱۵ درصد برای اعتبارسنجی و ۱۵ درصد دیگر برای تست و آزمون شبکه مورد استفاده قرار بگیرد. در مطالعه حاضر نیز نحوه تقسیم‌بندی و درصد سهم آنها مطابق با تقسیم‌بندی فوق‌الذکر صورت گرفته است.

ب- مجموعه داده‌های مورد بررسی: ماتریسی چهار ستونی (طول و عرض جغرافیایی، عیار مس و مولیبدن) از نمونه‌های جدید که قرار است فرآیند پیش‌بینی بر روی آنها صورت پذیرد.

سپس شبکه مذکور را طراحی و با توجه به آن در مورد نوع نمونه‌های مورد بررسی تصمیم گرفته و در نهایت نمونه‌های آنومال را به عنوان خروجی در اختیار کاربر قرار می‌دهد.

خطای *Resubstitution* و *K-Fold* برای شبکه طراحی شده با توجه به تعداد نمونه‌های مورد بررسی در محدوده مطالعاتی پرکام، به ترتیب برابر با ۰/۰۲۳۹ و ۰/۰۳۷۱ گزارش شده‌اند که مقدار بسیار قابل‌قبولی هستند و نشان می دهند که ترکیب این دو روش پیشگویی در مورد مقادیر آنومال را با دقت بسیار بالا انجام می‌دهد.

در نهایت نیز به منظور نمایش عملکرد ترکیب این دو روش، مجدداً از داده‌های محدوده پرکام (عیار مس و مولیبدن به همراه مختصات جغرافیایی همان ۳۷۷ نمونه) به عنوان داده‌هایی جدید برای شبکه مذکور استفاده شده که نتایج آن در شکل ۸ قابل مشاهده است.



شکل ۸: نمایش نمونه‌های آنومال جدا شده به کمک ترکیب دو روش *KNN* و فواصل ماهالانوبیس

۴-۳- پیش‌بینی مقادیر آنومال با استفاده از الگوریتم درخت تصمیم‌گیری

به منظور اعمال روش درخت تصمیم‌گیری بر روی داده‌ها، الگوریتم روش مذکور نیز به شکل برنامه‌ای در نرم‌افزار متلب نوشته شده است. قابل ذکر است که همانند قسمت ۴-۲ برای محاسبه خطا، از دو معیار ارزیابی استفاده شده است.

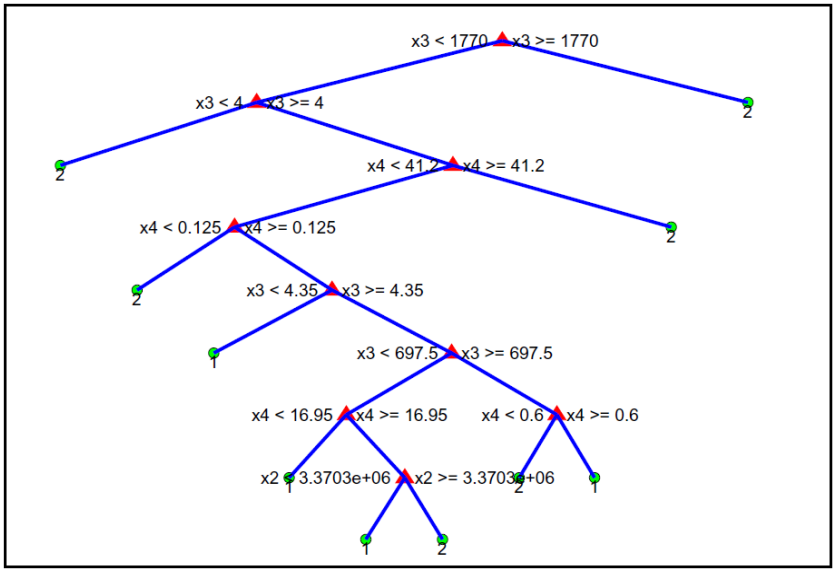
برنامه مورد نظر به گونه‌ای طراحی شده است که با توجه به چهار مقدار طول و عرض جغرافیایی و عیار دو عنصر مس و مولیبدن برای هر نمونه و همچنین وزنی که از روش فاصله ماهالانوبیس به آنها اختصاص داده شده، فرآیند پیش‌بینی مقادیر آنومال را بر روی مقادیر جدید ورودی انجام می‌دهد.

در واقع عملکرد این برنامه نیز به گونه‌ای است که در ابتدا موارد زیر را به عنوان ورودی دریافت می‌نماید:

الف- مجموعه داده‌های یادگیری: ماتریسی پنج ستونی از مقادیر عیار مس و مولیبدن به همراه طول و عرض جغرافیایی نمونه‌ها و وزن ارائه داده به آنها توسط روش ماهالانوبیس.

ب- مجموعه داده‌های مورد بررسی: ماتریسی چهار ستونی (طول و عرض جغرافیایی، عیار مس و مولیبدن) از نمونه‌های جدید که قرار است فرآیند پیش‌بینی بر روی آنها صورت پذیرد.

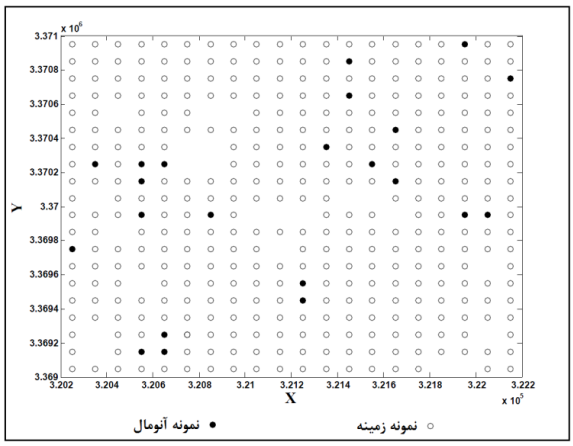
سپس درخت مورد نظر را طراحی و با توجه به آن در مورد نوع نمونه‌های مورد بررسی تصمیم گرفته و در نهایت نمونه‌های آنومال را به عنوان خروجی در اختیار کاربر قرار می دهد. نتیجه حاصل از اجرای برنامه مذکور، طراحی درختی (شکل ۹) برای ارزیابی نمونه‌های نامشخص از نظر آنومال بودن یا نبودن تحت تأثیر چهار پارامتر طول و عرض جغرافیایی و عیار عناصر مس و مولیبدن هست.



شکل ۹: درخت تصمیم‌گیری در مورد پیش‌بینی مقادیر آنومال

همچنین مشاهده گردید که خطای *Resubstitution* و *K-Fold* برای شبکه طراحی شده با توجه به تعداد نمونه‌های مورد بررسی در محدوده مطالعاتی پرکام، به ترتیب برابر با ۰/۰۰۵۳ و ۰/۰۲۹۲ گزارش شده‌اند که مقدار بسیار قابل‌قبولی می باشند و نشان می‌دهند که ترکیب این دو روش، دقت پیشگویی در مورد مقادیر آنومال را به مراتب بالا می‌برد.

سپس به منظور نمایش عملکرد و کارایی ترکیب فوق‌الذکر نیز مجدداً داده‌های مربوط به ۳۷۷ نمونه محدوده پرکام در اختیار شبکه طراحی شده قرار گرفته و نمونه‌های آنومال تعیین شده توسط آن به عنوان خروجی، در شکل ۱۰ به نمایش در آورده شده است.



شکل ۱۰: نمایش نمونه‌های آنومال جدا شده به کمک ترکیب دو روش درخت تصمیم‌گیری و فواصل ماهالانوبیس.

۴-۴- پیش‌بینی مقادیر آنومال با استفاده از الگوریتم طبقه‌بندی ساده بیز

در این قسمت به منظور استفاده از روش طبقه‌بندی ساده بیزی برای داده‌ها، الگوریتم این روش نیز به شکل برنامه‌ای در نرم‌افزار متلب نوشته شده است که تمام روابط موجود به منظور طراحی شبکه را شامل می‌شود.

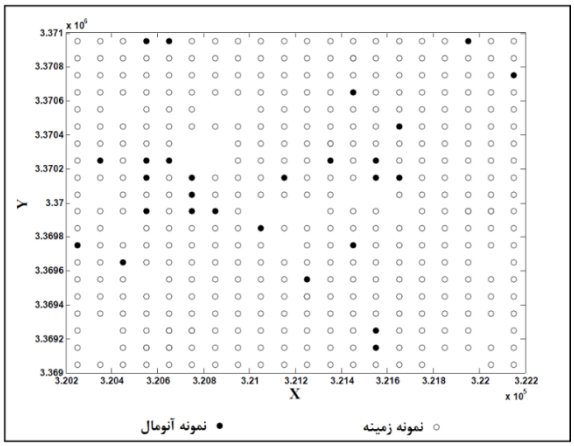
در روش طبقه‌بندی ساده بیزی همانند دو روش قبل، برنامه مورد نظر به گونه‌ای طراحی شده است که چهار مقدار طول و عرض جغرافیایی و عیار دو عنصر مس و مولیبدن برای هر نمونه را به همراه وزنی که از روش فواصل ماهالانوبیس به آنها اختصاص داده شده، به عنوان ورودی دریافت نموده و سپس فرآیند پیش‌بینی مقادیر آنومال را بر روی مقادیر جدید ورودی انجام می دهد. در حقیقت این برنامه همانند برنامه های قبل، دو مجموعه داده را دریافت

می‌نمایـد، یکی مجموعـه داده‌هـای یـادگیری و دیگری مجموعه داده‌های مورد بررسی. مجموعه داده‌های یادگیری، ماتریسی پنج ستونی از مقادیر عیار مس و مولیبدن، طول و عرض جغرافیایی و وزن در نظر گرفته شده برای آنها بوده و مجموعه داده‌های مورد بررسـی، ماتریسـی چهار سـتونی (طول و عرض جغرافیایی، عیار مس و مولیبدن) از نمونه‌های جدید است.

برنـامـه مورد نظر در ابتـدا بـه کمـک داده‌های آموزش شـبکه مورد نظر را طراحی نموده و پس از اعتبارسـنجی شـبکه طراحی شـده به کمک داده‌های اعتبارسـنجی، در نهایت عملکرد شـبکه مذکور را به کمک مجموعه داده‌های آزمایش مورد بررسی قرار می‌دهد.

نتیجه حاصـل از اجرای برنامه مذکور، طراحی شبکه ای برای ارزیـابی نمونـه‌های نامشـخص از نظر آنومال بودن یا نبودن تحـت تـأثیر چهار پارامتر طول و عرض جغرافیایی و عیار عناصـر مس و مولیبدن هسـت. خطای شـبکه طراحی شـده در مورد نمونـه‌هـای زمینـه و آنومال به ترتیب برابر ۰/۰۳۹۴ و ۰/۴۰۹۱ و خطـای کلی شـبکـه برابر ۰/۰۶۱ گزارش شـده اسـت که در مقایسـه با ترکیب‌های گذشـته عملکرد به مراتب ضعیف‌تری را نشان می‌دهد.

شـکل ۱۱ نتیجه حاصل از اجرای برنامه مذکور در مورد ۳۷۷ نمونـه محـدوده پرکـام را بـه عنوان خروجی شـبکه طراحی شده و عملکرد آن نمایش می‌دهد.



شکل ۱۱: نمایش نمونه‌های آنومال جدا شده به کمک ترکیب دو روش طبقه‌بندی ساده بیز و فواصل ماهالانوبیس.

۵- بحث

همـان‌طور که مشـاهده گردید در انتهای اجرای برنامه مربوط به هر روش، به منظور بررسـی عملکرد و کارایی هر روش و همچنین جهت تشـریح محاسـبه نوع خطای مورد نظر، مجدداً داده‌های حاصله از عملیات نمونه‌برداری سطحی پرکـام یعنی ۳۷۷ نمونـه مورد نظر، به عنوان ورودی جدید در اختیـار برنـامـه‌های تهیه شـده قرار گرفتند. با توجه به خروجی‌های حاصل شده، مشاهده می‌شود که روش درخت تصـمیم‌گیری تنهـا در مورد ۲ نمونـه از ۳۷۷ نمونـه مورد بررسـی (۲۲ عدد نمونه آنومال و ۳۵۵ عدد نمونه زمینه) به اشـتباه قضاوت کرده است. یعنی در یک مورد نمونه آنومال را بـه عنوان نمونـه زمینه و در موردی دیگر، نمونه زمینه را به عنوان نمونه آنومال معرفی کرده است. این در حالی است کـه روش *KNN* از ۳۷۷ نمونـه مورد بررسـی، در مورد ۹ نمونه از جامعه آنومال به اشـتباه قضـاوت کرده و ۹ نمونه آنومـال را بـه همراه ۳۵۵ عـدد نمونه دیگر زمینه به عنوان زمینه معرفی کرده اسـت. در نهایت نیز دیده می‌شـود که روش طبقـه‌بنـد سـاده بیز بر خلاف دو روش قبل عملکرد بسیار ضعیف‌تری را به نمایش گذاشته و از ۳۷۷ نمونه مورد بررسی، در مورد ۲۳ نمونه، اشتباه قضاوت کرده است.

با تقسـیم تعداد نمونه‌های به اشـتباه پیش‌بینی شده بر تعداد کل داده‌ها، خطای *Resubstitution* طبق تعریف آن به‌سادگی و به شرح زیر قابل محاسبه هست:

- برای ترکیب روش درخت تصـمیم‌گیری با روش ماهالانوبیس: ۲ ÷ ۳۷۷ ≈ ۰/۰۰۵۳
- بـــرای ترکیـــب روش *KNN* بـــا روش ماهالانوبیس: ۹ ÷ ۳۷۷ ≈ ۰/۰۲۳۹
- برای ترکیـب روش طبقه‌بند سـاده بیز با روش ماهالانوبیس: ۲۳ ÷ ۳۷۷ ≈ ۰/۰۶۱

همان‌طور که مشـاهده می‌شود، خطای محاسبه شده در بالا دقیقاً برابر با مقدار خطای گزارش شده توسط برنامه‌های پیشنهادی (دستور *ResubLoss*) در قسمت‌های ۴-۲، ۴-۳ و ۴-۴ است.

همچنین قابل ذکر اسـت که محدوده‌های آنومال تعیین شـده توسـط درخت تصـمیم‌گیری (شـکل ۱۰)، به واسطه خطای محاسـبه شـده برای شـبکه‌های آن، با محدوده‌های آنومالی تعیین شده در شکل ۶، مطابقت بسیار قابل‌قبولی را بـه همراه دارد. در حـالی کـه این موضـوع در مورد نتـایج حاصله از روش‌های *KNN* و بیز، صادق نبوده و محدوده‌های آنومـالی تعیین شـده در این روش‌ها با شـکل ۶ مطابقت مطلوبی را به همراه نداشته‌اند.

با مقایسه عملکرد این سه ترکیب، مشاهده می‌شود که ترکیب دو روش درخت تصمیم‌گیری و روش فواصل ماهالانوبیس، ترکیبی به مراتب قوی‌تر نسبت به دو ترکیب دیگر هست و می‌توان ترکیب دو روش مذکور را روشی بسیار مناسب در پیش‌بینی مقادیر آنومال در مبحث جدایش مقادیر آنومالی از زمینه معرفی کرد. همچنین قابل ذکر است که مزیت دیگر ترکیب این روش‌ها، تعیین مقادیر آنومال بر اساس چند متغیر (چند عنصر) به دلیل انتخاب روش چند متغیره فواصل ماهالانوبیس هست.

۶- نتیجه گیری

به منظور جدایش مقادیر آنومال از زمینه از نظر دو عنصر مس و مولیبدن، از روش چند متغیره فواصل ماهالانوبیس استفاده شد که مزیت بارز آن، تعیین مقادیر آنومال بر اساس چند متغیر است. به این ترتیب با رسم نمودار کای - اسکور برای دو عنصر مس و مولیبدن محدوده پرکام، نمونه‌های آنومال مشخص شدند. سپس نتایج حاصل از آن به همراه ۴ پارامتر عیار مس، مولیبدن، طول و عرض جغرافیایی نمونه‌ها، به منظور قضاوت در مورد نمونه‌های دیگر، در اختیار روش‌های داده‌کاوی مورد مطالعه قرار گرفت. به این ترتیب با مقایسه میزان خطای شبکه‌های طراحی شده، مشاهده گردید که روش درخت تصمیم‌گیری و روش *KNN* در مقایسه با روش دیگر در ترکیب با روش جدایش فواصل ماهالانوبیس، از دقت به مراتب بالاتری برخوردار هستند. خطای *Resubstitution* محاسبه شده برای درخت تصمیم‌گیری و *KNN* به ترتیب برابر با ۰/۰۰۵۳ و ۰/۰۲۳۹ و برای روش طبقه‌بند ساده بیز برابر با ۰/۰۶۱ گزارش شده است. بنابراین در بین دو روش مذکور نیز روش درخت تصمیم‌گیری برتری نسبی را به خود اختصاص داده است و می‌توان ترکیب دو روش جدایش و داده‌کاوی فوق‌الذکر را روشی مؤثر در پیش‌بینی مقادیر آنومال در نظر گرفت، زیرا مقادیر آنومال از نظر دو عنصر تائید شده‌اند و خطای روش پیش‌بینی نیز بسیار کوچک است. در نهایت نیز نمونه‌های آموزشی به عنوان نمونه‌های تست در اختیار هر یک از شبکه‌ها قرار گرفت و نشان داده شد که شبکه برتر طراحی شده (توسط روش درخت تصمیم‌گیری) تنها دو نمونه آنومال از بین ۳۷۷ نمونه را اشتباهاً به عنوان زمینه معرفی کرده است. این تعداد برای شبکه‌های طراحی شده توسط دو روش *KNN* و بیز به ترتیب برابر با ۹ و ۲۳ گزارش شده است. بنابراین با نوشتن برنامه رایانه‌ای مختص به روش ترکیبی معرفی شده با استفاده از نرم‌افزار *MATLAB*، به خوبی این ترکیب قابل استفاده برای کارهای مشابه نیز هست.

مراجع

[1] Ghannadpour, S. S. and A. Hezarkhani (2012). "Lead Geochemical Behavior with respect to those of Zinc and Iron based on Clustering Method Applications in Parkam Porphyry Copper System, Shahr Babak, Kerman." Journal of Researches in Earth Sciences 3(9): 64-77 (In Persian).

[2] Cheng, Q. (1999). Spatial and scaling modelling for geochemical anomaly separation. Journal of Geochemical Exploration, 65(3), 175-194.

[3] Ghannadpour, S. S., Hezarkhani, A., Maghsoudi, A., Farahbakhsh, E. (2015). Assessment of prospective areas for providing the geochemical anomaly maps of lead and zinc in Parkam district, Kerman, Iran. Geosciences Journal, 19(3), 431-440.

[4] Sinclair, A. J. (1991). A Fundamental Approach to Threshold Estimation in Exploration Geochemistry, probability plots revisited. journal of Geochemical Exploration, 41(1-2), 1–22.

[5] Mehrgini, B. and H. Memarian (2010). " Evaluation of Mahalanobis Distance method's performance in separating oil facies, in one of hydrocarbon fields in Iran." 14th Iranian geophysics conference, Iran's geopolitical forum.

[6] Zhao, X., Li, Y., Zhao, Q. (2015). Mahalanobis distance based on fuzzy clustering algorithm for image segmentation. Digital Signal Processing, 43, 8-16.

[7] Long, B., Xian, W., Li, M., Wang, H. (2014). Improved diagnostics for the incipient faults in analog circuits using LSSVM on PSO algorithm with Mahalanobis distance. Neurocomputing, 133(10), 237-248.

[8] Patil, N., Das, D., Pecht, M. (2015). Anomaly detection for IGBTs using Mahalanobis distance. Microelectronics Reliability, 55(7), 1054-1059.

[9] Hulten, G., Spencer, L., Domingos, P. (2001, August). Mining time-changing data streams. KDD, Processing of the seventh ACM SIGKDD international conference on knowledge discovery and data mining.

[10] Ghannadpour, S. S., Hezarkhani, A. (2012). A developed software to calculate the additive constant number of average in three-variable normal

logarithm. Global Journal of Computer sciences, 3(1), 1-6.

[11] Ghannadpour, S. S., Hezarkhani, A., Eshqi, H. (2012). Average and variance estimation programming in normal logarithmic distribution. Global Journal of Computer sciences, 2(1), 7-13.

[12] Ghannadpour, S. S., Hezarkhani, A., Sabetmobarhan, E. (2015). Some statistical analyses of Cu and Mo variates and geological interpretations for Parkam Porphyry Copper system, Kerman, Iran. Arabian Journal of Geosciences, 8(1), 345–355.

[13] Ghorbani, M (2002). "The Economic Geology of Iran." Arian Earth Press, Tehran (In Persian).

[14] Ghannadpour, S. S. and A. Hezarkhani (2015). "Assessment of prospective areas for providing the geochemical anomaly maps of Cu and Mo in Parkam district, Kerman, Iran." Journal of Researches in Earth Sciences 6(21), 40-50 (In Persian).

[15] Berberian, M., and King, G. C. (1981). Towards a Paleogeography and Tectonic Evolution of Iran. Canadian Journal of Earth Sciences, 18(2), 210–265.

[16] Saric, A., Diordjevic, M., Dimitrijevic, M. N. (1971). Geological map of Shahre-e-Babak, 1:100,000 Seri. Geological Survey of Iran, Tehran, Iran.

[17] Filzmoser, P., Garrett, R. G., Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. Computers & Geosciences, 31(5), 579–587.

[18] Hassani Pak, A. A. and M. Sharafaddin (2011). "Exploration data analysis." The second edit, Tehran University Press (In Persian).

[19] Yang, Y., Liu, X. (1999, August). A re-examination of text categorization methods. In proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99).

[20] Yang., Y. (1999). An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1(1), 69-90.

[21] He, J., Tan, A., Tan, C. (2000, August). Comparative Study on Chinese Text Categorization Methods. On the PRICAI 2000 Workshop on Text and Web Mining, Melbourne.

[22] Verbiest, N., Cornelis, C., Jensen, R. (2012, June). Fuzzy rough Positive region based Nearest Neighbor Classification. WCCI 2012 IEEE World congress on Computational Intelligence.

[23] Tan, K. C., Yu, Q. (2006). A coevolutionary algorithm for rules discovery in data mining. International Journal of Systems Science, 37(12), 835-864.

[24] Chan, C., Lewis, B. (2002). A basic primer on data mining. Information Systems Management, 19(4), 56-60.