نشریه علمی

علوم و مهندسی آبخیزداری ایران
Iran-Watershed Management
Science & Engineering

سال شانزدهم- شماره ۵۷- تابستان ۱۴۰۱ Vol. 16, No. 57, Summer 2022

پایش داده‌های مؤثر بر گرد و غبار با استفاده از طبقه‌بندی شورایی ایران مرکزی

محمد هاشمی نژاد[1]

تاریخ دریافت: ۰۹/۲۸/ ۱۴۰۰ تاریخ پذیرش: ۱۴۰۰/۱۲/۰۵

چکیده

امروزه استفاده از داده‌های ثبت شده در ایستگاه‌های سینوپتیک کشور یکی از مهم‌ترین منابع تحقیقات کاربردی برای پژوهش‌گران است. ایستگاه‌های سینوپتیک، اقلیم‌شناسی و غیره برای واکاوی‌های آماری مورد بررسی قرار می‌گیرند. در این تحقیق با استفاده از داده‌کاوی به روش طبقه‌بندی شورایی به پایش داده‌های موثر بر پدیده گرد و غبار ایستگاه‌های سینوپتیک ایران مرکزی پرداخته شد. در این مطالعه از داده‌های ۳۶ ایستگاه سینوپتیک واقع در استان‌های اصفهان، کرمان، یزد، سیستان و بلوچستان، سمنان، مرکزی، خراسان رضوی، قم و خراسان جنوبی استفاده شد. پارامترهای دمای متوسط روزانه، بارش روزانه، ارتفاع ایستگاه، موقعیت جغرافیایی ایستگاه، سرعت حداکثر باد، جهت سرعت حداکثر باد، نقطه شبنم در طبقه‌بندی شورایی مورد استفاده قرار گرفت. هم‌چنین مهم‌ترین عامل مؤثر در بین این پارامترها برای گرد و غبار عامل حداکثر سرعت باد است که در همه روش‌های طبقه‌بندی به‌عنوان مهم‌ترین عامل نشان داده شد. هم‌چنین سه طبقه‌بند KNN، SVM با کرنل RBF و شبکه عصبی MLP به عنوان اعضای شورا انتخاب شدند که با دقت ۹۰/۷ درصد منشأ تولید گرد و غبار (از منظر داخلی و خارجی بودن گرد و غبار) را به درستی تشخیص می‌دهد.

کلیدواژه‌ها: *داده‌کاوی، ایستگاه‌های سینوپتیک، گرد و غبار، طبقه‌بندی شورایی، فلات مرکزی ایران*

مقدمه

با توجه به اینکه کشور ایران بر روی کمربند خشکی‌های جهان قرار گرفته و بیش از ٦٦ درصد آن را اقلیم خشک و نیمه خشک تشکیل داده است بنابراین یکی از مهم‌ترین چالش‌های زیست محیطی غالب در کشور ایران پدیده گرد و غبار است. طوفان‌های گردوغبار همیشه به‌عنوان یکی از مخاطرات طبیعی شناخته می‌شوند که بخش‌های مختلفی شامل سلامت، کشاورزی و حمل‌ونقل را تحت تأثیر قرار می‌دهند و عواقب بسیار گسترده‌ای به‌ویژه کاهش حاصلخیزی خاک، آسیب زدن به محصولات کشاورزی، خشک شدن پوشش‌های گیاهی طبیعی، اختلال در سامانه‌های ارتباطی، اختلال در سامانه‌های مکانیکی [۱] و افزایش خطر بیماری‌های تنفسی [۲ و ۳] را به همراه دارد. به‌طورکلی به نظر می‌رسد عوامل طبیعی و انسان‌ساز طی سال‌های اخیر سبب افزایش فراوانی و شدت طوفان‌های گردوغبار در کشور شده‌اند. گرد و غبارها تأثیر قابل‌توجهی بر روی بودجه تابشی، چرخه‌های بیوژئو شیمیایی جهانی، ساختمان خاک ، ترکیبات شیمیایی اتمسفر، کیفیت هوا و سلامت و بهداشت عمومی می‌گذارند [۳، ٤، ٥ و ٦] پژوهش‌گران به‌دلیل اهمیت این پدیده و اثرات مهم آن بر روی زندگی انسان‌ها به صورت ضروری به مطالعه و تحلیل این پدیده به عنوان یک خطر طبیعی می‌پردازند. اکثریت تحلیل‌ها و بررسی‌ها از منظر آمار و اطلاعات برگرفته از داده‌های ثبت شده اقلیمی در ایستگاه‌های سینوپتیک است. دقت و صحت این داده‌ها می‌تواند نقش عمده‌ای در تحلیل نتایج و دستاوردها داشته باشد به همین منظور استفاده از داده‌کاوی و روش‌های مبتنی بر الگوریتم‌های عددی می‌تواند به تحلیل درست این پدیده کمک شایانی کند. به طور کلی داده‌کاوی فرآیند یافتن ناهنجاری ها، الگوها و همبستگی‌ها در مجموعه داده‌های بزرگ برای پیش‌بینی نتایج است. با استفاده از طیف وسیعی از تکنیک‌ها، می‌توانید از این اطلاعات برای افزایش درآمد، کاهش هزینه‌ها، بهبود روابط با مشتری، کاهش خطرات و موارد دیگر استفاده کنید. داده‌کاوی ، استخراج دانش در پایگاه داده‌ها نامیده می‌شود و روشی برای کشف اطلاعات سودمند جدید و بالقوه از بین حجم انبوهی از اطلاعات است [۸] . مفهوم داده‌کاوی شامل الگوریتم‌ها و روش‌هایی است که به استخراج اطلاعات از داده‌ها منجر می‌شود [۹ تا ۱۱] در مطالعه‌ای در زمینه منابع طبیعی از الگوریتم‌های داده‌کاوی برای تعیین حساسیت داده‌ها در کلاس‌های مختلف استفاده کردند. [۱۲] با استفاده از روش‌های داده‌کاوی از قبیل جنگل تصادفی و رگرسیون لجستیک به پهنه‌بندی گرد و غبار

۱ – نویسنده مسئول و استادیار مهندسی برق، گروه مهندسی برق، دانشکده فنی مهندسی، دانشگاه جیرفت. پست الکترونیکی: mhn@ujiroft.ac.ir

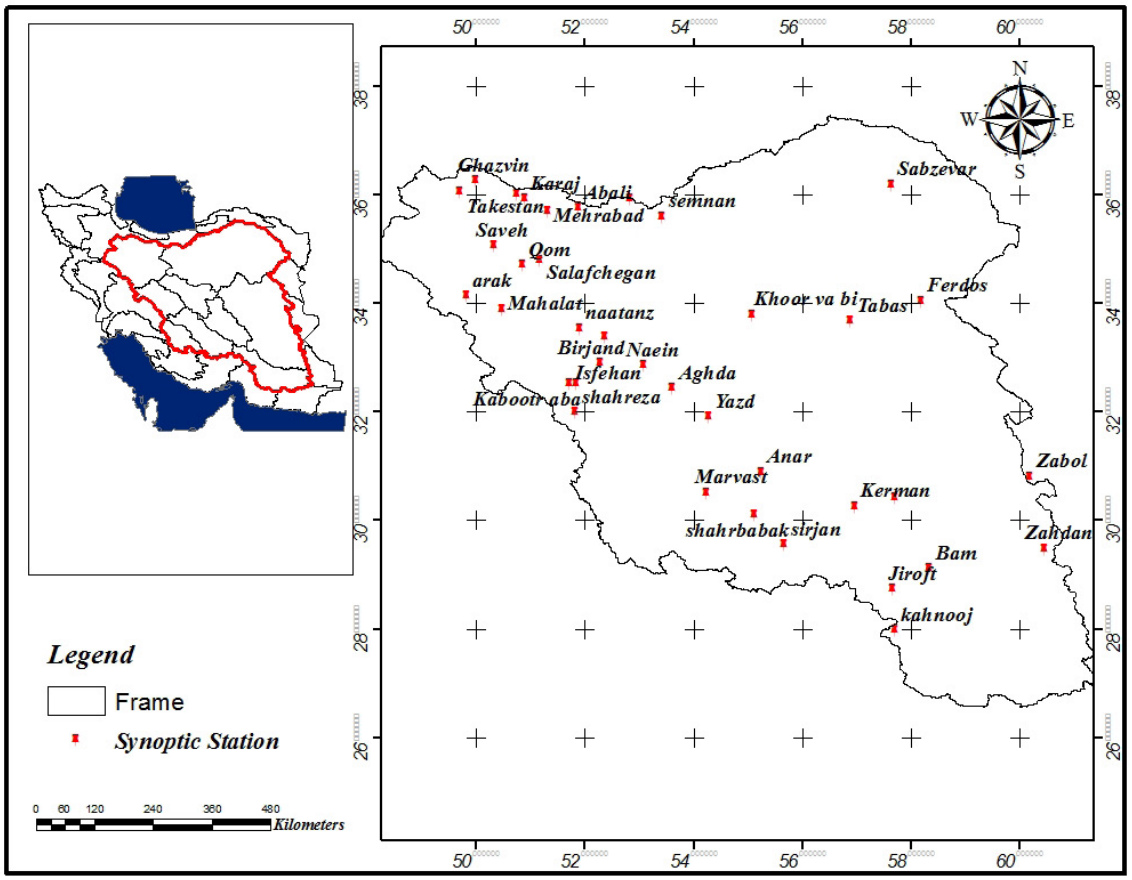
در استان خراسان رضوی پرداختند. [۱۳] با استفاده از سنجش از دور و علم داده‌کاوی خاستگاه هواویزهای اتمسفری در استان یزد را شناسایی کردند آن‌ها در این تحقیق ابتدا متغیرهای اقلیمی مختلف (از تصاویر ماهواره‌ای مختلف) از جمله سرعت باد در ارتفاع ده متری سطح زمین (Vs)، رطوبت خاک (Soil)، بارش تجمعی (Pr)، شاخص خشکسالی پالمر (Pdsi)، شاخص پوشش گیاهی نرمال شده (NDVI)، خشکی خاک یا کمبود آب خاک (Def)، تبخیر و تعرق مرجع (Pet) و واقعی (Aet)، بعد توپوگرافی (TD)، رادیانس طول موج کوتاه رسیده به زمین (Srad)، حداقل دمای هوا (Tmmn)، حداکثر دمای هوا (Tmmx)، فشار بخار (Vap)، کمبود فشار بخار(Vpd) و درصد رس (Clay) با استفاده از کدنویسی در سامانه آنلاین گوگل ارت انجین (GEE) استخراج شدند. سپس نمونه‌ها از مناطق بحرانی و مستعد گرد و غبار در سیستم اطلاعات جغرافیایی و به کمک تصاویر AOD مودیس استخراج شدند و این ویژگی و هم‌چنین سایر ویژگی‌ها در متغیرهای اقلیمی وارد سه مدل داده‌کاوی الگوریتم درختان رگرسیون و طبقه‌بندی (CART)، رگرسیون انطباقی چندمتغیره اسپیلاین (MARS) و درختان رگرسیون چندگانه جمع‌شدنی (TreeNet) شدند.

هدف از این تحقیق انتخاب مناسب‌ترین روش داده‌کاوی برای داده‌های ثبت شده در ایستگاه سینوپتیک کشور برای پدیده گرد و غبار، مشخص کردن مهم‌ترین پارامتر موثر بر روی پدیده گرد و غبار و در گام آخر با استفاده از فاکتورهای موجود صحت سنجی منشا گرد و غبار از منظر داخلی و یا خارجی بودن آن است.

مواد و روش ها

منطقه مورد مطالعه

یکی از مهم‌ترین چالش‌های فلات مرکزی ایران پدیده گرد و غبار است. وجود بیابان‌های کویر و لوت و هم‌چنین اقلیم طاقت فرسا در این منطقه باعث افزایش فراوانی این پدیده در این منطقه شده است. اقلیم فلات مرکزی ایران خشک و نیمه خشک بوده و هم‌چنین خاک آن خشک و پوشش گیاهی تنک دارد. نزولات جوی در این منطقه اندک بوده و در بسیاری از مناطق زیر ۱۰۰ میلی‌متر و گاهی حتی به زیر ۲۵ میلی‌متر نیز می‌رسد. تبخیر سالیانه بسیار بالا بوده و در مواردی به بیش از ٤۰۰۰ میلی‌متر در سال نیز می‌رسد. میانگین دما در این منطقه بین ۱۵ تا ۳۰ درجه متوسط روزانه تغییر می‌کند و دمای حداقل به ۱۸- درجه و دمای حداکثر گاهی تا ۵۱ درجه نیز افزایش پیدا می‌کند. موقعیت جغرافیایی منطقه مورد مطالعه در شکل ۱ نشان داده شده است.



شکل ۱ - موقعیت جغرافیایی منطقه مورد مطالعه

Fig 1. Geographical location of the study area

داده‌های مورد استفاده

به‌منظور تحلیل و بررسی داده‌های ثبت شده در ایستگاه‌های سینوپتیک از ۳۶ ایستگاه واقع در استان‌ها اصفهان، کرمان، یزد ، سیستان و بلوچستان، سمنان، مرکزی، خراسان رضوی، همدان، قم و خراسان جنوبی استفاده شد که اطلاعات آن‌ها در جدول ذیل آمده است. در این مطالعه از داده‌های مکانی (طول و عرض جغرافیایی ایستگاه، ارتفاع ایستگاه) و پارامتر های اقلیمی (متوسط دمای روزانه، سرعت باد، نقطه شبنم، سرعت حداکثر باد، جهت سرعت حداکثر باد) به عنوان پارامترهای مؤثر بر روی پدیده گرد و غبار استفاده شد.

جدول۱- مشخصات مربوط به ایستگاه‌های سینوپتیک منطقه مورد مطالعه

Table 1. Specifications related to synoptic stations in the study area

ردیف	نام ایستگاه Station name	استان Province	ارتفاع ایستگاه Elevation	عرض جغرافیایی Latitude	طول جغرافیایی Longitude
1	کهنوج Kahnooj	کرمان Kerman	499	57.71	27.99
2	جیرفت Jiroft	کرمان Kerman	722	57.76	28.70
3	بم Bam	کرمان Kerman	1067	58.35	29.10
4	شهربابک Shahrbabak	کرمان Kerman	834	55.13	30.10
5	کرمان Kerman	کرمان Kerman	1754	56.96	30.25
6	سیرجان Sirjan	کرمان Kerman	1739	55.68	29.46
7	انار Anar	کرمان Kerman	1409	55.25	30.88
8	زاهدان Zahedan	سیستان و بلوچستان Sistan and Baluchestan	1370	60.90	29.47
9	زابل Zabol	سیستان و بلوچستان Sistan and Baluchestan	1127	61.54	29.47
10	مروست Marvast	یزد Yazd	1547	54.20	30.40
11	یزد Yazd	یزد Yazd	1230	54.28	31.90
12	عقدا Aghada	یزد Yazd	1150	53.64	32.44
13	طبس Tabas	خراسان جنوبی Khorasan Jonoobi	711	56.95	33.60
14	اصفهان Isfahan	اصفهان Isfahan	1551.9	51.70	32.51
15	نایین Naein	اصفهان Isfahan	1573	53.07	32.85
16	کبوتر آباد Kabootarabad	اصفهان Isfahan	1542	51.83	32.51
17	خور و بیابانک Khour va biabanak	اصفهان Isfahan	842.2	55.08	32.77
18	شهر رضا Shahreza	اصفهان Isfahan	1858	51.81	31.98
19	نطنز Natanz	اصفهان Isfahan	1685	51.90	33.53
20	فردوس Ferdows	خراسان رضوی Khorasan Razavi	1292	58.18	34.03
21	سبزوار Sabzevar	خراسان رضوی Khorasan Razavi	963	57.64	36.20

22	اراک Arak	مرکزی Markazi	1702.8	49.78	34.07
23	محلات Mahalat	مرکزی Markazi	1622	50.48	33.88
24	ساوه Saveh	مرکزی Markazi	1111	50.37	35.08
25	قم Qum	قم Qum	879	50.85	34.77
26	سلفچگان Salafchegan	قم Qum	1381	50.46	34.48
27	سمنان Semnan	سمنان Semnan	1127	53.42	35.58
28	هشتگرد Hashtgerd	البرز Alborz	1612	50.74	36.00
29	کرج Karaj	البرز Alborz	1292	50.95	35.80
30	مهرآباد Mehrabad	تهران Tehran	1192	51.30	35.69
31	آبعلی Abali	تهران Tehran	2465	51.88	35.77
32	بیرجند Birjand	خراسان جنوبی Khorasan Jonoubi	1491	59.28	32.89
33	تاکستان Takistan	قزوین Qazvin	1283	50.60	36.05
34	قزوین Qazvin	قزوین Qazvin	1279	49.67	35.70

داده‌کاوی

داده‌کاوی فرآیند یافتن ناهنجاری ها، الگوها و همبستگی‌ها در مجموعه داده‌های بزرگ برای پیش بینی نتایج است. با استفاده از طیف وسیعی از تکنیک‌ها، می‌توانید از این اطلاعات برای افزایش درآمد، کاهش هزینه‌ها، بهبود روابط با مشتری، کاهش خطرات و موارد دیگر استفاده کنید.

داده‌ها را اغلب می‌توان به صورت یک ماتریس داده $n \times d$ با n ردیف و d ستون نشان داد یا خلاصه کرد، جایی که ردیف‌ها با عناصر موجود در مجموعه داده مطابقت دارند و ستون‌ها نشان دهنده ویژگی‌ها یا ویژگی‌های مربوطه هستند. هر ردیف در ماتریس داده، مقادیر مشخصه مشاهده شده را برای یک موجودیت معین ثبت می‌کند. ماتریس داده $n \times d$ به صورت زیر داده می‌شود:

$$D = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathrm{X}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathrm{X}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{X}_n & x_{n2} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \quad (۱)$$

که در آن x_i نشان دهنده ردیف i است که یک d تایی است که به صورت زیر است:

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{id}) \quad (۲)$$

و x_j نشان‌دهنده ستون j است که یک n تایی است که به صورت داده شده است:

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj}) \quad (۳)$$

بسته به دامنه کاربرد، ردیف‌ها ممکن است به عنوان موجودیت‌ها [۱٤]، نمونه‌ها [۱۵]، رکوردها [۱٦]، تراکنش‌ها [۱۷]، اشیا [۱۸]، نقاط [۱۹]، بردارهای ویژگی [۲۰] و غیره نیز نامیده شوند. به همین ترتیب، ستون‌ها را می‌توان ویژگی‌ها، خصوصیات، ابعاد، متغیرها، فیلدها و غیره نیز نامید. تعداد نمونه‌های n به عنوان اندازه داده‌ها نامیده می‌شود، در حالی که تعداد ویژگی‌های d را ابعاد داده می‌گویند. تجزیه و تحلیل یک ویژگی منفرد، تحلیل تک متغیره و تجزیه و تحلیل هم‌زمان دو ویژگی را تجزیه و تحلیل دو متغیره و تجزیه و تحلیل هم‌زمان بیش از دو ویژگی را تحلیل چند متغیره می‌نامند.

هم‌چنین شایان ذکر است که تحلیل داده‌های سنتی فرض می‌کند که هر موجودیت یا نمونه مستقل است. با این حال، با توجه به ماهیت به هم پیوسته جهانی که در آن زندگی می‌کنیم، این فرض ممکن است همیشه صادق نباشد. نمونه‌ها ممکن است از طریق انواع مختلفی از روابط به نمونه‌های دیگر متصل شوند، که منجر به ایجاد یک نمودار داده می‌شود، جایی که یک گره یک موجودیت را نشان می‌دهد و یک لبه نشان‌دهنده رابطه بین دو موجودیت است.

طبقه‌بندی شورایی

طبقه‌بند شورایی به گروهی از طبقه‌بندها اطلاق می‌شود که در یک مسئله طبقه‌بندی نظارت شده، به‌طور مشارکتی روی یک مجموعه داده آموزش داده می‌شوند.

در یک مسئله نظارت شده، از تعدادی نمونه برای آموزش استفاده می‌شود که یک برچسب کلاس به هر نمونه تعلق دارد. بسته به وضعیت داده‌های آموزشی و نوع طبقه‌بند، ممکن است که یک طبقه‌بند به خوبی آموزش داده نشده باشد. در چنین حالتی کارایی روی داده‌های آزمون ضعیف خواهد بود [۲۱] .

یک راه حل برای این مشکل آموزش چند طبقه‌بند روی یک مسئله است. در متون تحقیقاتی به این روش طبقه‌بندی شورایی[1] می‌گویند. هم‌چنین به هر کدام از این طبقه‌بندهایی که در این گروه هستند طبقه‌بند پایه یا ضعیف یا خبره گفته شده است [۲۱] .

در طول پیش‌بینی، طبقه‌بندی‌کننده‌های پایه تصمیمی را در مورد یک الگوی آزمایشی ارائه می‌کنند. سپس یک روش هم‌جوشی تصمیمات تولید شده توسط طبقه‌بندی‌کننده‌های پایه را ترکیب می‌کند. تعداد زیادی روش ادغام در ادبیات وجود دارد، از جمله رای اکثریت، شمارش بوردا، ترکیب کننده‌های جبری و غیره [۲۲].

فلسفه طبقه‌بندی‌کننده گروهی این است که طبقه‌بندی‌کننده پایه دیگری خطاهای یک طبقه‌بندی‌کننده پایه را جبران می‌کند. با این حال، آموزش طبقه‌بندی کننده پایه به روشی ساده، این مشکل را حل نمی کند. همان‌طور که در [۲۳] اشاره شد، اگر طبقه‌بندی‌کننده‌های پایه دقیق و متنوع باشند، یک طبقه‌بندی‌کننده گروهی بهتر از همتای پایه خود عمل می‌کند. واژه تنوع به این واقعیت اشاره دارد که خطاهای طبقه‌بندی کننده‌های پایه همبستگی ندارند. روش‌های خوبی برای محاسبه تنوع وجود دارد از جمله اندازه‌گیری‌های تنوع زوجی (آمار Q، ضریب همبستگی، اندازه‌گیری اختلاف، اندازه‌گیری خطای دوگانه) و اندازه‌گیری‌های تنوع غیر زوجی (معیار آنتروپی، واریانس کوهاوی–ولپرت، اندازه‌گیری قرارداد بین ارزیاب) [۲۳] .

هدف روش‌های مختلف تولید شورای طبقه‌بندی دستیابی به تنوع در میان طبقه‌بندی‌کننده‌های پایه است. برخی از طبقه‌بندی‌کننده‌های شورایی نیز با هدف مواجهه با یک مشکل یا برنامه خاص توسعه یافته‌اند. بخش زیر طبقه‌بندی‌کننده‌های پایه مختلف، طبقه‌بندی‌کننده‌های گروهی و کاربردهای آن‌ها را توضیح می‌دهد.

طبقه‌بندی پایه

طبقه‌بندی‌کننده‌های پایه به هر کدام از طبقه‌بندی کننده‌هایی اشاره دارند که برای ساخت شورای طبقه‌بندی استفاده می‌شود. شبکه‌های عصبی، ماشین‌های بردار پشتیبان و k-NN برخی از طبقه‌بندی‌کننده‌های پایه هستند که معمولاً مورد استفاده قرار می‌گیرند. به منظور کامل بودن، به طور خلاصه روند آموزش و آزمون این دسته‌بندی کننده‌های پایه را توضیح می‌دهیم. در طبقه‌بندی k-NN، فاصله بین یک الگوی آزمایشی و همه الگوهای مجموعه آموزشی محاسبه می‌شود. فاصله را می‌توان با استفاده از فاصله‌هایی از قبیل اقلیدسی یا فاصله منهتن محاسبه کرد. طبقات احتمالی حائز رأی هر یک از k الگو که فاصله کم‌تری به الگوی آزمون دارند، می‌شوند. کلاسی که بالاترین رای را کسب کند، کلاس الگوی آزمون در نظر گرفته می‌شود.

اگر x داده آزمون باشد، مجموعه k نزدیک‌ترین همسایه x را به صورت S_x نشان دهید. طبق تعریف S_x به این صورت تعریف می‌شود:

$$S_x \subseteq D \; s.t. |S_x| = k \\ \forall (x', y') \in D \setminus S_x, \quad (٤)$$

$$dist(x, x') \geq \max_{(x'', y'') \in S_x} dist(x, x'') \quad (٥)$$

(یعنی هر نقطه در D اما نه در S_x حداقل به اندازه دورترین نقطه در S_x از x دور است). سپس می‌توان طبقه‌بندی‌کننده h0 را به عنوان تابعی تعریف کرد که رایج‌ترین برچسب را در S_x برمی‌گرداند:

$$h(x) = mode(\{y'' : (x'', y'') \in Sx\}) \quad (٦)$$

که در آن $mode(\cdot)$ به معنای انتخاب برچسب بالاترین رخداد است.

(نکته: در صورت تساوی، یک راه حل خوب این است که نتیجه k-NN را با k کوچکتر برگردانید.)

یک شبکه عصبی [۲٤] را می‌توان به عنوان یک سیستم محاسباتی متشکل از تعدادی عناصر پردازشی ساده و بسیار به هم پیوسته در نظر گرفت که اطلاعات را با پاسخ حالت پویای خود به ورودی‌ها پردازش می‌کند. ساختار شبکه‌های عصبی شامل تعدادی لایه است. لایه‌ها از تعدادی گره به هم پیوسته تشکیل شده‌اند که هر کدام حاوی یک تابع فعال‌سازی است. الگوها از طریق لایه ورودی به شبکه ارائه می‌شوند، که با یک یا چند لایه پنهان ارتباط برقرار می‌کند، جایی که پردازش واقعی از طریق سیستمی از اتصالات وزنی انجام می‌شود. بعد از لایه‌های پنهان یک لایه خروجی است که در آن پاسخ خروجی است. اکثر شبکه‌های عصبی حاوی یک قانون یادگیری هستند که وزن اتصالات را با توجه به الگوهای ورودی که با آن ارائه می‌شود تغییر می‌دهد. تابع فعال‌سازی یک نورون که نتیجه جمع را به صفر یا یک نگاشت می‌دهد در معادله **Error!** **Reference source not found.** نشان داده شده است.

$$f(X) = \begin{cases} 1 \; W^T X + w_0 > 0 \\ 0 \; W^T X + w_0 \leq 0 \end{cases}$$

$$where \; W = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ w_m \end{bmatrix} \& X = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_m \end{bmatrix} \quad (٧)$$

1. Ensemble Classification

$$net_j = \sum_{i=1}^{d} x_i \omega_{ij} + \omega_{j0} = \sum_{i=0}^{d} x_i \omega_{ij} \Rightarrow y_j = f(net_j) \quad (۸)$$

که در آن f تابع فعال‌سازی است.

$$net_k = \sum_{j=1}^{nH} y_j \omega_{kj} + \omega_{k0} = \omega_k^T y \quad (۹)$$

که در آن nH نشان دهنده تعداد پرسپترون‌ها در لایه پنهان است و w_o واحدهای بایاس هستند.

بنابراین نورون‌های خروجی z را می‌توان به صورت معادله (۱۰) استخراج کرد.

$$z_k = f(net_k) = sgn(net_k) \Rightarrow z_k = f\left(\sum_{j=1}^{nH} \omega_{kj} f\left(\sum_{i=1}^{d} x_i \omega_{ji} + \omega_{j0}\right) + \omega_{k0}\right) \quad (۱۰)$$

یک SVM [۲۵] با استفاده از یک تابع کرنل داده‌ها را به داده‌های جدید با ابعاد بالاتر تبدیل می‌کند. سپس با یافتن بهترین ابر صفحه‌ای را که الگوهای یک کلاس را از کلاس دیگر جدا می‌کند، طبقه‌بندی را انجام می‌دهد. بهترین ابرصفحه برای SVM به صفحه‌ای اطلاق می‌شود که بیشترین حاشیه بین کلاس‌ها را داشته باشد. حاشیه به معنای حداکثر عرض صفحه موازی با ابر صفحه است که هیچ الگویی داخل آن نیست. بردارهای پشتیبان داده‌هایی هستند که نزدیک‌ترین به ابر صفحه جداکننده هستند. این نقاط در مرز صفحه موازی قرار دارند. در این مقاله از SVM با دو نوع تابع کرنل مختلف به عنوان طبقه‌بندی کننده پایه استفاده شده است. در حالت n بعدی فرمول این ابر صفحه که در SVM خطی به این صورت است:

$$\sum_{i=1}^{n} w_1 . x_1 + b = 0 \quad (۱۱)$$

انواع مختلفی از کرنل برای SVM وجود دارد که از آن جمله می‌توان به کرنل‌های چند جمله‌ای، گوسی و تابع پایه شعاعی[1] اشاره کرد که به این صورت تعریف می‌شود:

$$k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \quad \gamma > 0 \quad (۱۲)$$

روش‌ها تولید طبقه‌بندی‌کننده شورایی را می‌توان به طور کلی به شش گروه [۲۶] تقسیم کرد که بر اساس (۱) دستکاری پارامترهای آموزشی که با استفاده از آن می‌توان به گوناگونی در خبره‌ها رسید. به عنوان مثالی از این نوع می‌توان به استفاده از وزن‌های مختلف در [۲۷] اشاره کرد. (۲) دستکاری تابع خطا، در این نوع اگر خطاهای دو طبقه‌بند در الگوهای یکسان، مشابه هم بود مقداری را به خطا اضافه می‌کند از این نوع می‌توان به Negative correlation learning اشاره کرد که در [۲۸] استفاده شده است. (۳) گوناگونی در فضای ویژگی با استفاده از زیرمجموعه‌های مختلف از داده‌ها [۲۹] ، (٤) گوناگونی در برچسب‌های خروجی [۳۰] ، (۵) خوشه‌بندی داده‌های آموزشی به بخش‌ها بدون همپوشانی [۳۱]، و (٦) گوناگونی در الگوهای تولید داده‌های آموزشی [۳۲] .

همه طبقه‌بندهای پایه و قاعده ترکیب با استفاده از کتابخانه sklearn در محیط برنامه‌نویسی پایتون پیاده‌سازی شدند. بهترین پارامترها برای همه طبقه‌بندهای پایه با استفاده از سعی و خطا به‌دست آمدند. KNN یک پارامتر برای تنظیم دارد. این پارامتر تعداد نزدیک‌ترین همسایه‌ها است که در مقدار ۳ تنظیم شد. برای درخت تصمیم حداقل تعداد نمونه برای تقسیم یک گره داخلی که ۲ در نظر گرفته شد و حداکثر عمق درخت که محدودیتی برای آن در نظر گرفته نشد. برای جنگل تصادفی، از تعداد ۱۰ درخت‌ها در جنگل عمق نا محدود استفاده شد. پارامتر دیگر جنگل تصادفی تعداد ویژگی‌هایی که هنگام جستجوی بهترین تقسیم باید در نظر گرفته شود است که مقدار یک به این منظور اختصاص داده شد. برای دو روش SVM با کرنل خطی و تابع کرنل پایه شعاعی با مدل و تنظیمات پیش‌فرض استفاده شدند.

تعداد کل نمونه‌های استفاده شده ۵۲۲٦٦ نمونه است. از این تعداد ۷۰ درصد برای آموزش و ۳۰ درصد برای ارزیابی طبقه‌بندهای پایه استفاده شد. هر کدام از اعضای شورا یک احتمال تعلق منشأ گرد و غبار را به هر کدام از کلاس‌ها (منشأ داخلی و منشأ خارجی) نسبت می‌دهد و مجموع از دو احتمال یک است. در نتیجه امتیاز شباهت (احتمال) به کلاس منشأ داخلی را به عنوان یک بردار ورودی، که ابعاد آن با تعداد طبقه‌بندهای پایه یکسان است، به ورودی شبکه عصبی می‌دهیم تا امتیاز احتمال هر کلاس را مشخص کند. در این حالت تعداد ۱۵٦۸۰ بردار ورودی استفاده شد که ۵۰ درصد آن‌ها برای آموزش شبکه عصبی (به عبارتی قاعده ترکیب شورا) و ۵۰ درصد باقی‌مانده برای ارزیابی طبقه‌بند شورایی نهایی درنظر گرفته شد تا دقت نهایی بدست آید. شبکه عصبی که به عنوان قاعده ترکیب استفاده شد دارای ۱۰ لایه‌های مخفی، تابع فعال‌سازی RELU[2] و نرخ یادگیری ۰/۰۰۱ استفاده شد.

نتایج

شش روش مختلف طبقه‌بندی شامل درخت تصمیم، جنگل تصادفی، شبکه عصبی MLP، k نزدیک‌ترین همسایه (KNN) و ماشین‌های بردار پشتیبان (SVM) با دو نوع کرنل خطی و RBF پیاده‌سازی شد که نتایج آن در نمایش داده شده است. برای سادگی نمایش SVM با کرنل RBF را به اختصار RBF SVM می‌نامیم.

1. Radial Basis Function.

2. Rectified Linear Unit

جدول ۲ – نتایج صحت سنجی داده‌ها با استفاده از شش طبقه‌بند مختلف

Table 2. Results of data validation using six different classifiers

طبقه‌بند Classifier / معیار Criterion	ریشه میانگین مربعات خطا Root mean square error	دقت آموزش Training accuracy	دقت ارزیابی Evaluation accuracy
KNN	0.3294	94%	89%
Decision tree	0.3587	100%	87%
MLP	0.3224	89%	90%
Linear SVM	0.3385	89%	89%
RBF SVM	0.3125	90%	90%
Random forest	0.3476	88%	88%

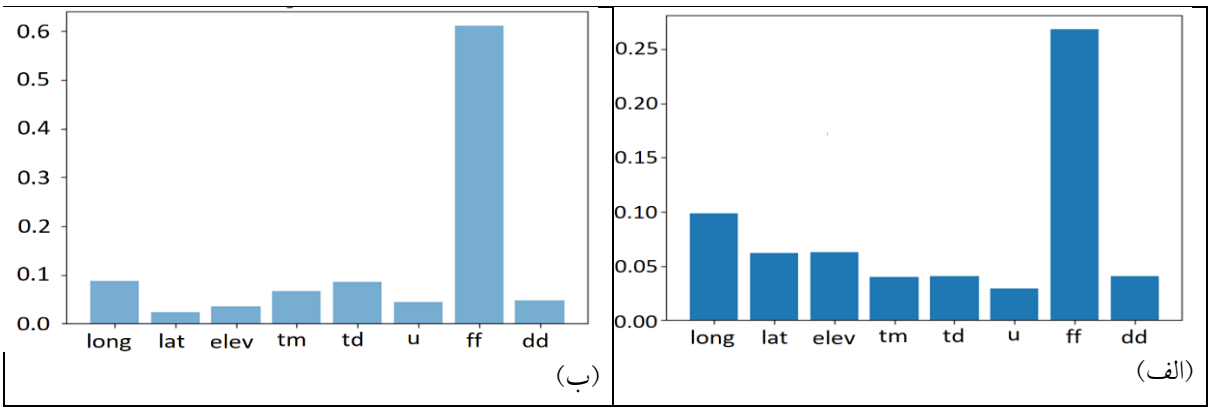
معیارهای ارزیابی روش‌های مختلف را به طور تقریبی نشان می‌دهد. روش‌های شبکه عصبی مصنوعی و RBF SVM دارای بیش‌ترین دقت هستند. همان‌گونه که از **Error! Reference source not found.** مشخص است، تعداد کل نمونه‌های آزمون ۱۵٦۸۰تا است. از این تعداد ۱٤۰۵۸ تشخیص درست در روش MLP ثبت شده است. در نتیجه دقت طبقه‌بندی برای این روش به طور دقیق‌تر ۸۹/٦٦ درصد است. با همین محاسبه برای روش RBF SVM به دقت ۹۰/۲٤ درصد می‌رسیم که بالاترین دقت در بین همه روش‌ها است.

در تخمین نوع گرد و غبار در این تحقیق از ۸ ویژگی مختلف استفاده شده است. که عبارتند از مشخصات مکانی ایستگاه طول جغرافیایی (longitude)، عرض جغرافیایی (latitude)، ارتفاع (Elevation)، پارامترهای اقلیمی متوسط دمای روزانه(t_m)، نقطه شبنم(t_d)، ارتفاع باد (u)، ماکزیمم سرعت باد (ff_{max}) و جهت سریعترین باد (dd_{max}). همان‌گونه که به طور ضمنی انتظار می‌رود، هر کدام از ویژگی‌های استفاده شده اهمیت متفاوتی را در طبقه‌بندی دارند که در شکل ۲ میزان اهمیت هر کدام از پارامترهای بر حسب روش مورد استفاده نشان داده شده است.

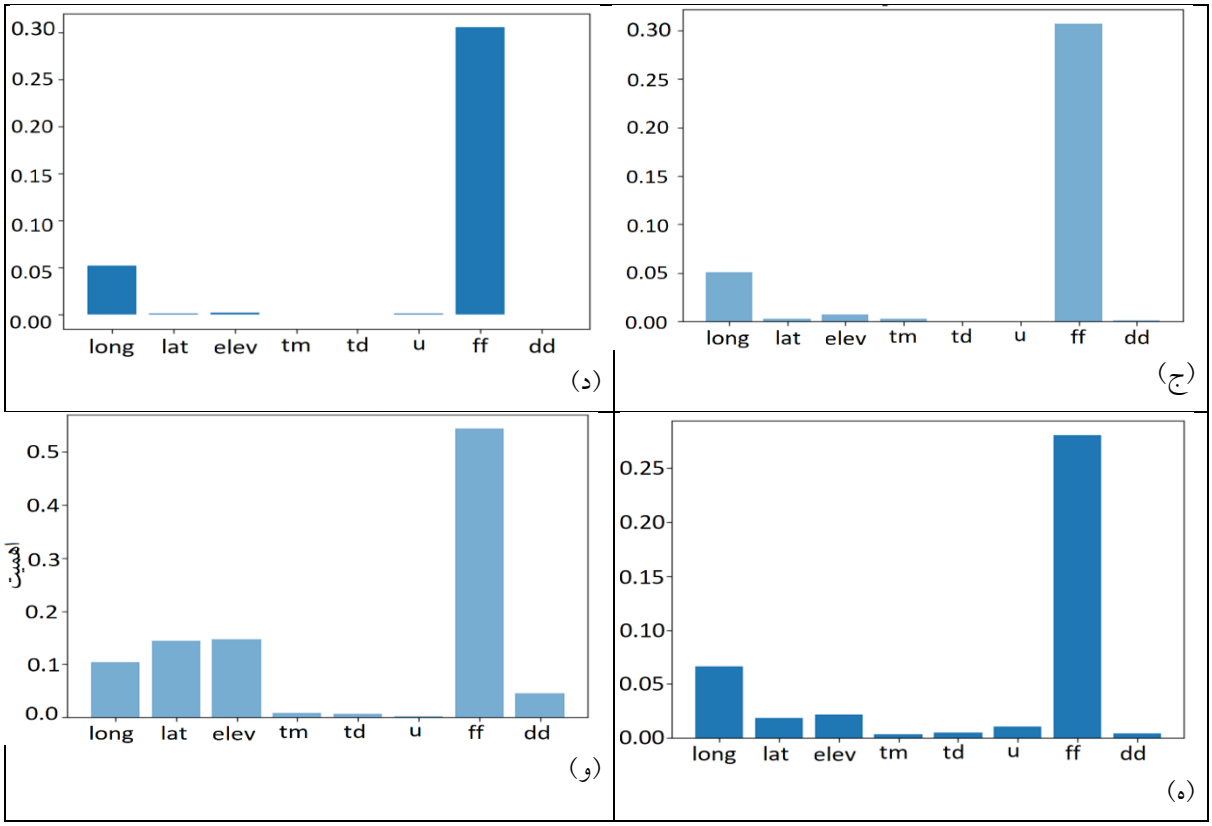
شاخص ROC

یکی دیگر از معیارهایی که برای ارزیابی کارایی روش‌های طبقه‌بندی مهم و مورد استفاده است، نمودار ROC[1] است. در تشخیص تعلق یک نمونه به یک کلاس معمولا یک امتیاز شباهت به دست می‌آید. اگر این امتیاز از یک حد آستانه بیشتر باشد نمونه آزمون به کلاس مربوطه نسبت داده می‌شود. نمودار گرافیکی ROC قابلیت تشخیص یک سیستم طبقه‌بندی دودویی را در حدود آستانه مختلفی که برای تشخیص در نظر می‌گیریم نشان می‌دهد. این نمودار برای استفاده از مزیت طبقه‌بندی شورایی انواع مختلفی از آن را با استفاده از ایجاد تنوع در نوع طبقه‌بند ایجاد کردیم. در ابتدا هر شش طبقه‌بند را به عنوان عضو شورا استفاده کردیم.

در این تحقیق ما از شبکه عصبی مصنوعی برای نتیجه‌گیری نهایی از روی همه امتیازهای به‌دست آمده استفاده کردیم. برای آموزش این شبکه عصبی ۵۰ درصد داده‌های آزمون استفاده شد و ۵۰ درصد مابقی برای ارزیابی شبکه ایجاد شده استفاده شد. با توجه به اینکه در قاعده ترکیب امتیاز طبقه‌بندهای پایه به عنوان ورودی سیستم استفاده می‌شوند، تحلیل اهمیت ویژگی در این قسمت اعضای تأثیر گذار شورا را مشخص می‌کند. اولین آزمایش‌های شورایی را با استفاده از همه امتیازهای به دست آمده ده بار تکرار کردیم میانگین دقت این آزمایش ۹۰/۳۲ درصد را نشان می‌دهد. این دقت از دقت بهترین روش پایه که RBF SVM است بالاتر است که در شکل ۳ نشان داده شده است.

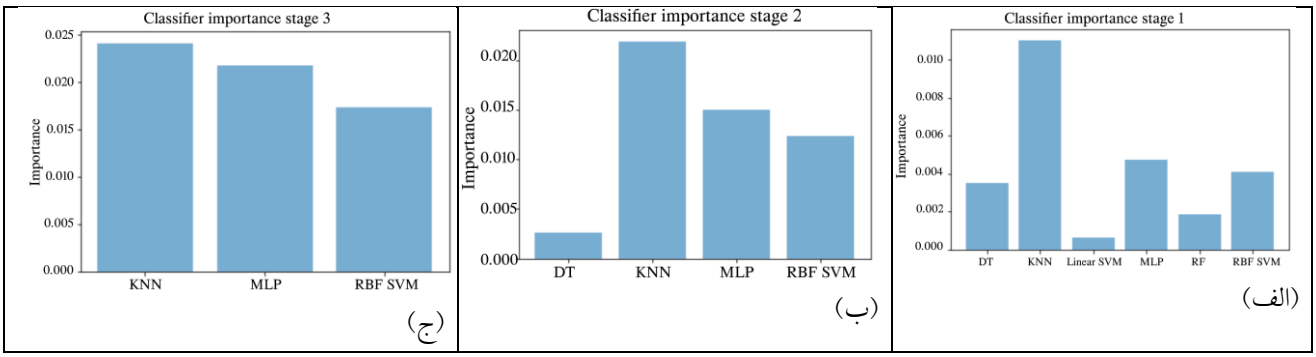


1. Receiver Operating Characteristic



شکل ۲ – نمودار اهمیت ویژگی برای روش (الف): kNN، (ب): درخت تصمیم، (ج): شبکه عصبی مصنوعی، (د): linear SVM، (ه): RBF SVM، (و): جنگل تصادفی

Fig 2. Graph of feature Importance for method (a): kNN, (b): decision tree, (c): artificial neural network, (d): linear SVM, (e): RBF SVM, (f): random forest



شکل ۳- اهمیت روش‌های پایه در طبقه‌بندی شورایی با استفاده از (الف) همه شش روش اولیه (ب) چهار روش منتخب RBF SVM، شبکه عصبی، KNN، درخت تصمیم و (ج) سه روش RBF SVM، شبکه عصبی و KNN

Fig 3. Importance of basic methods in council classification using (a) all six primary methods (b) four selected methods RBF SVM, neural network, KNN, decision tree and (c) three methods RBF SVM, neural network and KNN

با این حال نمودار شکل (الف) نشان می‌دهد که دو روش SVM خطی و درخت تصادفی تأثیر زیادی در شورا ندارند. در آزمایش بعدی با حذف این دو روش از شورا، یک شورای جدید را ایجاد کردیم که دقت میانگین ۹۰/۴۳ را در ده آزمایش انجام شده داشت.

مجدداً با رسم نمودار میله‌ای اهمیت ویژگی که در اینجا همان میزان تأثیر گذاری روش پایه در شورا راست را با چهار روش باقی‌مانده، در شکل نمایش داده‌ایم. از این نمودار مشخص است که تأثیر گذاری درخت تصمیم کم‌تر از دیگر اعضای شورا است. بنابراین در

آزمایشی دیگر تأثیر حذف این طبقه‌بند را در شورا بررسی کردیم. در این آخرین آزمایش دقت تشخیص به ۹۰/۷ درصد ارتقاء می‌یابد. شکل (ج) اثر بخشی اعضای شورای نهایی را نشان می‌دهد.

بحث و نتیجه‌گیری

با توجه به اینکه پدیده گرد و غبار یکی از مهم‌ترین بلایایی طبیعی در فلات مرکزی ایران است، آمار و اطلاعات دریان باره از اهمیت خاصی برخوردار است. به‌همین منظور روش‌های داده‌کاوی برای صحت‌سنجی داده‌های ایستگاه‌های سینوپتیک برای تحلیل و بررسی بسیار ضروری است. در این تحقیق ما از روش‌های مختلفی برای پیش‌بینی منشأ گرد و غبار استفاده کردیم. تحقیقات ما نشان می‌دهد استفاده از روش ماشین‌های بردار پشتیبان (SVM) به همراه کرنل پایه شعاعی بهترین نتایج را در این پیش بینی دارد. هم‌چنین ما نتیجه گرفتیم که چند طبقه‌بند مختلف می‌توانند به عنوان مکمل یکدیگر و به عنوان یک شورا کارایی بهتری را از همه روش‌های پایه در تشخیص منشأ گرد و غبار داشته باشند. نتایج نشان داد که وجود برخی از طبقه‌بندها در شورا نه تنها به کارایی شورا کمکی نمی‌کند که دقت را پایین می‌آورد در حالی‌که کارایی قابل قبول را به صورت منفرد از خود نشان می‌دهند. در نهایت مشخص شد روش‌های RBF SVM، شبکه عصبی و KNN می‌توانند مکمل‌های مناسبی برای تشخیص منشأ گرد و غبار از روی باشند. در ادامه به تحقیقاتی استفاده از روش داده‌کاوی برای پدیده گرد و غبار را تایید می‌کند اشاره می‌شود: [۳۳] با استفاده از روش‌های داده‌کاوی شبکه عصبی به پیش بینی و مدل‌سازی پدیده گرد و غبار در غرب ایران پرداختند و دریافتند که روش‌های داده‌کاوی میتوانند به خوبی و با صحت بالا این پدیده را پیش بینی و مدل‌سازی کنند. هم‌چنین مقایسهٔ دو مدل شبکهٔ عصبی ANFIS و RBF در بهترین شرایط نشان داد که مقدار RMSE مدل ANFIS برابر با ۱۱/۶۷ و مدل RBF برابر با ۲/۱۹ است. بنابراین، قدرت دقت RBF در پیش‌بینی گرد و غبار در سال‌های شبیه‌سازی‌شده بیشتر است. بروغنی و همکاران در [۱۲] با استفاده از روش داده‌کاوی مدل درخت رگرسیون تقویت شده BRT به مدل‌سازی مکانی حساسیت کانون‌های گرد و غبار در شرق ایران پرداختند با توجه به مطالعات صورت گرفته، در این تحقیق هشت عامل مؤثر شامل کاربری اراضی، زمین‌شناسی، درجه‌ی شیب، ارتفاع، شاخص پوشش گیاهی نرمال شده (NDVI)، فاصله از رودخانه، سرعت باد و بارش شناسایی و نقشه‌ی این عوامل در محیط GIS تهیه و جهت ارزیابی نتایج از منحنی ROC استفاده شد. نتایج نشان داد که مدل BRT با مساحت سطح زیر منحنی (۷۹/۶ درصد) کارایی نسبتاً بالایی در تهیه‌ی نقشه‌ی حساسیت گردوغبار در منطقه‌ی مورد مطالعه را داراست.

منابع

1. P. J. Prakash, G. Stenchikov, S. Kalenderski, S. Osipov, and H. Bangalath, 2014, The impact of dust storms on the Arabian Peninsula and the Red Sea. Atmospheric Chemistry & Physics Discussions, 14(13), 199-222 .

2. M. B. Lyles, H. L. Fredrickson, A. J. Bednar, H. B. Fannin, D. W. Griffin, and T. M. Sobecki, 2012, Medical geology in the Middle East: potential health risks from mineralized dust exposure. in EGU General Assembly Conference Abstracts, p. 1668.

3. S. J. A. Ebrahimi, L. Ebrahimzadeh, A. Eslami, and F. Bidarpoor, 2014, Effects of dust storm events on emergency admissions for cardiovascular and respiratory diseases in Sanandaj, Iran. Journal of Environmental Health Science and Engineering, 12(1): 1–5.

4. A. Chappell, J. Sanderman, M. Thomas, A. Read, and C. Leslie, 2012, The dynamics of soil redistribution and the implications for soil organic carbon accounting in agricultural south-eastern Australia. Global Change Biology, 18(6): 2081–2088.

5. O. A. Chadwick, L. A. Derry, P. M. Vitousek, B. J. Huebert, and L. O. Hedin, 1999, Changing sources of nutrients during four million years of ecosystem development. Nature, 397(6719): 491–497.

6. Jickells TD, An ZS, Andersen KK, Baker AR, Bergametti G, Brooks N, Cao JJ, Boyd PW, Duce RA, Hunter KA, Kawahata H., 2005, Global iron connections between desert dust, ocean biogeochemistry, and climate. science, 308(5718):67-71.

7. R. Reynolds, J. Belnap, M. Reheis, P. Lamothe, and F. Luiszer, 2001, Aeolian dust in Colorado Plateau soils: nutrient inputs and recent change in source, Proceedings of the National Academy of Sciences, 98(13): 7123–7127.

8. U. Fayyad, G. S. Piatetsky-Shapiro, and P. Smyth, 1996, From data mining to knowledge discovery in databases. AI magazine, 17(3): 37–54.

9. M. Akbari and M. Bashiri, 2017, Application of Data Mining Algorithms to Appreciate Sensitivity and Spatial Zoning Prone to Floating View in Khorasan Razavi Display Basins. Journal of Environmental Erosion Research, 7(26): 16-42. (In Persian)

10. T. Can, H. A. Nefeslioglu, C. Gokceoglu, H. Sonmez, and T. Y. Duman, 2005, Susceptibility assessments of shallow earthflows triggered by heavy rainfall at three catchments by logistic regression analyses. Geomorphology, 72(1–4): 250–271.

11. M. J. García-Rodríguez, J. A. Malpica, B. Benito, and M. Díaz, 2008, Susceptibility assessment of earthquake-triggered landslides in El Salvador using logistic regression. Geomorphology, 95(3–4): 172–191.

12. M. Boroughani and S. Pourhashemi, 2019, Susceptibility Zoning of Dust Source Areas by Data Mining Methods over

Khorasan Razavi Province. Environmental Erosion Research Journal, 9(3):1-22. (In Persian)

13. M. Kazemi, A. R. Nafarzadegan, F. Mohammadi, and A. Rezaei Latifi, 2021, Identifying origins of atmospheric aerosols using remote sensing and data mining (Case study: Yazd province). Journal of RS and GIS for Natural Resources, 12(1): 13-16. (In Persian)

14. P. Panov, L. Soldatova, and S. Džeroski, 2014, Ontology of core data mining entities, Data Mining and Knowledge Discovery, 28(5): 1222-1265.

15. C. Clifton, 2000, Protecting against data mining through samples. in Research Advances in Database and Information Systems Security, Springer: 193–207.

16. F. M. Bianchi, A. Rizzi, A. Sadeghian, and C. Moiso, 2016, Identifying user habits through data mining on call data records. Engineering Applications of Artificial Intelligence 54: 49-61.

17. M. E. Lokanan, 2019, Data mining for statistical analysis of money laundering transactions. Journal of Money Laundering Control.

18. K. Chitra and D. Maheswari, 2017, A comparative study of various clustering algorithms in data mining. International Journal of Computer Science and Mobile Computing, 6(8): 109-115.

19. W. Lu, 2020, Improved K-means clustering algorithm for big data mining under Hadoop parallel framework. Journal of Grid Computing, 18(2): 239-250.

20. E. Atagün and I. D. Argun, 2020, Performance analysis of data mining software with parametric changes. International Journal of Forensic Software Engineering, 1(2-3): 115-143.

21. M. Hasheminejad and H. Farsi, 2018, Sample-specific late classifier fusion for speaker verification. Multimedia Tools and Applications, 77(12): 15273-15289.

22. J. C. Heckelman and R. Ragan, Symmetric scoring rules and a new characterization of the Borda count. Economic Inquiry, 59(1): 287-299.

23. Sun, T., & Zhou, Z. H., 2018, Structural diversity for decision tree ensemble learning. Frontiers of Computer Science, 12(3): 560-570.

24. F. Cheshmberah, H. Fathizad, G. A. Parad, and S. Shojaeifar, 2020, Comparison of RBF and MLP neural network performance and regression analysis to estimate carbon sequestration. International Journal of Environmental Science and Technology, 17(9): 3891-3900.

25. A. S. A. Yahya et al., 2019, Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios. Water, 11(6): 1231.

26. Rahman, A., & Verma, B., 2013, Ensemble classifier generation using non-uniform layered clustering and Genetic Algorithm. Knowledge-Based Systems, 43: 30-42.

27. Berkhahn, S., Fuchs, L., & Neuweiler, I., 2019, An ensemble neural network model for real-time prediction of urban floods. Journal of hydrology, 575: 743-754.

28. Alobaidi, M. H., Ouarda, T. B., Marpu, P. R., & Chebana, F., 2021, Diversity-driven ANN-based ensemble framework for seasonal low-flow analysis at ungauged sites. Advances in Water Resources, 147: 103814.

29. Chen, W., Hong, H., Li, S., Shahabi, H., Wang, Y., Wang, X., & Ahmad, B. B., 2019, Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. Journal of Hydrology, 575: 864-873.

30. P. A. Gutiérrez, M. Pérez-Ortiz, and A. Suárez, 2017 Class Switching Ensembles for Ordinal Regression, In International Work-Conference on Artificial Neural Networks, pp. 408-419. Springer, Cham, 2017

31. A. Rahman, B. Verma, and X. Yao, 2010, Non–uniform Layered Clustering for Ensemble Classifier Generation and Optimality, In International Conference on Neural Information Processing, pp. 551-558. Springer, Berlin, Heidelberg.

32. M. H. D. M. Ribeiro and L. dos Santos Coelho, 2020, Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Applied Soft Computing, 86, 105837.

33. S. Sobhani, B., Safarian Zangir, V., Fizoallahzade, 2020, Modeling and Prediction of Dust in Western Iran. Physical Geography Research Quarterly, 52(1): 17-35.

نشریه علمی
علوم و مهندسی آبخیزداری ایران
Iran-Watershed Management
Science & Engineering

Vol. 16, No. 57, Summer 2022
سال شانزدهم- شماره ۵۷- تابستان ۱۴۰۱

Validation of Synoptic Station Data Using Ensemble Classification on Central Iran

M. Hasheminejad[1]
Received: 19-12-2021 Accepted: 24-02-2022

Abstract

Today, the use of data recorded in synoptic stations of the country is one of the most significant sources of applied research for researchers. Data recorded automatically or manually at synoptic, climatological, and other stations are analyzed for statistical analysis. In this research, the data recorded in the synoptic stations of Iran, which are used to determine the days of dust, were analyzed using the science of monitoring and data analysis using ensemble classification. In this study, data from 36 synoptic stations, were used. These stations are in Isfahan, Kerman, Yazd, Sistan and Baluchestan, Semnan, Markazi, Khorasan Razavi, Hamedan, Qom, and South Khorasan. The parameters of daily average temperature, daily rainfall, station height, geographical location of the station, maximum wind speed, maximum wind speed, and dew point were used for the classification. The results showed that the most important factor among these parameters for dust is the maximum wind speed, which was identified as the most significant factor in all classification methods. Also, three classifiers, KNN, SVM with RBF kernel, and MLP neural network, were selected as members of the ensemble, which accurately detects 90.7 percent of the source of dust production (from the inside and outside the dust).

Keywords: Data mining, Synoptic stations, Dust, Ensemble classification, Central plateau of iran

1. Corresponding Author and Assistant Professor, Department Electrical Engineering, University of Jiroft, Email: mhn@ujiroft.ac.ir