مجله پژوهش آب ایران
جلد ۱۰/ شماره ۳/ پیاپی ۲۲/ پاییز ۱۳۹۵ (۲۱-۲۹)

ارزیابی و مقایسه عملکرد روش‌های داده کاوی در تخمین شاخص کیفی SAR (مطالعه موردی: رودخانه آجی چای آذربایجان شرقی)

علی رضازاده جودی[۱*] و محمدتقی ستاری[۲]

چکیده

آب پاک یکی از عوامل مهم توسعه هر منطقه است. با توجه به قرارگیری ایران در منطقه گرم و خشک و کمبود منابع آب، حفاظت و تأمین کیفیت آب لازم برای مصارف مختلف اهمیتی دو چندان دارد. به طورمعمول ارزیابی کیفی آب‌های سطحی پرهزینه و زمان‌بر بوده و انتخاب روشی که در آن با حداقل پارامترهای هیدروشیمیایی بتوان پیش‌بینی به نسبت دقیقی از کیفیت آب داشت، ترجیح داده می‌شود. یکی از مهم‌ترین پارامترهای کیفی آب در زمینه فعالیت‌های کشاورزی، نسبت جذبی سدیم (SAR) است که تخمین و ارزیابی دقیق مقدار آن بسیار ضروری است. در این بررسی، امکان‌سنجی تخمین شاخص کیفی SAR در رودخانه آجی چای در منطقه آذربایجان شرقی با استفاده از پارامترهای هیدروشیمیایی مختلف با مدل درختی قوانین M5 و ماشین بردار پشتیبان بررسی شد. برای بررسی دقت مدل‌های M5 و ماشین بردار پشتیبان از چهار آماره ضریب همبستگی (R)، نش- ساتکلیف (NSC)، جذر میانگین مربعات خطا (RMSE) و میانگین خطای مطلق مقادیر (MAE) استفاده شد. مقادیر این آماره‌ها برای روش ماشین بردار پشتیبان (R=۰/۹۸، N-SC=۰/۹۷، RMSE=۶/۲۲ (mg/l) و MAE=۶/۰۶(mg/l)) و برای مدل M5(R=۰/۹۸، N-SC=۰/۹۶، RMSE=۷/۳۳ (mg/l) وMAE=۳/۹ (mg/l)) محاسبه شد. نتایج مقایسه نشان داد هر دو روش عملکرد خوبی در تخمین میزان SAR داشته‌اند، اما مدل درختی قوانین M5 در محدوده داده‌های مورد استفاده روابط خطی ساده و کاربردی‌تر ارائه می‌کند.

واژه‌های کلیدی: کیفیت آب، ماشین بردار پشتیبان، مدل درختی قوانین M5، نسبت جذبی سدیم.

ارجاع: رضازاده جودی ع. و ستاری م. ن. ۱۳۹۵. ارزیابی و مقایسه عملکرد روش‌های داده کاوی در تخمین شاخص کیفی SAR (مطالعه موردی: رودخانه آجی چای آذربایجان شرقی). مجله پژوهش آب ایران. ۲۲: ۲۱-۲۹.

۱- کارشناس ارشد عمران آب، باشگاه پژوهشگران جوان و نخبگان، واحد مراغه، دانشگاه آزاد اسلامی، مراغه.
۲- استادیار گروه مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز.
*** نویسنده مسئول: alijoudi66@gmail.com**
تاریخ دریافت: ۱۳۹۲/۰۶/۲۶ تایخ پذیرش: ۱۳۹۲/۱۱/۰۴

مقدمه

کیفیت آب از ویژگی‌های اصلی یک رودخانه است، حتی وقتی که هدف از آن بهره‌برداری برای مصارف و منابع انسانی نباشد. بنابراین ارزیابی کیفیت آب‌های سطحی دارای اهمیت بالایی در مدیریت منابع آب است (دوگان و همکاران، ۲۰۰۹). کیفیت آب‌های سطحی در یک منطقه به‌طور گسترده‌ای به طبیعت، گستردگی صنعت، کشاورزی و دیگر فعالیت‌های انسانی در آن حوضه آبریز بستگی دارد. انسان‌ها آب را از چرخه هیدرولوژیک آن برای نیازهای اقتصادی خود می‌گیرند و پس از استفاده دوباره به آن چرخه باز می‌گردانند (بانژاد و علیائی، ۲۰۱۱). با توجه به این که کشور ایران از نقطه نظر جغرافیایی در منطقه خشک و نیمه‌خشک قرار گرفته و با کم‌آبی روبرو است، از این‌رو شناخت و برآورد دقیق پارامترهای اساسی و تأثیرگذار در کیفیت منابع آب برای دست‌یابی به منابع آب پاک و به دنبال آن تبیین سیاست‌های مدیریتی مناسب امری ضروری و غیر قابل اجتناب به نظر می‌رسد. رودخانه‌ها نیز به‌عنوان مهم‌ترین منابع تأمین و انتقال آب مصرفی بخش‌های صنعت، کشاورزی و شهری دارای اهمیت خاصی بوده و به‌دلیل این که از بسترها و مناطق مختلفی می‌گذرند، ممکن است پساب‌های مختلف شهری، صنعتی و کشاورزی در طول مسیر به آن‌ها تخلیه شود. پس این عوامل می‌توانند سبب نوسانات کیفی زیادی در آن‌ها شوند. از این‌رو، بررسی و پیش‌بینی تغییرات پارامترهای کیفی آب در طول یک رودخانه باید مورد توجه قرار گیرد (جعفرزاده حقیقی و همکاران، ۱۳۸۵). از این جهت تلاش‌های بسیاری برای ارزیابی کیفیت آب انجام شده است که توسعه مدل‌های QUAL2E، WASP و HEC-5Q در زمینه مدیریت بهتر برای حفظ کیفیت آب و پیش‌بینی شاخص‌های کیفی آب‌های سطحی از این جمله هستند. در حالت کلی، با توجه به نوع کاربری آب، استانداردهای مختلفی برای آب تعریف شده و ناظر به آن پارامترهای مختلفی نیز در بررسی شاخص‌های کیفی اهمیت پیدا می‌کنند. از جمله این شاخص‌ها می‌توان به شاخص کیفی آب[1]، شاخص نسبت جذبی سدیم[2]، اکسیژن مورد نیاز بیوشیمیایی[3]، اکسیژن محلول[4]، میزان کل جامدات محلول[5]، هدایت الکتریکی[6] و اکسیژن مورد نیاز شیمیایی[7] اشاره کرد. از جمله مهمترین پارامترهای کیفی آب، شاخص نسبت جذبی سدیم است که برای تعیین مطلوب بودن آب برای آبیاری استفاده می‌شود. به طور کلی، مقادیر بالای این شاخص سبب نامناسب شدن آب آبیاری می‌شود. آبیاری با آبی که دارای میزان بالای این شاخص است، مستلزم اصلاح خاک برای جلوگیری از آسیب‌های دراز مدت به خاک است (میکائیل و همکاران، ۲۰۰۸). به طور کلی، پارامتر SAR با واحد mg/l تابعی از نسبت غلظت سدیم به غلظت کاتیون‌های دو ظرفیتی مانند کلسیم و منیزیم است:

$$SAR = \frac{Na^{+}}{\left[\frac{Ca^{++}+Mg^{++}}{2}\right]^{\frac{1}{2}}} \qquad (1)$$

با توجه به رابطه بالا، یکی از عواملی که بر میزان SAR نهایی تأثیر دارد، تغییر غلظت کلسیم و منیزیم در اثر رسوب یا انحلال کربنات‌های قلیایی است (علیزاده، ۱۳۶۸). با توجه به مشکلات بسیاری که در مرحله جمع‌آوری داده‌های مربوط به پارامترهای مؤثر بر کیفیت آب وجود دارد، عدم وجود امکانات لازم در کلیه ایستگاه‌ها، و لزوم صرفه‌جویی در زمان و هزینه‌ها، استفاده از روش‌های جایگزین نوین می‌تواند راه‌کار مناسبی برای پیش‌بینی کیفیت آب در همه نقاط و در کمترین زمان ممکن باشد. امروزه با پیشرفت علوم در زمینه‌های مختلف و پیدایش روش‌های محاسباتی نرم و ابزارهای نوین داده‌کاوی، از جمله ماشین بردار پشتیبان[8] و مدل درختی قوانین M5، مهندسین تلاش در حل مسایل پیچیده با این روش‌ها دارند. در مهندسی منابع آب نیز استفاده از این روش‌ها در سال‌های اخیر پیشرفت چشم‌گیری داشته و مسایل مختلف و پیچیده با این روش‌ها مدل‌سازی شده و تلاش بر ساده‌سازی شده است. در زمینه پیش‌بینی و ارزیابی کیفیت آب و شاخص‌های مربوط به آن می‌توان به موارد زیر اشاره کرد. علیائی و همکاران (۱۳۸۹) کارایی شبکه عصبی مصنوعی را در پیش‌بینی شاخص‌های کیفی (BOD و DO) آب رودخانه دره مراد بیک همدان، ارزیابی کرده و مشاهده کردند که شبکه عصبی پرسپترون چندلایه، روش کارآمد برای شبیه‌سازی تغییرات این

1- Water Quality Index
2- Sodium Adsorption Ratio
3- Biochemical Oxygen Demand
4- Dissolved Oxygen
5- Total Dissolved Solids
6- Electrical Conductivity
7- Chemical Oxygen Demand
8- Support Vector Machine

شاخص‌ها است. اسدالله‌فردی و همتی (۱۳۹۱) به پیش‌بینی میزان شاخص SAR در رودخانه چهل گزی، با استفاده از شبکه‌های عصبی مصنوعی پرداختند و دریافتند که فقط پارامترهای کلسیم و pH سبب تغییر در پیش‌بینی SAR شده و بقیه پارامترهای ورودی در پیش‌بینی SAR حساسیت مهمی نداشته‌اند و شبکه عصبی مصنوعی با اطمینان خوبی به پیش‌بینی شاخص SAR پرداخته‌اند. سلطانی و همکاران (۱۳۹۱) به ارزیابی و مقایسه عملکرد شبکه عصبی مصنوعی پرسپترون چندلایه و پیشخور تعمیم یافته در شبیه‌سازی شاخص کیفی SAR در رودخانه سیستان با استفاده از پارامترهای کلر، شوری، قلیائیت و کل مواد محلول پرداختند و به این نتیجه رسیدند که این شبکه توانایی زیادی در شبیه‌سازی کیفی شاخص SAR داشته و بهتر از دیگر مدل‌ها جواب می‌دهد. باچاریا و سولوماتین (۲۰۰۵) از شبکه عصبی مصنوعی و مدل درختی قوانین M5 برای مدل سازی رابطه دبی-اشل استفاده و مشاهده کردند که هم شبکه عصبی مصنوعی و هم مدل درختی قوانین M5 از مدل‌های سنتی، مانند منحنی سنجه، عملکرد بهتری دارند. پالانی و همکاران (۲۰۰۸) از شبکه‌های عصبی مصنوعی برای پیش‌بینی ویژگی‌های کیفی آب‌های ساحلی سنگاپور بهره گرفتند. نتایج نشان داد که شبکه عصبی توانایی زیادی در شبیه‌سازی پارامترهای کیفی آب دارد. سینگ و همکاران (۲۰۰۹) مدل شبکه عصبی مصنوعی را برای تخمین میزان اکسیژن محلول و اکسیژن‌خواهی بیوشیمیایی رودخانه گومتی در هند استفاده کردند و مشاهده کردند نتایج مدل هماهنگی خوبی با مقادیر اندازه‌گیری شده و مورد انتظار برای غلظت‌های رودخانه دارد. سارانی و همکاران (۲۰۱۲) عملکرد شبکه عصبی مصنوعی و رگرسیون خطی چندمتغیره را در پیش‌بینی نسبت جذبی سدیم در رودخانه سیستان مقایسه کردند و به این نتیجه رسیدند که شبکه عصبی پرسپترون چندلایه عملکرد بالاتری نسبت به رگرسیون خطی چندمتغیره دارد و می‌تواند به خوبی برای این مسأله به کار رود. قراشی و عبدالله (۲۰۱۲) به پیش‌بینی شاخص کیفیت آب در رودخانه گمبک در مالزی، با استفاده از شبکه عصبی و الگوریتم پس‌انتشار خطا پرداختند و مشاهده کردند که این شبکه توانایی بالایی در این زمینه دارد. نورانی و همکاران (۲۰۱۳) کاربرد شبکه‌های عصبی مصنوعی برای نظارت بر کیفیت آب تصفیه شده را بررسی کردند و دما، pH، کدورت آب، سختی کل و میزان کلسیم قبل از تصفیه به عنوان ورودی مدل درنظر گرفته شد و مقدار کل مواد جامد محلول و هدایت الکتریکی بعد از تصفیه به‌عنوان خروجی مدل درنظر گرفته شدند. نتایج نشان دهنده برتری شبکه عصبی مصنوعی نسبت به سایر مدل‌های کاربردی در پیش‌بینی پارامترهای کیفیت آب بودند. ثاقبیان و همکاران (۲۰۱۳) با روش درختان تصمیم‌گیری به طبقه‌بندی کیفیت آب‌های زیرزمینی در منطقه اردبیل پرداختند و مشاهده کردند که این روش می‌تواند کیفیت آب‌های زیرزمینی را فقط با استفاده از دو پارامتر هدایت الکتریکی و بارش تجمعی، به خوبی طبقه‌بندی کند. گرچه در سال‌های گذشته شبکه‌های عصبی مصنوعی کاربرد وسیعی در پیش‌بینی کیفیت آب داشته‌اند. اما بررسی‌های گسترده‌ای در زمینه امکان‌سنجی کاربرد ماشین بردار پشتیبان و مدل درختی قوانین M5 در این مورد صورت نشده است. ماشین بردار پشتیبان و مدل درختی قوانین M5 یکی از روش‌های نوین داده‌کاوی می‌باشند که دارای قابلیت یادگیری بالایی هستند و می‌توانند مسایل بسیار پیچیده را حل کنند.

هدف از این پژوهش، امکان‌سنجی پیش‌بینی شاخص کیفی SAR در رودخانه آجی چای در آذربایجان شرقی با استفاده از ماشین بردار پشتیبان و مدل درختی قوانین M5 است.

مواد و روش‌ها

منطقه مورد بررسی و داده‌های مورد استفاده

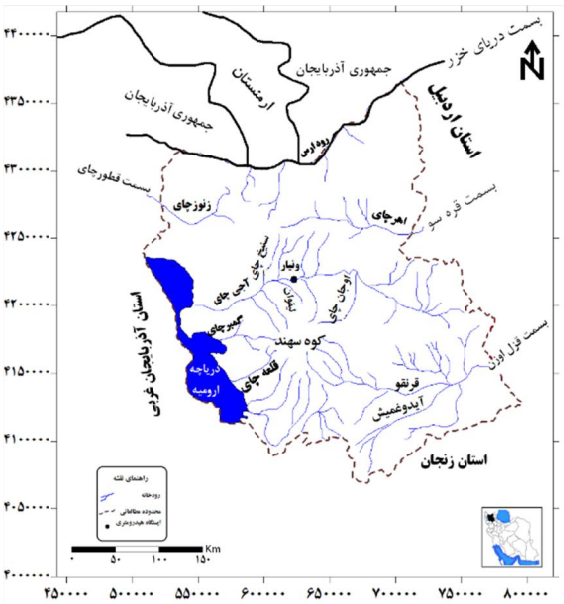
منطقه مورد بررسی رودخانه آجی چای در دامنه‌های شمال کوه سهند است. در این بررسی، برای ارزیابی و پیش‌بینی شاخص کیفی نسبت جذبی سدیم آب رودخانه آجی چای، از داده‌های هیدروشیمیایی ایستگاه هیدرومتری ونیار استفاده شد. موقعیت جغرافیایی ایستگاه ونیار ۴۶ درجه و ۲۴ دقیقه طول شرقی و ۳۸ درجه و ۷ دقیقه عرض شمالی بوده و ارتفاع این ایستگاه از سطح آب‌های آزاد برابر با ۱۴۶۰ متر است. در شکل ۱ موقعیت جغرافیایی منطقه مورد بررسی به همراه رودخانه‌ها و محل ایستگاه هیدرومتری ونیار نشان داده شده است. در این بررسی تأثیرگذاری پارامترهای میزان کل جامدات محلول (TDS)، هدایت الکتریکی (EC)، پ- هاش (pH)، کلر (Cl^-)،

سولفات (SO_4^{2+})، کلسیم (Ca^{2+})، منیزیم (Mg^{2+}) و سدیم (Na^{+}) برای پیش‌بینی میزان نسبت جذبی سدیم به کار رفته‌اند. محدوده تغییرات پارامترهای تأثیرگذار در کیفیت آب و مشخصات آماری آن‌ها در جدول ۱ ارائه شده‌اند.

جدول ۱- محدوده تغییرات و مشخصات آماری داده‌های ایستگاه ونیار

پارامتر	واحد	انحراف معیار	میانگین	حداکثر	حداقل
TDS	mg/l	۱۵۴۸۳/۵۹	۱۳۷۰۰/۴۸	۹۶۲۵۱	۶۲۷
EC	µSiemens/cm	۲۴۴۳۷/۶۳	۲۱۴۲۳/۵۶	۱۴۸۰۰۰	۹۵۴
pH	-	۰/۴۶	۷/۵۷	۹/۲	۶/۱
Cl^-	mg/l	۲۲۲/۲۵	۱۹۲/۳۷	۱۲۵۰	۱/۳۸
SO_4^{2-}	mg/l	۴۷/۶۷	۲۴/۹۵	۶۷۸/۵۵	۰/۱۵
Ca^{2+}	mg/l	۲۱/۳۲	۲۱/۲۸	۱۵۷/۵	۲
Mg^{2+}	mg/l	۱۸/۰۲	۱۴/۳۸	۲۲۵	۰/۳۸
Na^+	mg/l	۲۱/۳۲	۱۸۴/۵۴	۱۱۹۳/۹	۱/۵۶
SAR	mg/l	۲۹/۷۰	۳۶/۸۱	۱۲۵/۵۱	۰/۶۸

این داده‌ها از شرکت آب منطقه‌ای استان آذربایجان شرقی گرفته شدند. در این پژوهش از تعداد ۴۱۲ سری مجموعه داده‌ی ایستگاه ونیار ۶۶٪ آن برابر با ۲۷۲ سری داده برای قسمت آموزش و ۳۴٪ کل داده‌ها برابر با تعداد ۱۴۰ سری داده برای قسمت آزمایش در نظر گرفته شدند. برای بررسی میزان تأثیر پارامترهای مختلف هیدروشیمیایی بر میزان SAR، تعداد ۱۰ سناریو متفاوت شامل ترکیب پارامترهای گوناگون درنظر گرفته شد که این سناریوها در جدول ۲ ارائه شده است.



شکل ۱- موقعیت جغرافیایی محدوده مطالعاتی و ایستگاه هیدرومتری ونیار

جدول ۲- سناریوهای ارائه شده برای بررسی میزان تأثیر پارامترهای مختلف بر میزان SAR

پارامترهای ورودی
Na^+, Mg^{2+}, Ca^{2+}, SO_4^{2+}, Cl^-, pH, EC, TDS
Na^+, Cl^-
EC, TDS
Na^+, Cl^-, EC, TDS
Mg^{2+} , Ca^{2+}, SO_4^{2+}, pH
TDS, Na^+
Cl^-, pH, EC, TDS
Na^+, Mg^{2+}, Cl^-
Na^+, Mg^{2+}, Ca^{2+}, Cl^-
Na^+, Mg^{2+}

عملکرد مدل درختی قوانین M5 و ماشین‌بردار پشتیبان در این پژوهش بر پایه محاسبه ضریب همبستگی[1]، نش-ساتکلیف[2]، جذر میانگین مربعات خطا[3]، و میانگین خطای مطلق[4] ارزیابی شد. فرمول‌های محاسبه آماره‌های بالا در روابط (۲) تا (۵) ارائه شده است. در این رابطه‌ها مقادیر X شامل مقادیر مشاهداتی و مقادیر Y شامل مقادیر محاسباتی هستند.

$$r_{pearson} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (۲)$$

$$E = 1 - \frac{\sum_{i=1}^{n}(X_i-Y_i)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \quad (۳)$$

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n}(y_i-x_i)^2}{N}} \quad (۴)$$

$$MAE = \frac{1}{n}\sum_{i=0}^{n}\left|X_i - Y_i\right| \quad (۵)$$

مدل درختی قوانین M5

مدل درختی قوانین M5 یکی از روش‌های نوین داده‌کاوی است که به تازگی بسیار مورد توجه قرار گرفته است و در مدل‌سازی مسایل مختلف به کار می‌رود. هدف کلی این مدل برگرفته از درختان رگرسیونی است با این تفاوت که به جای مقادیر ثابت و برچسب‌های طبقه‌بندی در برگ‌های خود دارای توابع رگرسیونی است. برتری عمده مدل درختی قوانین M5 نسبت به درختان رگرسیونی این است که مدل درختی قوانین M5 بسیار کوچک‌تر از درختان رگرسیون است و توابع رگرسیون به‌صورت طبیعی شامل بسیاری از متغیرها نمی‌شوند. یک درخت تصمیم به طورمعمول از چهار بخش ریشه[5]، شاخه[6]، گره‌ها[7] و برگ‌ها[8] تشکیل شده است. هر گره مربوط به یک ویژگی مشخص است و شاخه‌ها به معنای بازه‌ای از مقادیر هستند، این بازه‌ها مقادیر معلوم را برای هریک از ویژگی‌ها در نظر می‌گیرند. عمل انشعاب با یکی از متغیرهای پیش‌بینی کننده انجام می‌گیرد، بازه‌های انشعاب طوری انتخاب می‌شوند که مجموع مجذور انحراف از میانگین داده‌های هر گره را حداقل کنند (فلاحی و همکاران، ۱۳۹۰). معیار انشعاب بیانگر میزان خطا در آن گره است و مدل حداقل خطای مورد انتظار را به‌عنوان نتیجه آزمایش هر صفت در آن گره محاسبه می‌کند. خطای مدل به طور عموم با اندازه‌گیری دقت پیش‌بینی مقادیر هدف موارد دیده نشده سنجش می‌شود. فرمول محاسبه کاهش انحراف استاندارد (SDR) به صورت روابط (۶) و (۷) است:

$$SDR = Sd(T) - \sum_{i=1}^{N}\frac{|T_i|}{|T|}Sd(T_i) \quad (۶)$$

$$Sd(T) = \sqrt{\frac{1}{N}(\sum_{i=1}^{N} y_i^2 - \frac{1}{N}(\sum_{i=1}^{N} y_i)^2)} \quad (۷)$$

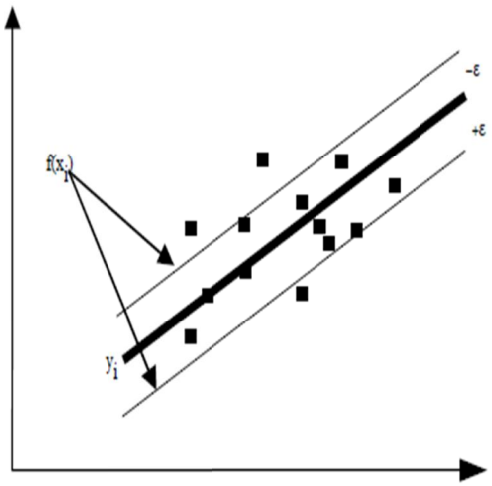
در این رابطه T مجموعه‌ای از نمونه‌ها است (موارد) که به هر گره وارد می‌شود، T_i نشان دهنده زیر مجموعه‌ای از آن می‌باشد، S_d بیانگر انحراف معیار، y_i مقدار عددی ویژگی هدف نمونه i و N شماره داده‌ها را نشان می‌دهد (آلبرگ و همکاران، ۲۰۱۲). فرآیند انشعاب در هر گره بارها تکرار می‌شود تا به گره پایانی (برگ) برسد که در برگ، مجموع مجذور انحراف از میانگین داده‌ها به طورتقریبی به صفر می‌رسد. با این کار درخت بزرگی توسعه پیدا خواهد کرد. کار با این درخت بزرگ که شاخه‌ها و گره‌های زیادی دارد سخت است، بنابراین برای رسیدن به یک درخت بهینه و کارآمد باید شاخه‌های اضافی درخت هرس شود. دو روش برای هرس کردن درخت[9] وجود دارد: ۱) هرس قبل از شکل‌گیری درخت حداکثر[10] و ۲) هرس بعد از شکل‌گیری درخت حداکثر[11]. در روش اول فرآیند هرس اجازه نمی‌دهد شاخه‌های اضافی تولید شوند، ولی در روش دوم ابتدا درخت حداکثر تشکیل می‌شود، سپس فرآیند هرس انجام می‌گیرد (برای اطلاعات بیشتر به کوییلن (۱۹۹۲) مراجعه شود). در این پژوهش برای مدل‌سازی روش M5 از نرم‌افزار WEKA[12] که در دانشگاه Waikato نیوزلند توسعه داده شده، استفاده شده است. مدل‌سازی مدل درختی قوانین M5 با استفاده از گزینه M5 Rules در این نرم‌افزار انجام شده است که قوانین ساده و خطی اگر آنگاه ارائه می‌کند.

1 - Correlation coefficient
2- Nash-Sutcliffe coefficient (E)
3- Root Mean Square Error
4- Mean Absolute Error
5- Root
6- Beach
7- Nodes
8- Leafs
9- Tree Pruning
10- Pre-Pruning
11- Past-Pruning
12- Waikato Environment for Knowledge Analysis

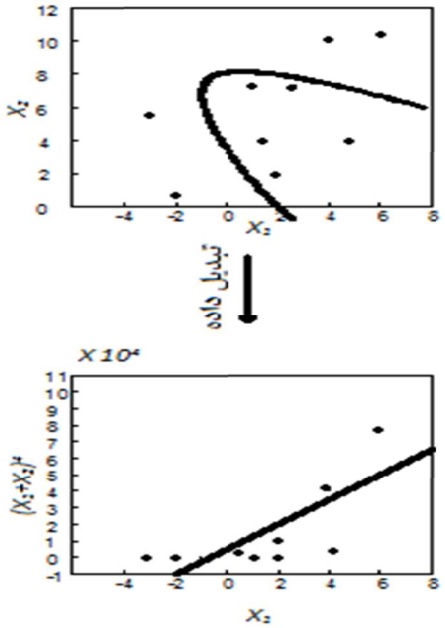
ماشین بردار پشتیبان

ماشین‌های بردار پشتیبان نیز همانند مدل درختی M5 و شبکه عصبی مصنوعی یک الگوریتم داده‌کاوی است. ماشین‌های بردار پشتیبان شامل دو دسته هستند: طبقه‌بندی کننده بردار پشتیبانی[1] و رگرسیون بردار پشتیبانی[2]. ماشین‌های بردار پشتیبان بر پایه مفهوم صفحات تصمیم هستند که مرز تصمیم را تعریف می‌کنند، که یک صفحه تصمیم، داده‌های با برچسب‌های مختلف را از هم جدا می‌کند (شهرابی و حجازی، ۱۳۹۰). در یک الگوریتم خطی‌سازی به کمک ماشین بردار پشتیبان با فرض مقادیر ورودی x_i و مقادیر خروجی y_i، هدف یافتن تابعی است که کمترین انحراف (ε) را از y_i ها داشته باشد، که در آن ε میزان انحراف است (واپنیک، ۱۹۹۵). در شکل ۲ که شمایی از خطی‌سازی به کمک الگوریتم ماشین‌بردار پشتیبان با در نظر گرفتن انحراف به مقدار ε نشان داده شده است.

در برخی مواقع، به ساختارهای پیچیده و غیرخطی برای جداسازی داده‌ها نیاز است. در این صورت، ماشین بردار پشتیبان، داده‌های اصلی را با به کارگیری مجموعه‌ای از توابع ریاضی که کرنل نام دارند در فضای جدیدی نگاشت و بازآرایی می‌کند که به این کار تبدیل (نگاشت) گفته می‌شود (شهرابی و حجازی، ۱۳۹۰). شکل ۳ نمونه‌ای از نگاشت داده را با یک تابع کرنل نشان می‌دهد.



شکل ۲- شمایی از خطی‌سازی به کمک الگوریتم ماشین بردار پشتیبان با درنظر گرفتن انحراف ε (هاشمی و همکاران، ۱۳۹۱)



شکل ۳- نمونه‌ای از نگاشت داده با ماشین بردار پشتیبان (شهرابی و حجازی، ۱۳۹۰)

اگر داده‌های آموزش به‌صورت k سری به‌صورت $(x_1,y_1),\ldots,(x_k,y_k)$ باشند، تابع خطی‌سازی مربوطه به صورت رابطه (۸) قابل بیان است:

$$f(x) = (w, x1) + b \text{with}(w, x) \in R^N, b \in R \qquad (۸)$$

که در آن x بردار ورودی، w وزن بردار و b میزان اختلال است. تابع $f(x)$ همواره باید به گونه‌ای تعیین شود که به طور همزمان میزان کمترین انحراف ε تعیین شده و همچنین تابع، مقدار مناسب w را اختیار کند. ایجاد چنین شرایطی با حل معادله بهینه‌سازی زیر قابل دست‌یابی است:

$$\text{Minimise}\frac{1}{2}\|w\|^2 \text{ Subjecto} \begin{cases} y_i - (w, x_i) - b \le \varepsilon \\ (w, x_i) + b - y_i \le \varepsilon \end{cases} \qquad (۹)$$

که در آن w، b، x_i و y_i همان پارامترهای معادله (۸) بوده و ε نیز میزان انحراف اعمال شده است. در نهایت برای تبدیل معادله بهینه‌سازی دارای یک مجموعه قیود نامعادلات و نامساوی‌ها به یک معادله صریح با درنظر گرفتن پارامترهای لاگرانژ λ_i, λ'_i معادله (۱۰) به دست خواهدآمد.

$$f(x, b) = \sum_{i=1}^{k} (\lambda'_i - \lambda_i)(x_i, x) + b \qquad (۱۰)$$

حال معادله خطی‌سازی بالا، با به کارگیری توابع کرنل، برای مسایل رگرسیون غیرخطی نیز قابل استفاده خواهد

1- Support Vector Classification (SVC)
2- Support Vector Regression (SVR)

شد (واپنیک، ۱۹۹۵). پس درنهایت معادله تابع رگرسیون، به شکل رابطه (۱۱) بازنویسی خواهد شد:

$$f(x,b) = \sum_{i=1}^{k} (\lambda'_i - \lambda_i) k(x_i, x) + b \qquad (۱۱)$$

که در رابطه (۱۱)، $k(x_i,x)$ نمایان‌گر تابع کرنل است. از انواع توابع کرنل می توان به توابع خطی[1]، چندجمله‌ای[2]، تابع شعاع محور[3] و سیگموئید[4] اشاره داشت. در این پژوهش برای مدل سازی میزان نسبت جذبی سدیم با رگرسیون بردار پشتیبان، از نرم‌افزار Statistica استفاده شده است.

نتایج و بحث

در مدل‌سازی میزان نسبت جذبی سدیم با مدل درختی قوانین M5 بهترین جواب در حالتی که ۶۶ درصد داده‌ها به آموزش و بقیه به آزمون اختصاص داده شده‌اند، به دست آمد. برای مدل‌سازی میزان نسبت جذبی سدیم با ماشین بردار پشتیبان، پس از آزمون توابع مختلف به‌عنوان تابع کرنل، مشخص شد که تابع RBF بهترین عملکرد را در مدل‌سازی میزان نسبت جذبی سدیم از خود نشان می‌دهد. از میان تعداد ۱۰ سناریو بررسی شده در این پژوهش، بهترین سناریو انتخاب شد، که این سناریو همراه با مقادیر نتایج آماری آن در جدول ۳ ارائه شده است.

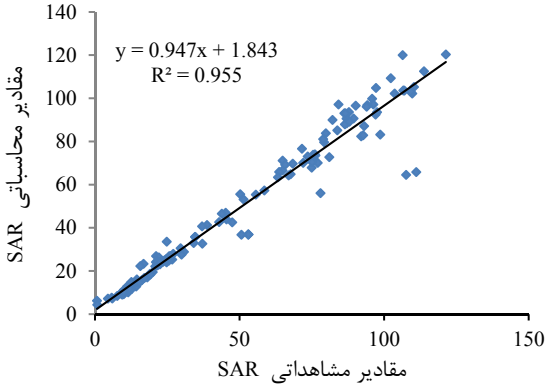
جدول ۳- مقادیر آماره های بررسی شده به ازای بهترین سناریو برای مدل درختی قوانین M5 و ماشین بردار پشتیبان

روش	M5	SVR
پارامترهای ورودی	Na^+, Mg^{2+}	Na^+, Mg^{2+}, Ca^{2+}, Cl^-
سناریو	۱۰	۹
R	۰/۹۸	۰/۹۸
N-S	۰/۹۶	۰/۹۷
RMSE	۷/۳۳	۶/۲۲
MAE	۳/۹	۶/۰۶
تعداد قوانین/بردار	۴	۲۰

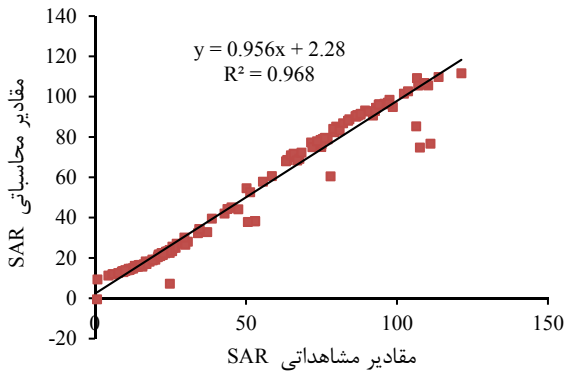
همان‌گونه که مشاهده می‌شود هر دو مدل به خوبی توانسته‌اند میزان نسبت جذبی سدیم را تخمین بزنند.

1- Linear
2- Polynomial
3- Radial Basis Function (RBF)
4- Sigmoid

پراکنش مقادیر مشاهداتی نسبت به مقادیر محاسباتی به دست آمده از مدل درختی قوانین M5، ماشین بردار پشتیبان به‌ترتیب در شکل‌های ۴ و ۵ دیده می‌شود.



شکل ۴- پراکنش مقادیر مشاهداتی و محاسباتی به دست آمده از مدل درختی M5 برای سناریوی ۱۰



شکل ۵- پراکنش مقادیر مشاهداتی و محاسباتی به دست آمده از رگرسیون بردار پشتیبان برای سناریوی ۹

روابط خطی ارائه شده توسط مدل درختی قوانین M5 در جدول ۴ قابل مشاهده می‌کند. همچنان که در جدول ۴ دیده می‌شود مدل درختی قوانین M5 با در اختیار گذاشتن چهار رابطه خطی صریح و ساده اقدام به تخمین میزان نسبت جذبی سدیم کرده است. به‌عنوان مثال در جدول ۴، قانون اول بیان می‌کند اگر میزان سدیم بزرگ‌تر از ۱۳۱ باشد، آنگاه میزان نسبت جذبی سدیم از رابطه (۱۲) قابل محاسبه است.

$$SAR = 0.1319 * Na - 0.437 * Mg + 27.2303 \qquad (۱۲)$$

همچنین اگر میزان سدیم ۴۷/۵ باشد آنگاه میزان نسبت جذبی سدیم از رابطه ناظر به قانون دوم قابل محاسبه است. این روند تا آخرین قانون یعنی قانون ۴ ادامه می‌یابد. همچنان که مشاهده می‌شود نتایج ارائه شده با هر

دو روش قابل قبول هستند، اما با توجه به روابط خطی صریح و روشن ارائه شده با مدل درختی قوانین M5، استفاده از این مدل به مهندسین منابع آب این امکان را فراهم می‌کند تا به سادگی با استفاده از این روابط و با در اختیار داشتن مقدار سدیم و منیزیم نسبت به تخمین مقدار جذبی سدیم اقدام کنند. در صورتی که ماشین بردار پشتیبان بهترین مدل‌سازی خود را با استفاده از چهار پارامتر هیدروشیمیایی سدیم، منیزیم، کلسیم و کلر ارائه کرده است.

جدول ۴- روابط خطی ارائه شده توسط مدل درختی قوانین M5

Rule: 1 IFNa > 131 THEN SAR= 0.1319 * Na - 0.437 * Mg + 27.2303
Rule: 2 IFNa > 47.5THEN SAR= 0.2451* Na-0.6183 * Mg+10.8804
Rule: 3 IFNa > 21.05 THEN SAR= 0.3238 * Na- 0.608 * Mg + 6.0023
Rule: 4 SAR = 0.5249 * Na - 0.5078 * Mg + 1.6935

نتیجه‌گیری

با توجه به قرارگیری ایران در منطقه گرم و خشک و کمبود منابع آبی، حفاظت و تأمین کیفیت آب لازم برای مصارف مختلف، اهمیتی دو چندان می‌یابد. به طور معمول ارزیابی کیفی آب‌های سطحی پرهزینه و زمان‌بر بوده و انتخاب روشی که در آن بتوان با حداقل پارامترهای هیدروشیمیایی، تخمین به نسبت دقیقی از کیفیت آب داشت، ترجیح داده می‌شود. در این پژوهش شاخص کیفی SAR در روش رگرسیون بردار پشتیبان، با استفاده از پارامترهای هیدروشیمیایی سدیم، منیزیم، کلسیم و کلر (سناریو ۹)، و در روش مدل درختی قوانین M5 با استفاده از پارامترهای سدیم و منیزیم (سناریو ۱۰) ارزیابی شد. در پژوهش مشابهی اسدالله‌فردی و همتی (۱۳۹۱) با استفاده از روش شبکه عصبی مصنوعی مقدار SAR در رودخانه چهل‌گزی را پیش‌بینی کرده و مقادیر ضریب همبستگی را برابر با ۰/۹۷ و ریشه میانگین مربعات خطا را برابر با ۰/۰۱۸ (mg/l) به دست آوردند. همچنین سلطانی و همکاران (۱۳۹۱)، نیز با روش شبکه عصبی مصتوعی مقدار SAR را در رودخانه سیستان شبیه‌سازی کرده و مقادیر ضریب همبستگی را برابر با ۰/۹۲ و ریشه میانگین مربعات خطا را برابر با ۱/۰۵ (mg/l) به دست آوردند. مقایسه نتایج به‌دست آمده از مطالعات بالا و این بررسی نشان می‌دهد هر دو روش بررسی شده رگرسیون بردار پشتیبان و مدل M5 دارای توانمندی زیادی در پیش‌بینی مقدار SAR در رودخانه آجی‌چای و در محدوده داده‌های مورد استفاده هستند ولی با توجه به ارائه روابط خطی ساده و کاربردی با مدل M5، استفاده از این مدل توصیه می‌شود.

منابع

۱. اسدالله‌فردی غ. و همتی آ. ۱۳۹۱. پیش‌بینی SAR با استفاده از شبکه‌های عصبی مصنوعی (مطالعه موردی: رودخانه چهل‌گزی). کنفرانس بین‌المللی جریان و آلودگی آب. دانشگاه تهران.

۲. جعفرزاده حقیقی ن. کعبی ه. نبی‌زاده ر. و سپهرفر ک. ۱۳۸۵. امکان‌سنجی کاربرد و انتخاب مناسب ترین شاخص کیفیت آب رودخانه (مطالعه موردی: رودخانه زهره). هفتمین سمینار بین‌المللی مهندسی رودخانه. دانشگاه شهید چمران اهواز.

۳. سلطانی ج. سارانی ن. و معاشری س. ع. ۱۳۹۱. ارزیابی عملکرد شبکه‌های عصبی مصنوعی MLP و GFF در شبیه‌سازی شاخص کیفی SAR (مطالعه موردی: رودخانه سیستان). کنفرانس بین‌المللی جریان و آلودگی آب. دانشگاه تهران.

۴. شهرابی ج. و حجازی ط. ح. ۱۳۹۰. داده‌کاوی. انتشارات جهاد دانشگاهی واحد صنعتی امیرکبیر. ۱۳۱ ص.

۵. علیائی ا. بانژاد ح. صمدی م. ت. رحمانی ع. و ساقی م. ح. ۱۳۸۹. ارزیابی کارایی شبکه عصبی مصنوعی در پیش‌بینی شاخص‌های کیفی (DO و BOD) آب رودخانه دره مرادبیک همدان. مجله دانش آب و خاک. ۲۰/۱(۳): ۱۹۹-۲۱۰.

۶. علیزاده ا. ۱۳۶۸. کیفیت آب در آبیاری. انتشارات آستان قدس رضوی. ۹۳ ص.

۷. فلاحی م. ر. و روانی ه. و گلیان س. ۱۳۹۰. پیش‌بینی بارش با استفاده از مدل رگرسیون درختی به منظور کنترل سیل. پنجمین کنفرانس

سراسری آبخیزداری و مدیریت منابع آب و خاک کشور. کرمان.

8. Alberg D. Last M. and Kandel A. 2012. Knowledge discovery in data streams with regression tree methods. WIREs Data Mining Knowl Discov 2: 69-78.
9. Banejad H. and Olyaie E. 2011. Application of an Artificial Neural Network Model to Rivers Water Quality Indexes Prediction-A Case Study. Journal of American Science. 7(1): 60-65.
10. Bhattacharya B. Solomatine D. P. 2005. Neural networks and M5 model trees in modelling water level–discharge relationship. Neurocomputing. 63: 381-396.
11. Dogan E. Sengorur B. and Koklu R. 2009. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. Journal of Environmental Management. 90: 1229-1235.
12. Gorashi F. and Abdollah A. 2012. Prediction of Water Quality Index Using Back Propagation Network Algorithm. (Case Study: Gombak River, Malaysia). Journal of Engineering Science and Technology. 7(4): 447– 461.
13. Michael A. M. Kherpar S. D. and Sondhi S. D. 2008. Water wells and pumps. McGraw-Hill. New Delhi.
14. Nourani V. Khanghah T. and Sayyadi M. 2013. Application of the Artificial Neural Network to monitor the quality of treated water. International Journal of Management & Information Technology. 3(1):38-45.
15. Palani S. Liong S. and Tkalich P. 2008. An ANN Application for Water Quality forecasting. Marine Pollution Bulletin. 56:1586-1597.
16. Quinlan J. R. 1992. Learning with continuous classes. Singapore. In proceedings AI, 92 (Adams & Sterling, Eds). World Scientific. 343-348.
17. Saghebian S. M. Sattari M. T. Mirabbasi R. and Pal M. 2013. Ground water quality classification by decision tree method in Ardebil region, Iran. Arabian Journal of Geosciences. DOI: 10.1007/s12517-013-1042-y.
18. Sarani N. Soltani J. Sarani S. and Moasheri A. 2012. Comparison of Artificial Neural Network and Multivariate Linear Regression Model to Predict Sodium adsorption ratio (SAR) (Case Study: Sistan River, Iran). International Conference on Chemical. Ecology and Environmental Sciences. 130-134.
19. Singh K.P. Basant A. Malik A. and Jain G. 2009. Artificial neural network modeling of the river water quality—A case study. Journal of Ecological Modelling. 220(6): 888-895.
20. Vapnik V.N. 1995. The nature of statistical learning theory. Springer. New York. 313 p.