نشریه حفاظت منابع آب و خاک، سال هفتم، شماره چهارم، تابستان ۱۳۹۷

ISSN 2251-7480

مدل‌سازی رواناب ماهانه با استفاده از روش‌های داده‌کاوی براساس الگوریتم‌های انتخاب ویژگی

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

۱) استادیار گروه مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران

۲) باشگاه پژوهشگران جوان و نخبگان، واحد مراغه، دانشگاه آزاد اسلامی، مراغه، ایران

*** نویسنده مسئول مکاتبات:** alijoudi66@gmail.com

تاریخ دریافت: ۹۶/۱۱/۰۵ تاریخ پذیرش: ۹۷/۰۲/۱۷

چکیده

باتوجه به اهمیت مقدار جریان خروجی از حوضه آبریز جهت مدیریت آب‌های سطحی، فهم دقیق ارتباط بین مقدار رواناب با پارامترهای اقلیمی همچون بارش و دما و شناسائی موثرترین پارامتر در فرآیند مدل‌سازی بسیار مهم می‌باشد. در این تحقیق پس از آزمون همگنی داده‌های بارش، دما و رواناب ماهانه حوضه آبریز ناورود، ابتدا براساس دو الگوریتم رلیف و همبستگی دو ترکیب مختلف از پارامترهای موثر در رواناب مورد توجه قرار گرفت. الگوریتم جدید رلیف با استفاده از میانگین بردار وزنی مرتبط بین داده‌ها و یک مقدار آستانه، ویژگی‌های موثر در بین یک مجموعه از داده‌ها را به ویژه در شرایطی که تعداد داده‌ها کم باشد، به ترتیب اهمیت شناسایی می‌کند. سپس با استفاده از دو روش رگرسیون بردار پشتیبان و نزدیک‌ترین همسایگی رواناب ماهانه مبتنی بر دو ترکیب ورودی پیشنهادی مدل‌سازی گردید. نتایج به دست آمده نشان داد، روش رگرسیون بردار پشتیبان با بهره‌گیری از تابع کرنل شعاع محور نسبت به روش نزدیک‌ترین همسایگی از دقت بالا و خطای کمتری در برآورد رواناب به خصوص در مقادیر جریان‌های سیلابی برخوردار است.

کلید واژه‌ها: مدل‌سازی رواناب؛ داده‌کاوی؛ رگرسیون بردارپشتیبان؛ نزدیک‌ترین همسایگی؛ الگوریتم رلیف؛ حوضه ناورود

مقدمه

بارش-رواناب یکی از پیچیده‌ترین فرآیندهای هیدرولوژیکی است. از آنجا که با وجود افزایش روز افزون تقاضای آب، سالانه حجم زیادی از رواناب حوضه‌های آبریز مناطق کشورمان به دلیل عدم کنترل، از دسترس خارج شده و هدر می‌رود، اهمیت و حساسیت مهار آب‌های سطحی جهت سیاست‌ها و برنامه‌ریزی‌های تأمین آب، بیش از پیش ضرورت می‌یابد. بدیهی است نخستین گام در استحصال این منابع و استفاده بهتر از پتانسیل موجود در حوضه‌های آبریز، برآورد دقیق جریان رواناب حوضه‌ها می‌باشد. اما یکی از مشکلات اساسی در این زمینه کمبود آمار و اطلاعات آبدهی رودخانه‌ها بدلیل عدم وجود ایستگاه‌های آب سنجی در خروجی حوضه‌ها است (ملکیان و همکاران، ۱۳۸۳). در اندازه‌گیری مستقیم جریان در ایستگاه‌های آب‌سنجی، مشکلاتی از قبیل کاهش یا افزایش سریع سرعت جریان، حمل مقدار زیاد رسوب، بارش‌های رگباری و غیره وجود دارد (رضایی و همکاران، ۱۳۸۶). اولین مطالعات صورت گرفته در زمینه برآورد مستقیم رواناب از رگبار، در قالب اشکال ترسیمی بود که محققین با افزایش پارامترهای تاثیرگذار در رواناب، سعی

در بهبود دقت نمودارهای ترسیمی داشتند (Kohler & Linsly, 1951). سازمان حفاظت خاک آمریکا[1] (SCS) نیز به منظور برآورد رواناب در سال ۱۹۷۲ رابطه جامع شماره منحنی[2] (SCS-CN) را ارائه داد (Dingman, 1994). لیکن این روش که روش عمومی برآورد رواناب از بارش است در نواحی کوهستانی با رژیم برفی دارای خطا می‌باشد (ولی خوجینی، ۱۳۷۷). مدل‌های تجربی دارای محدودیت های زیادی بوده و کاربرد آنها مستلزم تصحیح ضرایب هر رابطه تجربی با صرف زمان و هزینه برای هر منطقه است. در سال‌های اخیر محققان زیادی در سراسر جهان سعی کرده‌اند از روش‌های نوین داده‌کاوی[3] از جمله شبکه‌های عصبی مصنوعی[4]، سیستم استنتاج عصبی- فازی تطبیقی[5]، برنامه‌ریزی بیان ژن[6]، ماشین بردار پشتیبان[7]، مدل‌های درختی[8] و نزدیک‌ترین همسایگی[9] برای مدل‌سازی دقیق مقدار رواناب سطحی استفاده کنند که از جمله این تحقیقات می‌توان به موارد زیر اشاره کرد.

Eskandarinia و همکاران (۲۰۱۰) با استفاده از شبکه عصبی مصنوعی و نزدیک‌ترین همسایگی به پیش‌بینی جریان روزانه رودخانه بختیاری پرداختند. نتایج نشان داد روش شبکه عصبی مصنوعی برتری اندکی نسبت به روش نزدیک‌ترین همسایگی دارد. در این مطالعه از داده‌های بارش و رواناب رودخانه بختیاری به مدت ۲۱ سال استفاده گردیده که تعداد پارامتر ورودی کم و بازه نسبتاً مناسب داده‌های در دسترس از نقاط قوت این مطالعه و پیش‌بینی جریان در مقیاس نسبتاً کوچک روزانه از نقاط ضعف این مطالعه می‌باشد. Okkan و Serbes (۲۰۱۲) با استفاده از روش حداقل مربعات ماشین بردار پشتیبان پدیده بارش- رواناب را مدل‌سازی کردند. نتایج نشان داد مدل حداقل مربعات ماشین بردار پشتیبان با توجه به بکارگیری از توابع نرمال کردن متعدد سرعت اجرا و دقت به مراتب بالاتری نسبت به شبکه عصبی مصنوعی دارد. Sattari و همکاران (۲۰۱۲) با استفاده از شبکه‌های عصبی مصنوعی به پیش بینی جریان رودخانه سوهو در ترکیه پرداخته و کارایی بالای این روش را گزارش کردند. آنها در این مطالعه از داده‌های بارش، رواناب و حرارت مربوط به ۵ سال را برای پیش بینی جریان روزانه رودخانه سوهو استفاده نموده و توانایی بالای شبکه عصبی مصنوعی را در این طمینه گزارش نمودند. هرچند مدل‌سازی جریان رودخانه در مقیاس روزانه نیز مزیت‌ها و کاربردهای خود را دارد اما به نظر می‌رسد پیش‌بینی رواناب رودخانه در مقیاس‌های ماهانه و سالانه کاربردی‌تر باشد. Sattari و همکاران (۲۰۱۳) با استفاده از مدل درختی M5 و ماشین بردار پشتیبان به پیش‌بینی جریان رودخانه سوهو در ترکیه پرداختند. نتایج نشان داد مدل درختی M5 به خوبی ماشین بردار پشتیبان قادر به مدل‌سازی رواناب می‌باشد و با توجه به ارائه مدل‌های خطی ساده توسط مدل درختی M5 زمان محاسباتی کمتری نسبت به ماشین بردار پشتیبان لازم داشته و کاربرد آن توصیه می‌شود. Vafakhah و همکاران (۲۰۱٤) با استفاده از روش‌های ARX، ARMAX، شبکه عصبی مصنوعی، سیستم استنتاج عصبی- فازی تطبیقی و شبکه عصبی مصنوعی-موجکی به مدل‌سازی بارش- رواناب پرداختند. علی رغم استفاده از داده‌های بارش و رواناب به مدت ۲٦ سال و به کارگیری انواع روش‌های پیش پردازش داده‌ها در این تحقیق نتایج در اغلب مدلها خیلی دقیق نبوده و فقط روش سیستم استنتاج عصبی- فازی تطبیقی نتایج بسیار خوبی ارائه کرده است. Kumar Vyas و همکاران (۲۰۱٦) با استفاده از رگرسیون و شبکه عصبی مصنوعی به مدل‌سازی بارش رواناب در رودخانه بناس در هند پرداختند. نتایج نشان‌دهنده برتری روش شبکه عصبی مصنوعی نسبت به رگرسیون در پیش‌بینی

1 - Soil conservation number
2 - Curve number
3 - Data Mining
4 - Artificial Neural Network
5 - Adaptive Neuro-Fuzzy Inference System
6 - Gene Expression Programming
7 - Support Vector Machine
8 - Model Trees
9 - Nearest Neighbours

مقادیر رواناب است. آنها در این تحقیق مقادیر رواناب را فقط در ماه‌های مربوط به باران‌های موسمی مدل‌سازی کرده و گزارشی از عملکرد مدل‌های به کار رفته در سایر ماه‌ها ارائه نکردند. Patel و Joshi (۲۰۱۷) با استفاده از روش شبکه عصبی مصنوعی به مدل‌سازی ضرایب بارش رواناب در حوضه آبخیز داروی در هند پرداختند. با بررسی نتایج مشاهده گردید مدل شبکه عصبی خطای بسیار پایینی داشته و برای شبیه‌سازی ضرایب بارش رواناب در این حوضه پیشنهاد می‌شود. آنها در تحقیق خود از داده‌های بارش و رواناب به مدت ۲۹ سال استفاده نمودند که منجر به دستیابی به نتایج بسیار دقیق از جمله مقدار ۰/۹۹ برای ضریب همبستگی در این مطالعه شده است. از جمله تحقیقات صورت گرفته در این زمینه در ایران نیز می‌توان به موارد زیر اشاره نمود. نگارش و همکاران (۱۳۹۱) با استفاده از روش‌های آماری همچون رگرسیون چند متغیره به مدل‌سازی تولید رواناب حوضه آبریز رودخانه کشکان پرداختند. آنها بدین منظور از داده‌های بارش و دبی در مقیاس ماهانه، مساحت، زمان تمرکز، ضریب فشردگی و حداکثر ارتفاع استفاده نمودند. یکی از فاکتورهای مهم در مدل‌سازی پدیده‌های هیدرولوژیکی استفاده از تعداد پارامترها و متغیرهای ورودی کم می‌باشد که تعداد ٦ پارامتر ورودی در این مطالعه کمی زیاد به نظر می‌رسد. غلامی و درواری (۱۳۹۲) با استفاده از شبکه عصبی مصنوعی و مدل HEC-HMS فرآیند بارش-رواناب را در حوضه آبریز کسیلیان شبیه‌سازی نمودند. آنها از عامل هدر رفت اولیه خاک شامل سه فاکتور خاک، پوشش گیاهی و رطوبت پیشین خاک و همچنین میزان بارش به‌عنوان ورودی برای شبیه‌سازی مقدار دبی یا رواناب استفاده نمودند. در این مطالعه علی‌رغم استفاده از پارامترهای متعدد مشاهده می‌گردد ضریب همبستگی در بهترین حالت برابر با ۰/٦۳ به‌دست آمده است که مقدار خیلی مناسبی به نظر نمی‌رسد. خزایی و همکاران (۱۳۹۳) با استفاده از مدل‌های شبکه عصبی مصنوعی پرسپترون چندلایه و شبکه شعاع محور به مدل‌سازی بارش-رواناب در مقیاس‌های زمانی مختلف شامل ماهانه، فصلی و سالانه پرداختند. بررسی نتایج نشان داد در حالت کلی هر دو مدل شبکه عصبی مصنوعی استفاده شده در مقیاس سالانه نتایج بهتری ارائه داده و نتایج ارائه شده این مدل‌ها در مقیاس ماهانه چندان دقیق نمی‌باشد. سیدیان و همکاران (۱۳۹۳) با استفاده از روش‌های ماشین بردار پشتیبان و سری زمانی به پیش بینی دبی جریان رودخانه گرگانرود پرداختند و برتری روش ماشین بردار پشتیبان را نسبت به سری زمانی اعلام کردند عبداله‌پورآزاد و ستاری (۱۳۹٤) با استفاده از دو مدل سیستم استنتاج عصبی- فازی تطبیقی و شبکه عصبی مصنوعی جریان روزانه رودخانه اهرچای را برای یک روز بعد پیش‌بینی نمودند. نتایج بدست آمده نشان داد مدل سیستم استنتاج عصبی- فازی تطبیقی از دقت نسبتا بالاتری برخوردار است. در این مطالعه نیز مشاهده می‌شود که جریان رودخانه در بازه روزانه مدل‌سازی شده است که تحقیق حاضر که در بازه ماهانه صورت گرفته کاربردی‌تر به نظر می‌رسد. عظیمی و همکاران (۱۳۹٤) با استفاده از برنامه‌ریزی بیان ژن و مدل درختی M5 به برآورد دبی‌های روزانه رودخانه لیقوان پرداختند. نتایج نشان‌دهندهٔ کارایی قابل قبول این روش‌ها در زمینه مدل‌سازی جریان روزانه رودخانه است. آنها در این تحقیق جریان روزانه رودخانه را مدل‌سازی نمودند که با توجه به تغییرات کمتر جریان در بازه زمانی کوچکتر دقت بالای این مدل‌ها قابل پیش‌بینی بود. جودی حمزه آباد و همکاران (۱۳۹۵) به ارزیابی کارایی مدل‌های هیدرولوژیک ارزیابی آب و خاک[1] و ماشین بردار پشتیبان در شبیه‌سازی رواناب رودخانه لیقوان چای پرداختند. نتایج نشان داد علی‌رغم توانایی هر دو مدل در شبیه‌سازی رواناب رودخانه لیقوان چای مدل هیدرولوژیک ارزیابی آب و خاک دقت بیشتر و خطای کمتری نسبت به مدل ماشین بردار پیشتیبان دارد. دلیل این امر استفاده از تعداد زیادی متغیر ورودی شامل بارش

1 - SWAT

روزانه، درجه حرارت، تشعشع خورشیدی، پوشش گیاهی، کاربری اراضی، کاربرد کود، نقشه خاک، نقشه توپوگرافی و شبکه حریان در این مدل هیدرولوژیک می‌باشد. این در حالیست که دستیابی به دقیقترین نتیجه ممکن با کمترین تعداد پارامتر ورودی در اینگونه مطالعات دارای اهمیت فراوانی است. بررسی منابع نشان‌دهنده این است که علی‌رغم حجم بالای مطالعات در زمینه کاربرد شبکه‌های عصبی مصنوعی در زمینه مدل‌سازی بارش-رواناب مطالعات چندانی در زمینه کاربرد توام روش‌هایی همچون رگرسیون بردار پشتیبان و نزدیک‌ترین همسایگی برای مدل‌سازی رواناب به خصوص در ایران صورت نگرفته است.

با توجه به موارد ذکر شده هدف از مطالعه حاضر امکان سنجی کاربرد روش‌های رگرسیون بردار پشتیبان و نزدیک‌ترین همسایگی برای مدل‌سازی رواناب ماهانه در حوضه آبریز ناورود واقع در منطقه پر بارش گیلان و معرفی بهترین روش می‌باشد. همچنین بررسی مطالعات پیشین نشان داد تاکنون مطالعه‌ای در زمینهٔ تعیین پارامترهای موثر در مدل‌سازی بارش-رواناب و سایر فرآیندهای هیدرولوژیکی توسط الگوریتم رلیف[1] صورت نگرفته است، لذا در این مطالعه سعی خواهد شد تا کارایی الگوریتم رلیف در مقایسه با روش همبستگی برای تعیین پارامترهای موثر در مدل‌سازی بارش-رواناب نیز سنجیده شود.

مواد و روش‌ها

منطقه مورد مطالعه و داده‌های مورد استفاده

حوضه ناورود یکی از حوضه‌های آبریز مهم غرب استان گیلان در محدوده شهرستان تالش بین طول جغرافیایی ٤٨ درجه و ٣٥ دقیقه تا ٤٨ درجه و ٥٤ دقیقه شرقی و عرض جغرافیایی ٣٧ درجه و ٣٦ دقیقه تا ٢٧ درجه و ٤٥ دقیقه شمالی قرار گرفته است. ناورود یک حوضه جنگلی-کوهستانی با مساحتی معادل ٢٧٤ کیلومتر مربع، طول آبراهه اصلی ٣٢/٥ کیلومتر و دارای شیب متوسط ٣١/١٦ درصد است. ارتفاع حوضه در نقطه خروجی معادل ١٣٠ متر و در مرتفع‌ترین نقطه ٣٠١٦ متر است. از لحاظ ساختار زمین‌شناسی بیش از ٨٥ درصد از سطح حوضه را سنگ‌های آذرآواری و توفی همراه با سنگ‌های آتشفشانی و آهک ناخالص تشکیل می‌دهند که دارای نفوذپذیری بالایی می‌باشند (بی‌نام، ١٣٨٢). میانگین بارش سالانه در کل حوضه ٩٨٣ میلی متر بوده و اقلیم منطقه از روش دمارتن اصلاح شده و آمبرژه در ارتفاعات مرطوب و سرد و در پایین دست خیلی مرطوب می‌باشد (فتح ا... زاده، ١٣٩٤). در شکل ١ محدوده مورد مطالعه و در شکل ٢ فراوانی داده‌های مورد مطالعه نشان داده شده است.

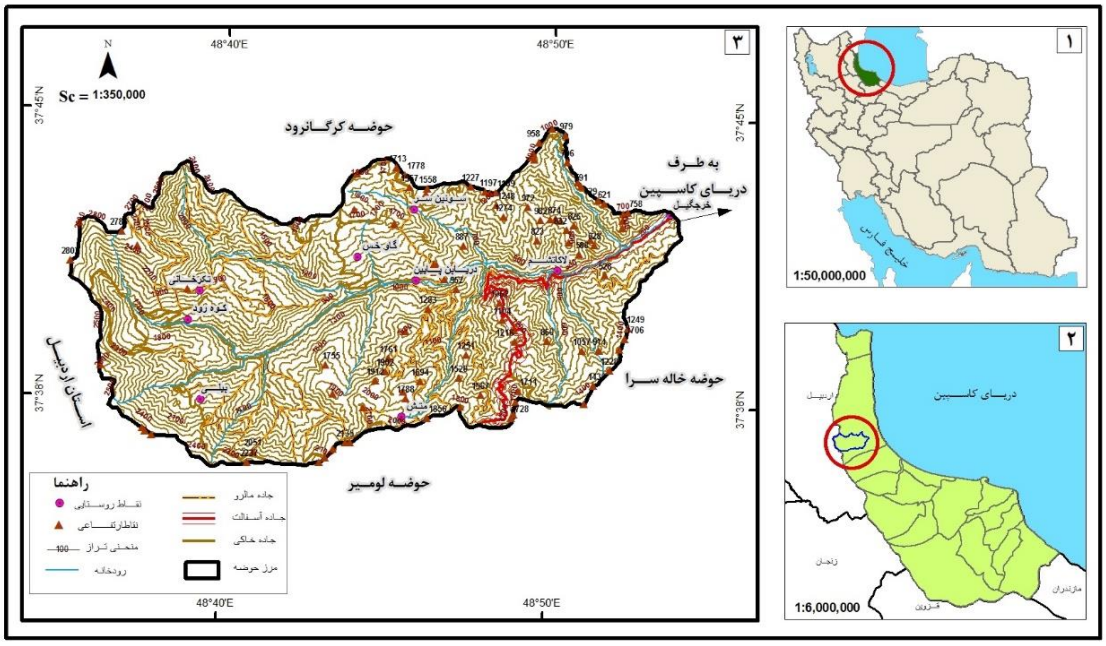
داده‌های مورد استفاده

در این تحقیق از داده‌های ماهانه شامل بارش، میانگین دما و رواناب با تاخیر ١ تا ٣ ماهه رودخانه ناورود در بین سال‌های ١٣٨٠ الی ١٣٩٠ استفاده گردید. اطلاعات مربوط به ایستگاه آبسنجی رودخانه ناورود، از شرکت آب منطقه‌ای استان گیلان اخذ گردید. از مجموع کل داده‌ها ٧٠ درصد از داده‌ها برای آموزش مدل‌ها و ٣٠ درصد باقیمانده برای آزمون در نظر گرفته شد. مشخصات آماری داده‌های استفاده شده در این تحقیق در جدول ١ ارائه گردیده است.

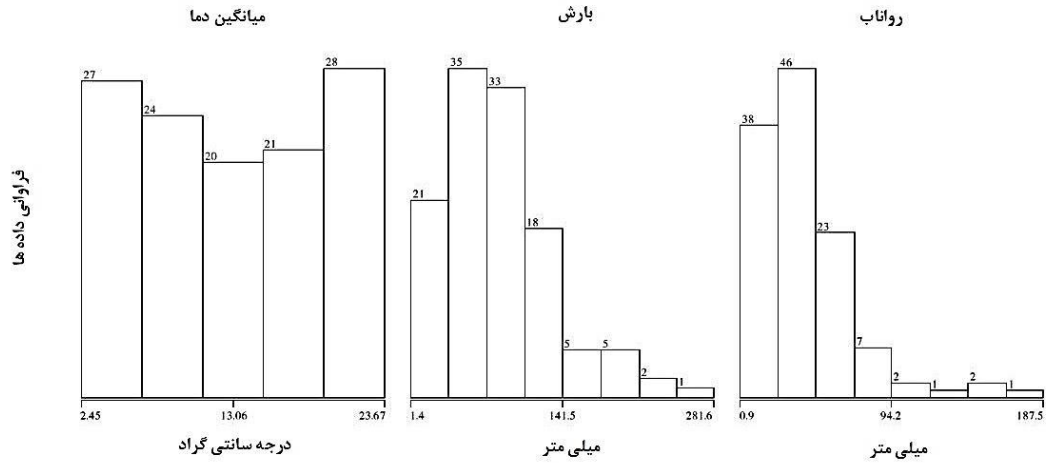
در این مطالعه قبل از انجام مدل‌سازی توسط روش‌های مذکور ابتدا همگنی و صحت کلیه داده‌ها توسط آزمون‌های همگنی پتیت[2]، نرمال استاندارد[3]، بیشند[4] و ون نیومن[5] بوسیله نرم‌افزار XLSTAT مورد بررسی قرار گرفتند.

1 - *RELIEF*

2 -*Pettit*
3 -*Standard Normal Homogeneity Test*
4 -*Buishand*
5 -*Von Neumann*



شکل ۱. موقعیت حوضه آبریز ناورود در استان گیلان و کشور ایران (فتح ا... زاده، ۱۳۹٤).



شکل۲.نمودار بافت‌نگار داده‌های مورد مطالعه.

جدول۱.مشخصات آماری داده‌های استفاده شده ماهانه.

مشخصات آماری	میانگین دما(C^{o})	بارش (mm)	رواناب (mm)
کمینه	۲/۴۵	۱/۴	۰/۹
بیشینه	۲۳/۶۷	۲۸۱/۶	۱۸۷/۵
میانگین	۱۲/۸۲۵	۸۳/۴۳۲	۳۹/۱۱۸
انحراف معیار	۶/۳۷۳	۵۱/۵۹۴	۳۱/۴۸

مدل رگرسیون بردار پشتیبان (SVR)

ماشین بردار پشتیبان یکی از روش‌های یادگیری با نظارت است که در سال ۱۹۹۲ میلادی توسط Vapnik و Chervonenkis بر پایه تئوری یادگیری آماری معرفی شد. منظور از رگرسیون، به دست آوردن یک ابرصفحه است که بر داده‌های مورد نظر برازش داده می‌شود. فاصله هر

نقطه از این ابرصفحه نشان دهنده خطای آن نقطه خاص است (شهرابی و شجاعی، ۱۳۹۰). بهترین روشی که تاکنون برای رگرسیون خطی پیشنهاد شده است، روش حداقل مربعات می‌باشد. با این وجود، برای مسائل رگرسیون، این امکان وجود دارد که استفاده از برآورد کننده کمترین مربعات در حضور داده‌های پرت، بطور کامل شدنی نباشد و درنتیجه رگرسور عملکرد ضعیفی را از خود به نمایش بگذارد. بنابراین می‌بایست یک برآورد کننده نیرومند که نسبت به تغییرات کوچک در مدل حساس نباشد را توسعه داد. تابع زیان غیرحساس بصورت رابطه (۱) تعریف می‌شود:

$$L^{\varepsilon}(x,y,f)=|y-f(x)|_{\varepsilon}=max\left(0,|y-f(x)|-\varepsilon\right) \qquad (۱)$$

اگر $|y-f(x)|\le\varepsilon$ باشد آنگاه تابع زیان برابر صفر و در غیر این‌صورت برابر $|y-f(x)|-\varepsilon$ خواهد بود. مجموعه داده‌های آموزشی در حالت کلی به صورت $S=\{(x_1,y_1),(x_2,y_2),\ldots(x_n,y_n)\}$ و کلاس تابع به‌صورت $f(x)=\{w^Tx+b, w\in R^m, b\in R\}$ می‌باشد. در صورتی که داده‌ها از مقدار مجاز تخطی نمایند، می‌بایست متناسب با مقدار تخطی، متغیر کمبود تعریف شود (شهرابی و شجاعی، ۱۳۹۰). مطابق با تابع زیان اشاره شده، کمینه‌سازی، مسأله اولیه، نقطه زینی[1] و مسأله دوگان[2] مربوطه با استفاده از توابع کرنل تشکیل داده و سپس شرایط کاروش کان تاکر[3] بررسی می‌شود:

$$min_{w,b}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\left(\overset{+}{\xi}_i+\overset{-}{\xi}_i\right) \qquad (۲)$$

subject to:

$$f(x)-y_i\le\varepsilon+\overset{+}{\xi}_i,\ f(x)-y_i\le\varepsilon+\overset{-}{\xi}_i,\ \overset{+}{\xi}_i,\overset{-}{\xi}_i\ge0,$$

$$i=1,2,3,\ldots n$$

در رابطه (۲)، $\|w\|^2$ نرم بردار وزن، $\overset{+}{\xi}_i,\overset{-}{\xi}_i$ متغیرهای کمبود کمکی هستند و پارامتر C ضریب تعادل پیچیدگی میان ماشین و تعداد نقاط تفکیک‌ناپذیر است که با سعی و خطا بدست می‌آید (شهرابی و شجاعی، ۱۳۹۰). برای حل مسئله با ابعاد خیلی بالا و تبدیل فرم غیرخطی به خطی در این روش از توابع کرنل استفاده می‌شود. در این راستا از توابع کرنل مختلفی از جمله کرنل‌های چندجمله‌ای ساده (رابطه ۳)، چند جمله‌ای نرمال شده (رابطه ٤)، شعاع محور (رابطه ٥) و پیرسون (رابطه ٦) می‌توان استفاده نمود. بنابراین کافی است در مسائل غیرخطی از کرنل مقادیر ورودی به جای خود تابع استفاده شود. با توجه به تئوری توضیح داده شده دقت در تعیین پارامترهای هموارساز C، میزان ε و مقدار پارامترهای موجود در تابع کرنل نقش بسزایی در کاهش خطای مسأله دارد. در روابط (۳) تا (٦) چهار تابع کرنل پرکاربرد ارائه شده است.

$$K(x_i,x_j)=(x_i^Tx_j+1)^p \qquad (۳)$$

$$k(x_i,x_j)=\frac{(x_i^Tx_j+1)^p}{\sqrt{(x_i^Tx)(x_j^Tx_j)}} \qquad (٤)$$

$$k(x_i,x_j)=\exp\left(-\Upsilon|x_i-x_j|^2\right) \qquad (٥)$$

$$k(x_i,x_j)=\frac{1}{\left[1+\left(2\sqrt{\|x_i,x_j\|^2}\sqrt{2^{1/w}-1/\sigma}\right)^2\right]^w} \qquad (٦)$$

پارامترهای ω، σ، Υ و P پارامترهای مخصوص هر کدام از توابع کرنل می‌باشند که توسط روش‌هایی مانند انجام آزمون و خطا مقدار بهینه آن‌ها تعیین می‌گردد.

مدل نزدیک‌ترین همسایگی

این تکنیک بر اساس مفهوم تشابه شکل گرفته است. نتایج حاصل از استدلال مبتنی بر حافظه[4]، بر پایه موقعیت‌های مشابهی که در گذشته اتفاق افتاده، بنا نهاده شده است. بدین معنی که این الگوریتم، تمامی موارد موجود را ذخیره کرده و بر اساس اندازه‌گیری شباهت‌ها به پیش‌بینی عددی هدف می‌پردازد. الگوریتم انجام پیش‌بینی با استفاده از روش نزدیک‌ترین همسایگی به صورت زیر

[1] - *Saddle* Point
[2] - *The* Dual *Problem*
[3] - *The Karush-Kuhn-Tucker*

4- Memory-Based Reasoning

است: بردار سطری m ستونه مقادیر متغیرهای پیش‌بینی‌کننده x_j در زمان t به صورت رابطه (۷) می‌باشد:

$$\mathrm{Pr}_{jt} = (x_{jt}) \quad j = 1...m \qquad (۷)$$

ماتریس m ستونه n سطری از مقادیر متغیرهای پیش‌بینی کننده x_j در سری زمانی تاریخی به صورت زیر است.

$$\mathrm{Pr}_{j,(t-i)} = (x_{j(t-i)}) \quad j = ۱...m, i = ۱...n \qquad (۸)$$

با استفاده از تابع فاصله‌سنج، فواصل بین بردار $P_{rj,t}$ با سطرهای ماتریس $P_{rj,(t-i)}$ استخراج می‌گردد.

$$Dist(t-i) = f(w_j, x_{j,(t-i)}, x_{jt}) \qquad (۹)$$

که در این رابطه اندیس j نشان‌دهنده متغیرهای پیش‌بینی کننده و اندیس i بیان کننده گام زمانی در سری تاریخی است. مقادیر W_j وزن‌هایی است که برای پیش‌بینی کننده‌ها، در نظر گرفته می‌شود. به طور کلی می‌توان برای بهبود عملکرد مدل‌های نزدیک‌ترین همسایگی اقداماتی از جمله انتخاب روشی جهت تخمین بهترین همسایه‌ها، توسعه توابع انتقال اطلاعات و توسعه توابع فاصله‌سنجی انجام داد. برای تخمین بهترین همسایگی‌ها در روش نزدیک‌ترین همسایگی، روش‌های مختلفی ارائه شده‌اند که بسته به دقت و مورد استفاده و پیچیدگی و حجم مسأله قابل استفاده هستند. از جمله این روش‌ها استفاده از رابطه تجربی $K = \sqrt{n}$ می‌باشد که در این رابطه n طول سری زمانی و k بهترین تعداد همسایگی مورد استفاده در این روش است (عزمی و عراقی نژاد، ۱۳۹۱).

میزان کارایی این رابطه با افزایش طول سری زمانی افزایش می‌یابد. سعی خطا روش دیگری است که می‌توان در این زمینه به کاربرد بدین ترتیب که با انتخاب همسایگی‌های مختلف در محدوده بهترین همسایگی و استخراج خطاهای پیش‌بینی، سعی در پیدا کردن بهترین همسایگی‌ها را داشت. قدم بعدی انتخاب تابعی برای انتقال اطلاعات می‌باشد که وظیفه این تابع همگن کردن فضای اطلاعات ورودی مدل به جهت جلوگیری از خطاهای مربوطه می‌باشد. تابع دیگری که نقش اساسی در عملکرد روش نزدیک‌ترین همسایگی دارا می‌باشد، تابع فاصله-سنجی است که از جمله این توابع می‌توان به توابع اقلیدسی (رابطه ۱۰)، مینکوسکی (رابطه ۱۱)، منهتن (رابطه ۱۲) و چبی‌شف (رابطه ۱۳) اشاره کرد (عزمی و عراقی نژاد، ۱۳۹۱).

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (۱۰)$$

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \qquad (۱۱)$$

$$\sqrt{\sum_{i=1}^{k}|x_i - y_i|} \qquad (۱۲)$$

$$\mathrm{Max}((|x_i - x_{i-1}|, |y_i - y_{i-1}|) \qquad (۱۳)$$

در این روابط x و y نقاط یا بردارهایی هستند که می‌خواهیم سایر نقاط فاصله‌ی معینی از آنها دارا باشند و k برابر با تعداد نقاط مورد نظر می‌باشد. در این تحقیق جهت مدل‌سازی میزان رواناب رودخانه ناورود، توسط روش‌های رگرسیون بردار پشتیبان و نزدیک‌ترین همسایگی از نرم‌افزار WEKA که در دانشگاه وایکاتو نیوزلند توسعه یافته، استفاده گردید.

الگوریتم‌های انتخاب متغیرهای ورودی مدل‌ها

انتخاب ویژگی توسط الگوریتم رلیف

رلیف یک الگوریتم انتخاب ویژگی است که برای کاهش ابعاد مسئله به کار می‌رود و اولین بار توسط Kira و Rendell (۱۹۹۲) پیشنهاد شده است. از جمله نکات قوت این الگوریتم می‌توان به ساده بودن اصول و عدم پیچیدگی آن، قابل حل بودن با توابع چند جمله‌ای مرتبه پایین، قابل استفاده بودن برای داده‌های پیوسته و نیاز به تعداد کم داده‌های آموزشی اشاره کرد. در یک مجموعه داده با تعداد N نمونه (داده مشاهداتی) و تعداد P ویژگی که مربوط به دو طبقه مختلف هستند، هر ویژگی باید در بازه (۱، ۰) قرار گیرد. الگوریتم مذکور m بار تکرار خواهد

شد که در هر مرتبه از یک بردار وزنی متفاوت که از صفر شروع می‌گردد، استفاده می‌کند.

در هر تکرار، الگوریتم مذکور بردار ویژگی X را که متعلق به یک نمونه تصادفی است و بردارهای ویژگی نزدیک‌ترین نمونه به نمونه X در طبقه مورد نظر را توسط تابع فاصله‌سنجی اقلیدسی انتخاب می‌کند. پس از m تکرار هر یک از عناصر بردار وزن توسط m تقسیم‌بندی می‌شوند نتیجه این عمل به دست آمدن یک بردار مرتبط است که اگر مقدار بردار مرتبط یک ویژگی از آستانه تعریف شده بیشتر گردد، آن ویژگی انتخاب می‌گردد (Kira & Rendell, 1992). همچنان که ذکر شد از مهم‌ترین ویژگی ها و برجستگی‌های الگوریتم رلیف مناسب بودن آن برای استفاده برای مجموعه داده‌های با تعداد نمونه آموزشی کم می‌باشد. از این رو در این مطالعه با توجه به طول نسبتاً کم آماری داده‌ها از این روش نیز برای تعیین پارامترهای تاثیرگذار برای مدل‌سازی رواناب استفاده شد.

انتخاب ویژگی بر اساس همبستگی[1] (CFS)

انتخاب ویژگی بر اساس همبستگی یک روش معمول و پرکاربرد برای انتخاب متغیرهای ورودی و کاهش ابعاد مسئله به کار می‌رود و توسط Hall (۱۹۹۹) معرفی شده است. روش همبستگی به زیرمجموعه‌هایی که دارای ویژگی‌هایی با بیشترین ضریب همبستگی با کلاس نمونه مورد نظر هستند، امتیاز می‌دهد و متغیرهایی که بیشترین امتیاز را دارا باشند، به عنوان متغیر اصلی در نظر می‌گیرد. این الگوریتم توانایی بالایی در تشخیص سریع داده‌های نامربوط، اضافی و دارای خطا دارد که عموماً منجر به حذف نیمی از داده‌ها می‌گردد. این ویژگی با کاهش ابعاد مسئله سبب افزایش بهره‌وری مدل‌ها می‌گردد.

معیارهای ارزیابی

عملکرد رگرسیون بردار پشتیبان و نزدیک‌ترین همسایگی در این تحقیق بر پایه محاسبه ضریب همبستگی[2]، جذر میانگین مربعات خطا[3]، و میانگین خطای مطلق[4] مورد ارزیابی قرار گرفت. روابط محاسبه آماره‌های فوق در روابط (۱۴) تا (۱۶) ارائه گردیده است. در این روابط مقادیر X شامل مقادیر مشاهداتی و مقادیر Y شامل مقادیر محاسباتی می‌باشند.

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (۱۴)$$

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n}(y_i - x_i)^2}{N}} \quad (۱۵)$$

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|x_i - y_i| \quad (۱۶)$$

نتایج و بحث

نتایج مربوط به آماده‌سازی داده‌ها

با توجه به اهمیت استفاده از داده‌های صحیح و قابل اعتماد برای دقت بیشتر در نتایج مدل‌سازی در مطالعات پیش‌رو در مرحله اول نسبت به انجام آزمون همگنی داده‌ها توسط چهار آزمون ذکر شده اقدام گردید. برای انجام این آزمون‌ها از نرم‌افزار آماری XLSTAT استفاده گردیده است. در جدول ۲ نتایج مربوط به آزمون همگنی داده‌ها برای داده‌های ایستگاه ناورود ارائه گردیده است.

در این آزمون فرض صفر بیانگر همگن بودن داده‌ها و فرض یک بیانگر غیر همگن بودن داده‌هاست. چنان‌که مقدار p-value از مقدار درجه اطمینان α بزرگ‌تر باشد، فرض صفر صحیح است در غیر این صورت فرض یک قابل قبول می‌باشد. نتایج جدول ۲ نشان‌دهنده این مطلب است که اگرچه همگنی داده‌های دما نسبت به بارش و همگنی داده‌های بارش هم نسبت به رواناب بیشتر است ولی در حالت کلی همه داده‌های مربوط به پارامترهای دما، بارش و رواناب در کلیه آزمون‌های استفاده شده همگن بوده و می‌توانند برای مدل‌سازی به‌کار برده شوند.

1 - Correlation based feature selection

2 - Correlation coefficient
3 - Root Mean Square Error
4 - Mean Absolute Error

جدول ۲. نتایج آزمون همگنی داده‌ها برای ایستگاه ناورود.

نام آزمون	α	رواناب	بارش	دمای میانگین
		p-value	p-value	p-value
پتیت	۰/۰۵	۰/۰۶۴	۰/۰۷۳	۰/۹۵۸
نرمال استاندارد	۰/۰۵	۰/۰۹۲	۰/۳۲۴	۰/۱۴۷
بیشند	۰/۰۵	۰/۰۵۹	۰/۱۱۱	۰/۹۶
ون نیومن	۰/۰۵	۰/۰۶۸	۰/۰۷۷	۰/۹۱۵

ترکیب پارامترهای ورودی مدل

چنان‌چه در بخش‌های قبلی اشاره شد در این مطالعه از پارامترهای میانگین دما، بارش و رواناب در مقیاس ماهانه با تاخیر یک تا سه ماه استفاده گردیده است. هدف از روش‌های انتخاب ویژگی اولاً کاهش ابعاد مسئله و دوماً تعیین تاثیرگذارترین پارامترها در مدل‌سازی دقیق میزان رواناب در حوضه آبریز ناورود و در آخر معرفی بهترین روش برای انتخاب ویژگی‌های ورودی می‌باشد. جدول ۳ نشان‌دهنده پارامترهای ورودی انتخاب شده برای مدل سازی رواناب در ماه جاری (Q_t) بر اساس روش‌های الگوریتم RELIEF و CFS می‌باشد. در تعریف سناریوها، (T) معرف دمای میانگین ماهانه، (P) معرف بارش ماهانه، (Q) معرف رواناب و اندیس (t) نشان‌دهنده ماه جاری و (t-i) نشان‌دهنده تاخیر بین یک تا سه ماه می‌باشد.

همچنان که در قسمت مواد و روش‌ها بیان گردید، الگوریتم انتخاب ویژگی رلیف برای هر ویژگی یک بردار وزنی مرتبط با ویژگی هدف تعریف کرده و پارامترهای ورودی مرتبط را بر اساس وزن هر ویژگی رتبه‌بندی می کند. در هنگام تعیین پارامترهای تاثیرگذار بر مدل سازی رواناب در ناورود مشخص شد، پارامترهای تاثیرگذار ارائه شده توسط الگوریتم رلیف (ترکیب S۱) به ترتیب اهمیت شامل بارش در ماه جاری، رواناب در دو ماه قبل، بارش در دو ماه قبل، رواناب در ماه قبل و رواناب در سه ماه قبل می‌باشند. همچنین این الگوریتم پارامتر دما را به عنوان ویژگی موثر شناسایی نکرده است که این مسئله با توجه به اینکه عامل دما در حوضه‌های برف‌گیر دارای تاثیر بیشتری بوده و در حوضه‌های پربارش مانند ناورود تاثیر چندانی ندارد، قابل توجیه است.

بررسی نتایج مدل‌سازی رواناب

با توجه به اهمیت انتخاب مواردی همچون نوع تابع کرنل در روش رگرسیون بردار پشتیبان و همچنین نوع تابع فاصله‌سنجی در روش نزدیک‌ترین همسایگی ابتدا عملکرد این مدل‌ها به ازای توابع مختلف در حالت استفاده از مقادیر پیش فرض پارامترهای هر تابع سنجیده شد تا تابع بهینه برای هر روش انتخاب و سپس نسبت به مدل‌سازی رواناب با هر یک از ترکیب‌های دوگانه براساس توابع بهینه اقدام گردد. برای این منظور ۷۰ درصد از داده‌ها برای قسمت آموزش مدل‌ها و ۳۰ درصد از داده‌ها برای آزمون مدل‌ها در نظر گرفته شد، استفاده گردید. نتایج مربوط به روش‌های رگرسیون بردار پشتیبان و نزدیک‌ترین همسایه با استفاده از توابع مختلف در حالت پیش فرض تنظیمات مربوط به ساختار هر تابع، در جدول ٤ ارائه گردیده است.

جدول ۳.پارامترهای ورودی تعیین شده توسط روش‌های انتخاب ویژگی.

ترکیب پارامتر	روش	پارامترهای ورودی تعیین شده
$S_۱$	الگوریتم RELIEF	$P_t, P_{t-2}, Q_{t-1}, Q_{t-2}, Q_{t-3}$
$S_۲$	الگوریتم CFS	$P_t, P_{t-1}, T_{t-1}, Q_{t-2}$

جدول ٤.نتایج مدل‌سازی رواناب توسط روش‌های داده‌کاوی با توابع مختلف در حالت تنظیمات ساختار پیش فرض.

روش	تابع مورد استفاده	ترکیب (S۱)			ترکیب (S۲)		
		MAE (mm)	RMSE (mm)	R	MAE (mm)	RMSE (mm)	R
رگرسیون بردار پشتیبان	چند جمله‌ای ساده	۱۵/۲۸	۲۸/۷۸	۰/۷۶	۱۴/۱۴	۲۶/۳۶	۰/۸۲
	چند جمله‌ای نرمال شده	۱۷/۳۲	۳۰/۶۰	۰/۶۴	۱۴/۹۱	۲۹/۲۸	۰/۶۶
	شعاع محور	**۱۰/۹۰**	**۱۹/۹۳**	**۰/۹۲**	**۹/۹۷**	**۱۶/۳۷**	**۰/۹۶**
	پیرسون	۱۱/۹۱	۲۰/۱۷	۰/۸۷	۱۴/۰۲	۲۳/۱۵	۰/۹۴
نزدیکترین همسایگی	**اقلیدسی**	**۱۲/۴۰**	**۲۰/۹۲**	**۰/۸۸**	**۱۰/۸۹**	**۱۵/۸۹**	**۰/۹۱**
	چبی‌شف	۱۳/۷۲	۱۸/۹۷	۰/۸۶	۱۴/۲۰	۱۸/۹۵	۰/۸۶
	منهتن	۱۳/۴۴	۱۸/۸۱	۰/۸۶	۱۲/۳۵	۱۶/۷۴	۰/۸۹

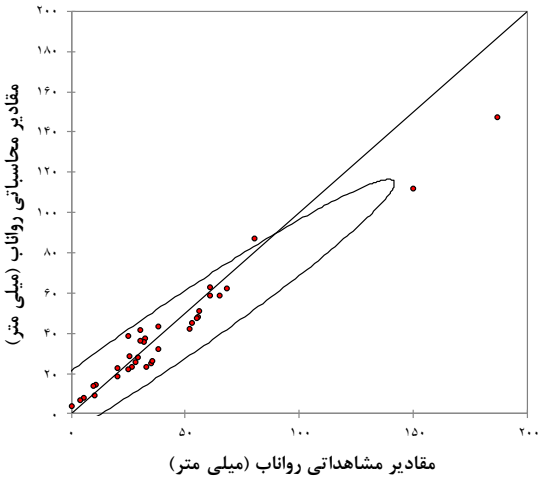
بر اساس نتایج ارائه شده در جدول ٤ مشاهده می‌گردد که در روش رگرسیون بردار پشتیبان بهترین نتایج با استفاده از تابع شعاع محور و در روش نزدیک‌ترین همسایگی بهترین نتایج با استفاده از تابع اقلیدسی حاصل می‌شود. در روش نزدیک‌ترین همسایگی برای تعیین تعداد همسایگی بهینه برای افزایش دقت مدل از روش آزمون و خطا استفاده گردید. برای این منظور تغییر در میزان دقت و خطای مدل به ازای تعداد همسایگی بین ۱ تا ۷ بررسی شده و مشاهده گردید با افزایش تعداد همسایگی مورد استفاده برای مدل‌سازی دقت مدل رفته رفته افزایش می‌یابد و بهترین نتیجه در هنگام استفاده از تعداد سه همسایه حاصل می‌گردد اما پس از آن با افزایش تعداد همسایه مورد استفاده دقت مدل رفته رفته کاهش پیدا می‌کند. در روش رگرسیون بردار پشتیبان نیز برای تعیین مقدار بهینه مقادیر ضریب هموارساز (C) و پارامتر گاما (γ) از روش آزمون و خطا استفاده شد. برای این منظور تغییرات مقدار آماره‌های دقت و خطای مدل به ازای مقادیر ۰/۰۱ تا ۲۰ مورد بررسی قرار گرفت و مشاهده شد که در روش رگرسیون بردار پشتیبان ساختار بهینه مدل با مقادیر (C=۵ و γ=۰/۲) به‌دست می‌آید. همچنین مشاهده گردید که روش رگرسیون بردار پشتیبان به هنگام استفاده از تابع کرنل شعاع محور حساسیت بسیار زیادی نسبت به مقدار دقیق پارامتر گاما (γ) دارد، زیرا با تغییر در مقدار این پارامتر دقت مدل بسیار تغییر پیدا کرد در صورتی که با تغییر مقدار ضریب هموارساز (C) تغییر بسیار زیادی در نتایج مشاهده نگردید. در جدول ۵ نتایج بهینه به‌دست آمده از هر دو روش نوین داده‌کاوی بررسی شده در این تحقیق برای هر کدام از مجموعه پارامترهای ورودی تعیین شده، ارائه گردیده است.

نتایج جدول ۵ نشان‌دهندهٔ این مطلب است که هر دو روش داده‌کاوی استفاده شده در این تحقیق توانایی بالایی در مدل‌سازی دقیق میزان رواناب در حوضه مورد مطالعه دارند. اما بررسی آماره‌های ضریب همبستگی، ریشه میانگین مربعات خطا و میانگین خطای مطلق نشان‌دهندهٔ برتری نسبی روش رگرسیون بردار پشتیبان می‌باشد. چنان‌که آماره میانگین خطای مطلق و آماره ریشه میانگین مربعات خطا حدوداً ۵۰ درصد در روش رگرسیون بردار پشتیبان نسبت به روش نزدیک‌ترین همسایگی کمتر می‌باشد. این اختلاف خطای نسبتاً چشم‌گیر در این دو روش می‌تواند مربوط به بهره‌گیری این روش از تابع کرنل باشد که توانایی حل مسائل غیرخطی را در این روش بسیار افزایش می‌دهد.

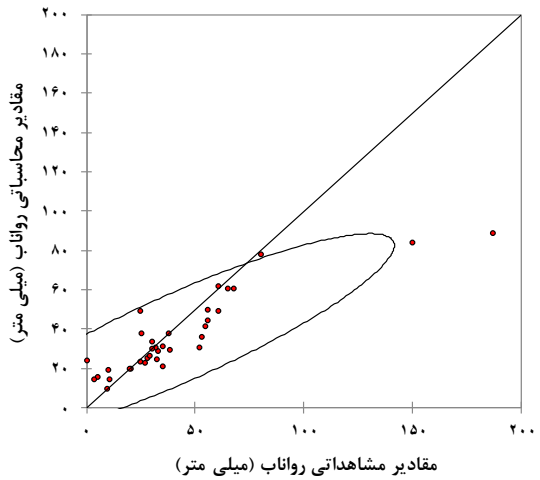
جدول۵. نتایج مدل‌سازی رواناب توسط مدل‌های داده‌کاوی به ازای مقادیر بهینه پارامترهای هر یک از توابع برای ترکیب‌های S۱ و S۲.

	پارامترهای ورودی	نزدیک‌ترین همسایگی			رگرسیون بردار پشتیبان		
		MAE	RMSE (mm)	R	MAE (mm)	RMSE (mm)	R
S_1	$P_t, P_{t-2}, Q_{t-1}, Q_{t-2}, Q_{t-3}$	۱۲/۵۲	۱۹/۰۶	۰/۸۹	۷/۳۹	۱۱/۴۲	۰/۹۷
S_2	$P_t, P_{t-1}, T_{t-1}, Q_{t-2}$	۱۱/۹۴	۱۷/۷۸	۰/۹۴	۶/۷۸	۹/۳۲	۰/۹۸

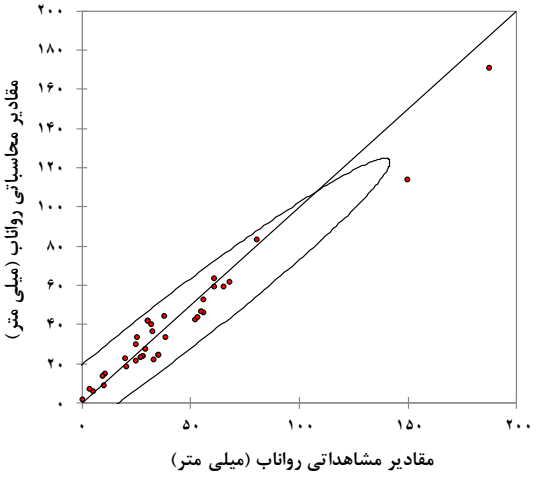
با بررسی نتایج جدول ۵ مشاهده می‌شود که اگرچه نتایج حاصل از دو ترکیب به‌همدیگر بسیار نزدیک می باشد اما ترکیب پارامترهای ورودی تعیین شده توسط روش همبستگی (CFS) شامل پارامترهای (P_t, P_{t-1}, T_{t-1}, Q_{t-2}) نسبتا دارای دقت بیشتر و خطای کمتری نسبت به ترکیب پارامترهای مشخص شده توسط الگوریتم رلیف است. برای درک بهتر نتایج به‌دست آمده در این پژوهش و بررسی بصری دقت مدل‌های استفاده شده در مدل‌سازی میزان رواناب، نمودارهای پراکنش مقادیر محاسباتی از روش‌های داده‌کاوی نسبت به مقادیر مشاهداتی رواناب و همچنین نمودارهای سری زمانی بارش-رواناب در بازهٔ داده‌های قسمت آزمون هم برای ترکیب پارامتر ورودی (S۱) و هم برای ترکیب پارامتر ورودی (S۲) در شکل‌های ۳ تا ۸ نشان داده شده است.



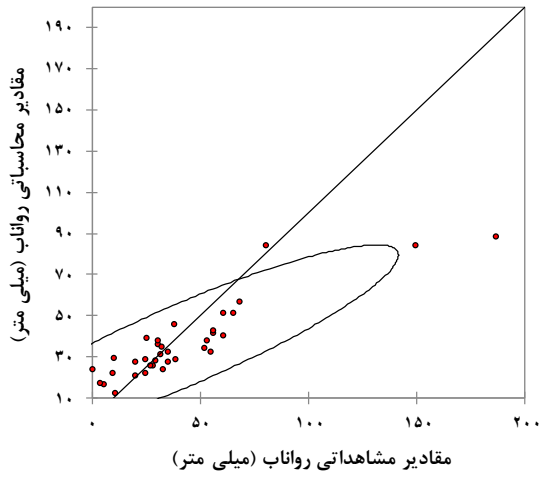
شکل۳. نمودار پراکنش مقادیر به‌دست آمده از رگرسیون بردار پشتیبان نسبت به مقادیر مشاهداتی (ترکیب پارامتر S۱).



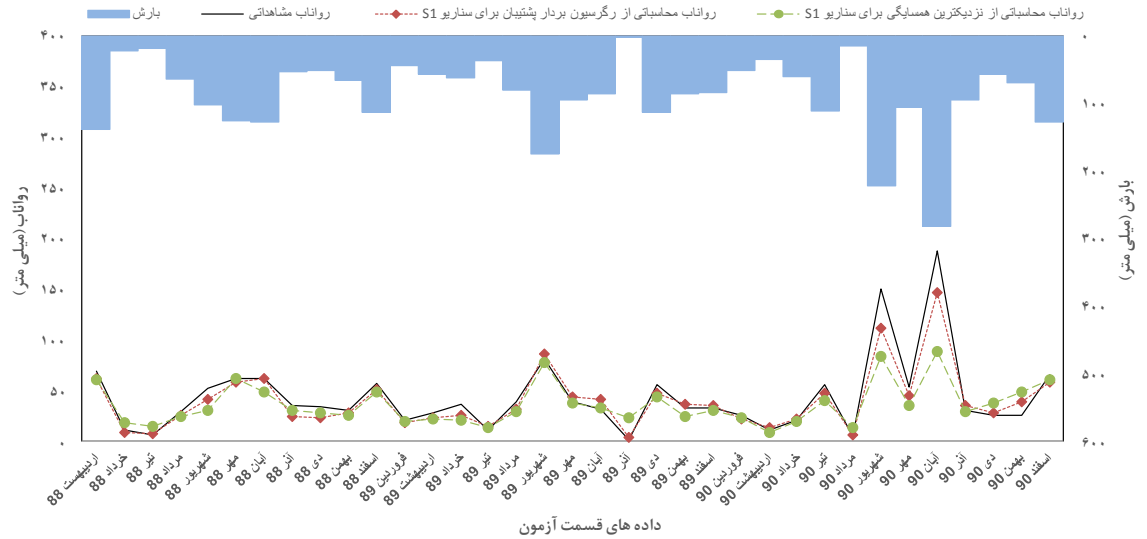
شکل٤. نمودار پراکنش مقادیر به‌دست آمده از نزدیک‌ترین همسایگی نسبت به مقادیر مشاهداتی (ترکیب پارامتر S۱).



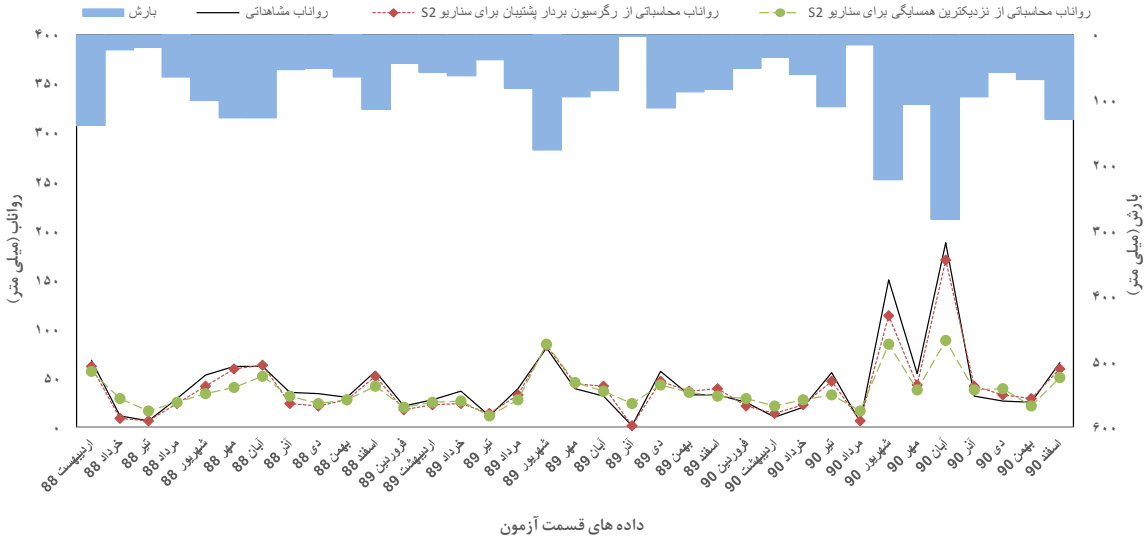
شکل۵. نمودار پراکنش مقادیر به‌دست آمده از رگرسیون بردار پشتیبان نسبت به مقادیر مشاهداتی (ترکیب پارامتر S۲).



شکل٦. نمودار پراکنش مقادیر به‌دست آمده از نزدیک‌ترین همسایگی نسبت به مقادیر مشاهداتی (ترکیب پارامتر S۲).



شکل۷. نمودار سری زمانی بارش–رواناب برای داده‌های قسمت آزمون مدل‌ها (ترکیب پارامتر S1).



شکل۸. نمودار سری زمانی بارش–رواناب برای داده‌های قسمت آزمون مدل‌ها (ترکیب پارامتر S۲).

از بررسی نمودارهای ارائه شده در شکل‌های ۳ تا ٦ مشاهده می‌گردد که مقادیر رواناب محاسباتی به دست آمده از روش رگرسیون بردار پشتیبان، دارای پراکنش و همخوانی بهتری با مقادیر مشاهداتی رواناب نسبت به روش نزدیک‌ترین همسایگی همسایگی در بازهٔ داده‌های آزمون دارد که تطابق بیشتر محدوده نقاط واقع در بازهٔ اطمینان در روش رگرسیون بردار پشتیبان با نیمساز ربع اول تأییدکنندهٔ این مطلب است. همچنین ترکیب (S۲) شامل پارامترهای ورودی $(P_t, P_{t-1}, T_{t-1}, Q_{t-2})$، هم در روش رگرسیون بردار پشتیبان و هم در روش نزدیک‌ترین همسایگی پراکنش و همخوانی بیشتری با مقادیر مشاهداتی داشته که نشان‌دهندهٔ عملکرد نسبتاً بهتر روش همبستگی در تعیین پارامترهای ورودی برای روش‌های داده‌کاوی است. همچنین با بررسی نمودار سری زمانی بارش–رواناب در شکل‌های ۷ و ۸ مشاهده می‌گردد که علی‌رغم دقت بالای هر دو روش داده‌کاوی بررسی شده در این

تحقیق، روش رگرسیون بردار پشتیبان دارای خطای کمتری به خصوص در مدل‌سازی مقادیر بحرانی و سیلابی رواناب می‌باشد. توانایی بالاتر روش رگرسیون بردار پشتیبان در مدل‌سازی مقادیر حدی رواناب مربوط به بهره‌گیری این روش از توابع کرنل است. توابع کرنل توانایی این روش در حل و مدل‌سازی مسائل غیرخطی را افزایش می‌دهند که این موضوع توسط محققینی همچون Adamowski (۲۰۱۳) و غظنفری‌هاشمی و شهیدی (۱۳۹۱) نیز گزارش شده است. این موضوع از این جهت بسیار حائز اهمیت است که، می‌تواند مدیریت بهتر منابع آب و کاهش خطرات احتمالی وقوع سیلاب را با پیش آگاهی و انجام اقدامات پیش‌گیرانه ممکن سازد.

نتیجه‌گیری

یکی از نیازهای اساسی در طراحی پروژه‌های آبیاری و زهکشی و مدیریت منابع آب سطحی، تعیین به هنگام و دقیق میزان رواناب حاصل از بارش است. برای این منظور در این مطالعه امکان کاربرد روش‌های داده‌کاوی شامل رگرسیون بردار پشتیبان و نزدیک‌ترین همسایگی در مدل سازی مقدار رواناب حاصل از بارش ماهانه در حوضه ناورود مورد سنجش قرار گرفت. پارامترهای استفاده شده در این تحقیق عبارت‌اند از میانگین دما، بارش و رواناب در مقیاس ماهانه که قبل از به کارگیری برای مدل‌سازی توسط آزمون‌های آماری پتیت، نرمال استاندارد، بیشند و ون نیومن از لحاظ همگن بودن بررسی شدند. نتایج آزمون ها نشان‌دهندهٔ صحت و همگنی داده‌ها است. نتایج بدست آمده نشان داد هر دو مدل از قابلیت‌های خوبی در مدل-سازی رواناب ماهانه برخوردارند. مرور مطالعات انجام شده در زمینه مدل‌سازی رواناب با روش‌های هوشمند و مقایسه نتایج آن‌ها با نتایج به‌دست آمده از تحقیق حاضر، نشان‌دهنده کارایی و دقت قابل قبول و در اکثر موارد بهتر روش‌های رگرسیون بردار پشتیبان و نزدیک‌ترین همسایگی است. به عنوان مثال سیدیان و همکاران (۱۳۹۳) با استفاده از ماشین بردار پشتیبان دبی رودخانه گرگانرود را در دو مقیاس هفتگی و ماهانه پیش‌بینی کردند. آنها با استفاده از ماشین بردار پشتیبان مقدار دبی را در ایستگاه قره شور در مقیاس ماهانه با ضریب همبستگی برابر با ۰/۸ و ریشه میانگین مربعات خطا برابر با ۰/۳۲ میلی متر و در مقیاس هفتگی با ضریب همبستگی برابر با ۰/۵۷ و ریشه میانگین مربعات خطا برابر با ۰/۵۰ میلی متر پیش‌بینی نمودند. همچنین در تحقیق دیگری Hu و همکاران (۲۰۱۱) با استفاده از ماشین بردار پشتیبان بارش-رواناب را در رودخانه فنحه واقع در چین که در یک منطقه نیمه خشک واقع شده است مدل‌سازی نمودند. نتایج نشان داد که در بهترین حالت در مقیاس ماهانه ضریب همبستگی برابر با ۰/۸۵ به‌دست آمد که این مساله کارایی این روش را در حوضه‌های آبریز دیگر با خصوصیات اقلیمی و نوع خاک متفاوت نیز نشان می‌دهد. در تحقیق حاضر نیز روش رگرسیون بردار پشتیبان در بهترین حالت (ترکیب S۲)، با ارائه مقادیر ضریب همبستگی برابر ۰/۹۸، ریشه میانگین مربعات خطا برابر ۹/۳۲ (میلی‌متر) و میانگین خطای مطلق برابر با ٦/۷۸ (میلی‌متر) و روش نزدیک‌ترین همسایگی در بهترین حالت (ترکیب S۲)، با ارائه مقادیر ضریب همبستگی برابر با ۰/۹٤، ریشه میانگین مربعات خطا برابر ۱۷/۷۸ (میلی‌متر) و میانگین خطای مطلق برابر با ۱۱/۹٤ (میلی‌متر) دارای دقت و کارایی نسبتا بالایی در این زمینه می‌باشند. بررسی نتایج حاصل مدل‌سازی رواناب توسط روش‌های داده‌کاوی با پارامترهای ورودی تعیین شده توسط الگوریتم‌های رلیف و همبستگی نشان‌دهندهٔ عملکرد نسبتاً بهتر روش همبستگی با تعیین پارامترهای بهینه ورودی در ترکیب پارامتر (S۲) شامل پارامترهای (P_t, P_{t-1}, T_{t-1}, Q_{t-2}) می‌باشد. در حالت کلی کاربرد روش رگرسیون بردار پشتیبان، با توجه به دقت بالاتر و خطای کمتر نسبت به روش نزدیک‌ترین همسایگی، به خصوص در مقادیر بالای رواناب (رواناب سیلابی)، برای

حوضه‌های دارای جریان‌های سیلابی، به عنوان روشی دقیق و کارامد برای مدل‌سازی رواناب حاصل از بارش توصیه می‌گردد. زیرا با برآورد دقیق سیلاب می‌توان ضمن بهره‌برداری بهینه از منابع آبی و ذخیره آن در سدهای مخزنی خطرات ناشی از سیل را در سیلابدشت‌ها کاهش داد. همچنین نتایج بدست آمده نشان داد، ترکیب پارامترهای مبتنی بر الگوریتم رلیف بعلت درنظر نگرفتن پارامتر دما در فرآیند مدل‌سازی با اختلاف قابل اغماض و بسیار اندک، رقابت نزدیکی با انتخاب ترکیب پارامترها براساس روش همبستگی داشته و به عنوان روشی کارامد می‌تواند برای تعیین پارامترهای موثر در فرآیندهای هیدرولوژیکی به خصوص مدل‌سازی رواناب حاصل از بارش در حوضه‌های فاقد طول آماری مناسب مورد استفاده قرار گیرد.

منابع مورد استفاده

بی نام، ۱۳۸۲. گزارش طرح جامع چند منظوره حوضه آبریز ناورود. اداره کل منابع طبیعی استان گیلان.

جودی حمزه آباد، آ.، کدخدا حسینی، م.، اخوان، س. و نوروزی، ح. ۱۳۹۵. ارزیابی مدلهای SWAT و SVM در شبیه‌سازی رواناب رودخانه لیقوان. دانش آب و خاک، ۲۶(۱): ۱۳۷-۱۵۰.

خزایی، م.، میرزایی، م. و ملکیان، آ. ۱۳۹۳. ارزیابی کارایی مدل‌های MLP و RBF در مدل‌سازی بارش-رواناب در مقیاس‌های زمانی مختلف. دو فصلنامه مدیریت آب در مناطق خشک، ۱(۱): ۱-۱۲.

رضایی، ع.، مهدوی، م.، لوکس، ک.، فیض نیا، س. و مهدیان، م. ۱۳۸۶. مدل‌سازی منطقه‌ای دبی‌های اوج در زیر حوضه‌های آبخیز سد سفیدرود با استفاده از شبکه عصبی مصنوعی. مجله علوم و فنون کشاورزی و منابع طبیعی، ۱: ۲۵-۴۰.

سیدیان، س م.، سلیمانی، م. و کاشانی، م. ۱۳۹۳. پیش‌بینی دبی جریان رودخانه با استفاده از داده‌کاوی و سری زمانی. اکوهیدرولوژی، ۳: ۱۶۷-۱۷۹.

شهرابی، ج. و ذوالقدر شجاعی، ع. ۱۳۹۰. داده‌کاوی پیشرفته (مفاهیم و الگوریتم‌ها). انتشارات جهاد دانشگاهی واحد صنعتی امیرکبیر، ۴۵۷ ص.

عبداله پور آزاد، م. و ستاری، م. ۱۳۹۴. پیش‌بینی جریان روزانه رودخانه اهرچای با استفاده از روش‌های شبکه عصبی مصنوعی (ANN) و مقایسه آن با سیستم استنتاجی فازی-عصبی تطبیقی (CANFIS). نشریه پژوهش‌های حفاظت آب و خاک، ۲۲(۱): ۲۸۷-۲۹۸.

عزمی، م. و عراقی نژاد، ش. ۱۳۹۱. توسعه روش رگرسیون k-نزدیکترین همسایگی در پیش‌بینی جریان رودخانه. آب و فاضلاب، ۲: ۱۰۸-۱۱۹.

عظیمی، و.، وکیلی فرد، ع. و اسدی، ا. ۱۳۹۴. ارزیابی برنامه‌ریزی بیان ژن و مدل M5 در برآورد دبی‌های روزانه، مطالعه موردی رودخانه لیقوان. فصلنامه بین المللی پژوهشی تحلیلی منابع آب و توسعه، ۳(۱۱): ۱۳۴-۱۴۲.

غظنفری هاشمی، س. و شهیدی، ا ا. ۱۳۹۱. پیش‌بینی عمق آبشستگی اطراف پایه پل با استفاده از ماشین‌های بردار پشتیبان. مجله عمران مدرس، ۱۲ (۲): ۲۳-۳۶.

غلامی، و. و درواری، ز. ۱۳۹۲. شبیه سازی فرایند بارش-رواناب با بکارگیری شبکه عصبی مصنوعی (ANN) و مدل HEC-HMS (مطالعه موردی حوضه آبخیز کسیلیان). نشریه علوم و مهندسی آبخیزداری ایران، ۷(۲۱): ۶۷-۷۰.

فتح ا... زاده، ط. ۱۳۹٤. بررسی انواع و شدت فرسایش و تولید رسوب در زیر حوضه های آبخیز ناورود. فصلنامه جغرافیای طبیعی، ۸(۲۷): ۲۵-۳۸.

ملکیان، آ.، محسنی ساروی، م. و مهدوی، م. ۱۳۸۳. بررسی کارایی روش شماره منحنی دربرآورد عمق رواناب. نشریه منابع طبیعی، ۵۷(٤): ٦۲۱-٦۳۳.

نگارش، ح.، طاوسی، ت. و مهدی نسب، م. ۱۳۹۱. مدل‌سازی تولید رواناب خوضه آبریز رودخانه کشکان بر اساس روش‌های آماری. دوفصلنامه پژوهش‌های بوم شناسی شهری، ٦: ۸۱-۹۲.

ولی خوجینی، ع. ۱۳۷۷. بررسی شماره منحنی (CN) روش SCS در برآورد عمق رواناب و بده اوج در حوضه‌های آبخیز معرف سلسله جبال البرز. پژوهش و سازندگی، ۳۸: ۱۲-۱۵.

Adamowski, J. 2013. Using support vector regression to predict direct runoff, base flow and total flow in a mountainous watershed with limited data in Uttaranchal, India, Annals of Warsaw University of Life Sciences-SGGW, Land Reclamation No, 45 (1): 71–83

Dingman, S.L. 1994. Physical Hydrology; Prenice Hall. 646 pages.

Eskandarinia, A., Nazarpour, H., Teimouri, M., Ahmadi, M. 2010. Comparison of K-nearest neighbor in daily flow forecasting. Journal of applied sciences, 10(11): 1006-1010

Hall, M. A. 1999. Correlation-based Feature Selection for Machine Learning, phd thesis, University of Waikato.

Hu, C., Wu, Z., Wang, J., Liu, L. 2011. Application of the Support Vector Machine on Precipitation-Runoff Modelling in Fenhe River. International Symposium on Water Resource and Environmental Protection (ISWREP),1099-1103.

Kira, K., and Rendell, L. A. 1992. The Feature Selection Problem: Traditional methods and a new algorithm. AAAI-92 Proceedings of the tenth national conference on Artificial intelligence, 129-134

Kohler, M.A., and Linsly, R.K. 1951. Predicting runoff from storm rainfall. U.S. Weather Bureau, Research, Paper, 34: 1-10.

Kumar Vyas, S., Prakash Mathur, Y., Sharma, G., Ghandvani, V. 2016. Rainfall-Runoff Modelling: Conventional regression and Artificial Neural Networks approach. Recent Advances and Innovations in Engineering (ICRAIE), DOI: 10.1109/ICRAIE.2016.7939532.

Okkan, U., Serbes, Z. A. 2012. Rainfall-Runoff modeling using least squares support vector machines. Journal of Environmentrics, DOI: 10.1002/env.2154

Patel, A.B., Joshi, G.S. 2017. Modeling of Rainfall-Runoff Correlations Using Artificial Neural Network-A Case Study of Dharoi Watershed of a Sabarmati River Basin, India. Civil Engineering Journal, 3(2): 78-87

Sattari, M. T., Apaydin, H., Ozturk, F. 2012. Flow estimations for the Sohu Stream using artificial neural networks. Environmental earth science, 66: 2031-2045

Sattari, M. T., Pal, M., Apaydin, H., Ozturk, F. 2013. M5 Model Tree Application in Daily River Flow Forecasting in Sohu Stream, Turkey. Journal of water resources and the regime of water bodies, 40(3): 233-242

Vafakhah, M., Janizadeh, S., Khosrobeigi Bozchaloei, S. 2014. Application of Several Data-Driven Techniques for Rainfall-Runoff Modeling. Journal of Ecopersia, 2(1): 455-469.

ISSN 2251-7480

Modelling monthly runoff by using data mining methods based on attribute selection algorithms

Mohammad Taghi Sattari[1] and Ali Rezazadeh Joudi[2*]

1) Assistant Professor, Department of Water Engineering, Agriculture Faculty, University of Tabriz, Tabriz, Iran
2) Young Researchers and Elite Club, Maragheh Branch, Islamic Azad University, Maragheh, Iran
[*] Corresponding author email: alijoudi66@gmail.com

Received: 24-01-2017 Accepted: 07-05-2018

Abstract

Given the importance of catchment basin output flow for surface water management, precise understanding of the relationship between the amount of runoff and climatic parameters such as precipitation and temperature is important. therefore the identification of parameters are important in the modeling process. In this paper, after homogeneity tests have been carried out for monthly precipitation, temperature, and runoff data in the Navroud Catchment Basin in Iran, two combinations of effective factors for runoff are considered according to Relief and Correlation algorithms. A new Relief Algorithm first identifies effective features within a set of data in an orderly manner especially when the amount of available data is low. The new method uses a data-related weight vector average and a threshold value. Applying support vector regression and the nearest neighbor method, monthly runoff was modeled based on the two proposed combinations. The results showed that support vector regression approach which utilizes a radial basis function kernel, yields higher accuracy and lower error than the nearest neighbor method for estimating runoff. The improvement is particularly noticeable for flooding situations.

Keywords: data mining, nearest neighbors, runoff modelling, support vector regression