مجله پژوهش آب ایران
سال ۸/ شماره ۱۵/ پاییز و زمستان ۱۳۹۳ (۱۷۷-۱۸۶)

ارزیابی مدل‌های داده‌کاوی در ریزمقیاس نمایی بارش براساس داده‌های مدل گردش عمومی NCEP (مطالعه موردی: ایستگاه سینوپتیک کرمانشاه)

خلیل قربانی[1*]

چکیده

طی سال‌های اخیر روش‌های زیادی برای ریزمقیاس‌نمایی متغیرهای هواشناسی از داده‌های بزرگ مقیاس مدل‌های گردش‌های عمومی جو ارائه شده است. در این پژوهش کارایی سه روش ماشین بردار پشتیبان (SVM)، درخت تصمیم (M5) و نزدیکترین k-همسایگی (KNN) از مدل‌های داده‌کاوی در ریزمقیاس‌نمایی بارش، به کمک پارامترهای خروجی مدل گردش عمومی جو NCEP در ایستگاه کرمانشاه بررسی شد. در دوره آماری ۱۹۹۱-۱۹۶۱ بارش روزانه ایستگاه کرمانشاه و پارامترهای خروجی NCEP مدل‌سازی شد و نتایج آن‌ها برای دوره آماری ۲۰۰۱-۱۹۹۲ آزمایش شد. نتایج این بررسی نشان داد که بارش شبیه‌سازی شده با هر یک از مدل‌های داده‌کاوی، دارای میانگین و انحراف معیار کمتری نسبت به داده‌های مشاهداتی هستند و مقادیر حدی را نمی‌توانند به خوبی پیش‌بینی کنند. با این وجود روش نزدیک‌ترین همسایگی نسبت به دیگر روش‌ها نتایج بهتری را ارائه کرد.

واژه‌های کلیدی: درخت تصمیم، ریزمقیاس نمایی، ماشین بردار پشتیبان، مدل گردش عمومی، نزدیک‌ترین k-همسایگی.

ارجاع: قربانی خ. ۱۳۹۴. ارزیابی مدل‌های داده‌کاوی در ریزمقیاس نمایی بارش براساس داده‌های مدل گردش عمومی NCEP (مطالعه موردی: ایستگاه سینوپتیک کرمانشاه). مجله پژوهش آب ایران. ۱۷۷-۱۸۶.

۱- استادیار گروه مهندسی آب، دانشگاه علوم کشاورزی و منابع طبیعی گرگان.
*** نویسنده مسئول:** ghorbani.khalil@yahoo.com
تاریخ دریافت: ۱۳۹۱/۰۶/۱۵ تاریخ پذیرش:۱۳۹۲/۰۳/۰۶

مقدمه

با توجه به صنعتی شدن و افزایش مصرف سوخت‌های فسیلی، انتشار گازهای گلخانه‌ای در جو افزایش یافته است. گازهای گلخانه‌ای امواج حرارتی مادون قرمز منتشر شده از سطح زمین را جذب می‌کنند و بخشی از آن را دوباره به سطح زمین بر می‌گردانند. این عمل سبب برهم خوردن توازن حرارتی زمین و در نتیجه گرم شدن زمین می‌شود. به دنبال گرمایش جهانی، گردش عمومی جو نیز تحت تأثیر قرار می‌گیرد و سبب تغییراتی در مقدار، شدت، مدت و زمان وقوع بارش در نقطه‌های مختلف کره زمین می‌شود. مدل‌های گردش عمومی جو، با در نظر گرفتن جو زمین به عنوان یک سیال و حل معادلات پیوستگی دینامیک سیال جو در مقیاس‌های گسسته زمانی و مکانی، توسعه یافتند. این مدل‌ها مجموعه معادلاتی هستند که معرف قانون‌های فیزیکی- دینامیکی و آماری حاکم بر جو هستند و به منظور مطالعهٔ رفتار جو یا اقلیم در زمان‌های مختلف تدوین می‌شوند و می‌توانند اقلیم حاضر کره زمین را شبیه‌سازی کنند و تغییرات اقلیمی آینده را پیش‌بینی نمایند (زو، ۱۹۹۹). یکی از مشکلات عمده در استفاده از خروجی مدل‌های گردش عمومی جو، بزرگ مقیاس بودن سلول‌های محاسباتی آن نسبت به منطقه مطالعاتی است. دو نوع تکنیک برای به دست‌آوردن متغیرها در مقیاس محلی (ریزمقیاس نمایی) از روی مقیاس جهانی وجود دارد، یکی روش دینامیکی که شامل حل صریح معادلات دینامیکی سیستم است و دیگری روش آماری که از رابطه‌های استخراج شده از داده‌های مشاهده شده استفاده می‌کند. روش‌های دینامیکی به دلیل پرهزینه بودن و دشواری اجرا به طور معمول در بررسی‌ها و مطالعات دانشگاهی استفاده می‌شوند (هیوتسان و کران، ۱۹۹۶). همچنین این روش‌ها امکان تولید داده برای انواع سناریوهای مختلف را ندارند، در مقابل روش‌های آماری با محاسبات ساده آماری، صرف زمان و هزینه کم امکان بررسی انواع سناریوهای اقلیمی و تحلیل عدم قطعیت آن‌ها را دارند. از معروف‌ترین مدل‌های آماری که برای ریزمقیاس نمایی داده‌های مدل‌های اقلیمی استفاده می‌شود مدل‌های LARS_WG و SDSM[1] هستند که به‌صورت بسته‌های نرم افزاری و رایگان در اختیار هستند.

قابلیت این مدل‌ها در ریزمقیاس نمایی پارامترهای دما و بارش توسط پژوهشگران مختلف ارزیابی شده است. بطور نمونه می‌توان به مطالعات دیبایک و کولیبالی (۲۰۰۶) اشاره کرد که دو روش شبکه‌های عصبی مصنوعی و SDSM را در ریز مقیاس کردن داده‌های بارش و درجه حرارت روزانه در حوضه ساگونوی در شمال ایالت کبک کانادا مقایسه کردند و نتیجه گرفتند شبکه عصبی مصنوعی در ریزمقیاس نمایی داده‌های بیشینه و کمینه دمای روزانه و همچنین بارش روزانه نسبت به روش SDSM کارایی بهتری دارد. چن و همکاران (۲۰۱۲)، ماشین بردار پشتیبان (SVM[2]) و شبکه عصبی مصنوعی را در ریزمقیاس نمایی داده‌های بارش مورد ارزیابی قرار دادند و نتیجه گرفتند ماشین بردار پشتیبان با تابع کرنل پایه شعاعی روش بهتری برای ارزیابی اثرات تغییر اقلیم در هیدرولوژی است. تریپاتی و همکاران (۲۰۰۶) نیز با مقایسه روش‌های شبکه عصبی مصنوعی و ماشین بردار پشتیبان برای ریزمقیاس نمایی بارش ماهانه در کل هندوستان، روش ماشین بردار پشتیبان را مناسب دانستند. سجاد خان و همکاران (۲۰۰۶) سه روش ریزمقیاس‌نمایی شامل مدل آماری SDSM، مدل تولید داده LARS-WG و شبکه عصبی مصنوعی را برای ریزمقیاس‌نمایی متغیرهای بارش روزانه و دماهای حداقل و حداکثر روزانه به کار بردند. آن‌ها از داده‌های ۴۰ ساله مشاهده شده در زیر حوضه چوت‌دو دایابل کانادا و داده‌های مدل گردش عمومی NCEP در دوره آماری ۲۰۰۱-۱۹۶۱ برای مدل‌سازی استفاده کردند. مدل SDSM بهترین نتایج را در بازسازی خصوصیات داده‌های مشاهده شده ارائه کرد و مدل شبکه عصبی مصنوعی از این نظر کمترین کارآیی را داشت. کروفورد و همکاران (۲۰۰۷) برای ریزمقیاس‌نمایی آماری بارش روزانه در شمال ایرلند از نرم‌افزار SDSM استفاده کردند. بررسی آن‌ها نشان داد، نتایج پیش‌بینی مدل SDSM برای فصل‌های بهار و تابستان ضعیف است. آکسورن سینگچای و سرینیلتا (۲۰۱۱) سه روش ریزمقیاس نمایی آماری را برای پیش‌بینی دما و بارش در ۴۵ ایستگاه هواشناسی در تایلند در دوره آماری ۲۰۰۷-۱۹۶۵ مقایسه کردند. ایشان از داده‌های دوباره آنالیز شده آزمایشگاه ژئوفیزیکی

1- Statistical Downscaling Model

2- Support Vector Machine

دینامیک سیال (GFDL[1]) استفاده کردند و نتیجه گرفتند روش ماشین بردار پشتیبان با تابع کرنل پایه شعاعی نسبت به روش‌های ماشین بردار پشتیبان با تابع کرنل چندجمله‌ای و روش رگرسیون چندمتغیره نتیجه بهتری را ارائه می‌کند. برخی دیگر از پژوهشگران روش نزدیکترین K-همسایگی (KNN[2]) را برای ریزمقیاس نمایی استفاده کردند و نتایج آن را با دیگر روش‌های آماری مورد مقایسه قرار دادند. دیپاشری و موجومدار (۲۰۱۱) نیز سه روش میدان تصادفی مشروط (CRF)، نزدیکترین K- همسایگی و ماشین بردار پشتیبان را برای ریزمقیاس نمایی بارش روزانه به کمک داده‌های مدل گردش عمومی NCEP در دوره آماری ۲۰۰۴-۱۹۵۱در ایالت پنجاب مورد مقایسه قرار دادند و نتیجه گرفتند مدل‌های CRF و KNN در مقایسه با مدل SVM از لحاظ آماری نتایج بهتری را ارائه می‌کند. در پژوهشی دیگر قمقامی (۱۳۸۹)، از یک روش ناپارامتری مبتنی بر برآوردگر کرنل برای شبیه‌سازی بارش ماهانه استفاده کرد. نتایج آن‌ها نشان داد که این روش قابلیت شبیه‌سازی مناسب میانگین‌ها و واریانس‌های ماهانه و همچنین وقایع حدی بارش نظیر بارندگی های با دوره بازگشت بالا را دارد. قابلیت بالای مدل‌های رگرسیونی درخت تصمیم موسوم به مدل M5 در مدل‌سازی و ایجاد رابطه‌های بین داده‌ها نیز سبب شده است تا کاربرد آن در مسائل مختلف رو به گسترش باشد. با توجه به مطالب ارائه شده، در این پژوهش کارایی سه روش مدل‌سازی داده‌های پیوسته در داده‌کاوی شامل مدل درخت تصمیم M5، ماشین بردار پشتیبان و نزدیکترین K- همسایگی در ریزمقیاس نمایی بارش روزانه به کمک پارامترهای خروجی حاصل از مدل گردش عمومی جو NCEP برای ایستگاه سینوپتیک کرمانشاه ارزیابی شود.

مواد و روش‌ها

داده‌های مورد استفاده

ایستگاه سینوپتیک کرمانشاه در موقعیت جغرافیایی به طول ۴۷ درجه و ۱۲ دقیقه شرقی و عرض جغرافیایی ۳۷ درجه و ۳۲ دقیقه شمالی و در ارتفاع ۱۳۱۸ متری از سطح دریا قرار دارد. این منطقه با میانگین بارندگی سالانه ۴۲۰ میلی‌متر و میانگین سالانه دمای ۱۳ درجه سانتی‌گراد دارای اقلیم نیمه خشک است. داده‌های بارش روزانه این ایستگاه طی سال‌های ۱۹۶۱ تا ۲۰۰۱ بر طبق داده‌های موجود مدل گردش عمومی جو NCEP استفاده شد. داده‌های NCEP مشتمل بر ۲۶ متغیر و شامل پارامترهایی مانند میانگین دما در ارتفاع ۲ متری، میانگین فشار سطح دریا، رطوبت نسبی و رطوبت ویژه در سطح‌های ۵۰۰، ۸۵۰ هکتوپاسکال و نزدیک سطح زمین می‌باشد.

ریزمقیاس نمایی با استفاده از مدل‌های داده‌کاوی

داده‌کاوی فرآیندی است که ابزارهای مختلف تحلیل داده را به کار می‌گیرد تا الگوها و رابطه‌های فیزیکی متغیرها را در مجموعه داده‌های مختلف کشف کند (تو کراوز، ۱۹۹۹). در داده‌کاوی الگوریتم‌های زیادی برای هدف‌های مختلف مانند طبقه‌بندی و پیش‌بینی استفاده می‌شود. در این پژوهش سه روش مدل درخت تصمیم M5، ماشین بردار پشتیبان و نزدیکترین K-همسایگی مورد استفاده و ارزیابی قرار گرفتند.

مدل درخت تصمیم M5

درخت‌های تصمیم روشی برای نمایش یک سری از قانون‌ها هستند که منتهی به یک رده یا مقدار می‌شوند. درخت‌های تصمیم به کمک جداسازی متوالی داده‌ها به یک سری گروه جداگانه تشکیل شده و سعی می‌شود در فرآیند جداسازی، فاصله بین گروه‌ها افزایش یابد. ساختار یک مدل درختی شامل ریشه، گره‌های داخلی و برگ است. از مدل‌های درخت تصمیم در حل بسیاری از مسائل طبقه‌بندی و رگرسیون استفاده شده است. برای اولین بار کوینلان (۱۹۹۲) مدل درخت تصمیم موسوم به M5 را برای پیش‌بینی داده‌های پیوسته ارائه کرد. این مدل، برخلاف مدل‌های درخت تصمیم معمول که کلاس یا رده‌های گسسته را به عنوان خروجی ارائه می‌کنند، یک مدل خطی چندمتغیره را برای داده‌ها در هر گره از مدل درختی می‌سازد. تشکیل ساختار مدل‌های درخت تصمیم‌گیری شامل مراحل ایجاد درخت و هرس کردن آن است (کوینلان، ۱۹۹۲ و ویتن و فرانک، ۲۰۰۵). در مرحله ساختن درخت، از یک الگوریتم استنتاجی یا معیار تقسیم (انشعاب) برای تولید یک درخت تصمیم استفاده می‌شود. معیار تقسیم برای الگوریتم مدل M5، ارزیابی انحراف معیار مقادیر کلاسی است که به عنوان کمیتی از خطا به یک گره می‌رسد و کاهش مورد انتظار در این خطا را به

1- Geophysical Fluid Dynamics Laboratory
2- K- nearest neighbor

عنوان نتیجه آزمون هر صفت در آن گره محاسبه می‌کند. کاهش انحراف معیار (SDR)[1] از رابطه (۱) به دست می‌آید:

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \qquad (۱)$$

که در آن T بیانگر یک سری نمونه‌هایی است که به گره می‌رسند، T_i بیانگر نمونه‌هایی است که iامین خروجی سری پتانسیلی را دارند و sd بیانگر انحراف معیار است. به دلیل فرآیند انشعاب، داده‌های قرار گرفته در گره‌های فرزند، انحراف معیار کمتری نسبت به گره مادر داشته و بنابراین خالص‌تر هستند. پس از بیشینه‌سازی تمامی انشعاب‌های ممکن، M5 صفتی را انتخاب می‌کند که کاهش مورد انتظار را بیشینه کند. این تقسیم بیشتر ساختار شبه‌درختی بزرگی را تشکیل می‌دهد که سبب بیش‌برازش می‌شود. برای غلبه بر مسئله بیش برازش، درخت تشکیل شده بایستی هرس می‌شود. این کار با جایگزینی یک درخت فرعی با یک برگ انجام می‌شود. بنابراین، مرحله دوم در طراحی مدل درختی شامل هرس کردن درخت رشد یافته و جایگزینی درختان فرعی با توابع رگرسیونی خطی است. این روش تولید مدل درختی، فضای پارامترهای ورودی را به نواحی یا زیر فضاهای کوچکتر تقسیم کرده و در هر کدام از آنها، یک مدل رگرسیونی خطی برازش می‌دهد. بعد از اینکه مدل خطی به دست آمد، برای کمینه کردن خطای تخمین با حذف کردن پارامترها، ساده‌سازی مدل انجام می‌شود. در مدل M5 از یک جستجوی حریصانه برای حذف متغیرهایی که مشارکت کمی در مدل دارند، استفاده می‌شود. البته گاهی وقت‌ها همه متغیرها حذف شده و فقط یک مقدار ثابت باقی می‌ماند (کوینلان، ۱۹۹۲).

ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان توسط واپنیک و چروننکیس (۱۹۷۴) و واپنیک (۱۹۸۲) بر پایه تئوری یادگیری آماری توسعه داده شد و یکی از روش‌های یادگیری نظارت شده است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند. جزییات رگرسیون بردار پشتیبان که استفاده از ماشین بردار پشتیبان در مسائل رگرسیونی است به نقل از باساک و همکاران (۲۰۰۷) به‌صورت زیر است:

فرض کنید داده‌های آموزشی به‌صورت $\{(x_1,y_1),\ldots,(x_l,y_l)\} \subset R^d \times R$ داده شده است که در آن R^d به فضای الگوهای ورودی اشاره دارد. در رگرسیون بردار پشتیبان، هدف پیدا کردن تابع $f(x)$ است به طوری که برای همه داده‌های آموزشی حداکثر به میزان ε از هدف‌های به دست آمده واقعی y_i انحراف دارد و در عین حال تا حد ممکن هموار است. تابع خطی f به شکل زیر بیان می‌شود:

$$f(x) = \langle \omega, x \rangle + b \qquad \omega \in R^d, b \in R \qquad (۲)$$

که عبارت $\langle .,. \rangle$ بیانگر ضرب نقطه‌ای در R^d است. هموار بودن در رابطه بالا به معنی ω کوچک است. بدین منظورلازم است تا نرم اقلیدسی یعنی $\|\omega\|^2$ حداقل شود. این می‌تواند به‌صورت یک مسئله بهینه‌سازی درجه دو به‌صورت زیر نوشته شود:

$$Minimize \frac{1}{2}\|\omega\|^2$$
$$\left.\begin{aligned} y_i - \langle \omega, x_i \rangle - b \le \varepsilon \\ \langle \omega, x_i \rangle + b - y_i \le \varepsilon \end{aligned}\right\} \qquad (۳)$$

مسئله بهینه‌سازی ذکر شده زمانی که مقدار واقعی f و تقریب همه جفت (x_i, y_i) با دقت ε وجود داشته باشد امکان‌پذیر است. گاهی وقت‌ها چندین خطا مجاز است. متغیرهای کمکی ξ_i و ξ^*_i برای غلبه بر قیود عملی نشدن مسئله بهینه‌سازی معرفی می‌شوند، بنابراین فرمول بالا به‌صورت زیر تبدیل می‌شود:

$$Minimize \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi^*_i) \qquad (۴)$$

مقدار ثابت $C \ge 0$ توازنی بین هموار بودن f و خطای تجربی برقرار می‌کند.

$$\left.\begin{aligned} y_i - \langle \omega, x_i \rangle - b \le \varepsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i \le \varepsilon + \xi^*_i \\ \xi_i, \xi^*_i \ge 0 \end{aligned}\right\} \qquad (۵)$$

وپنیک (۱۹۸۲) برای کاربرد ماشین‌های بردار پشتیبان در مسائل رگرسیونی از تابع جدیدی به نام $\varepsilon - insensitive$ استفاده کردکه خطاهایی که در یک فاصله معین از مقادیر واقعی هستند را نادیده می‌گیرد (رابطه ۶)

$$|\xi|_\varepsilon = \left\{\begin{aligned} 0 \quad & if \ |\xi| \prec 0 \\ |\xi| - \varepsilon \quad & otherwise \end{aligned}\right\} \qquad (۶)$$

این تابع خطای کمتر از ε را در نظر نمی‌گیرد. تابع خطا

1- Standard Deviation Reduction

$\varepsilon-insensitive$ به‌عنوان خطای تجربی جایگزین می‌شود. فرمول‌بندی دوگانه، کلیدی برای توسعه بردار پشتیبان برای توابع غیرخطی تهیه می‌کند. روش دوگانه‌سازی از تابع لاگرانژ ضربی توصیف شده به‌صورت زیر استفاده می‌کند:

$$\begin{aligned} L &= \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi^*_i) \\ &-\sum_{i=1}^{l} a_i(\varepsilon + \xi_i - y_i + \langle \omega, x_i \rangle + b) \\ &-\sum_{i=1}^{l} a^*_i(\varepsilon + \xi^*_i + y_i - \langle \omega, x_i \rangle - b) \\ &-\sum_{i=1}^{l}(\eta_i \xi_i + \eta^*_i \xi^*_i) \end{aligned} \qquad (۷)$$

با محاسبه ضریب‌های لاگرانژ، تابع رگرسیون به‌صورت زیر بهینه می‌شود:

$$\begin{aligned} &\omega = \sum_{i=1}^{l}(a_i - a_i^*)x_i \quad \rightarrow \\ &f(x) = \sum_{i=1}^{l}(a_i - a_i^*)\langle x_i, x \rangle + b \end{aligned} \qquad (۸)$$

این معادله توسعه بردار ماشین نامیده می‌شود یعنی اینکه ω می‌تواند بطور کامل بصورت یک ترکیب خطی از نمونه‌های آموزشی توصیف شود. از رابطه بالا مشخص می‌شود که فقط نقطه‌هایی که ضریب‌های لاگرانژ غیرصفر دارند یعنی نقطه هایی که روی خطوط و خارج از حدود $\varepsilon-insensitive$ قرار می‌گیرند در محاسبات $f(x)$ نقش دارند. این نقطه‌ها بردار پشتیبان نامیده می‌شوند.

در روش SVM در حالتی که نتوان یک تابع رگرسیون خطی در فضای ورودی به داده‌ها برازش داد، از یک نگاشت غیرخطی ϕ برای تبدیل داده‌ها به یک فضای بالاتر استفاده می‌شود و سپس در این فضای جدید الگوریتم SVM استاندارد اجرا می‌شود. بنابراین رگرسیون خطی در این فضای جدید متناظر با رگرسیون غیرخطی در فضای ورودی است. انجام محاسبات در فضای ویژگی چون ابعاد بیشتری دارد می تواند پرهزینه باشد. در حالت کلی این فضا بی‌نهایت است. برای غلبه بر این مشکل از تابع هسته (کرنل) استفاده می‌شود. تابع هسته، یک جداکننده خطی متکی بر ضرب داخلی بردارهاست که به صوت $k(x_i, x_j) = x_i^T x_j$ می‌باشد. اگر نقطه های با استفاده از انتقال $\varphi: x \rightarrow \varphi(x)$ به فضای ویژگی (فضای با ابعاد بالاتر) انتقال یابند، ضرب داخلی آن‌ها به‌صورت $k(x_i, x_j) = \varphi(x_i)^T . \varphi(x_j)$ تبدیل خواهد شد.

یک تابع کرنل تابعی است که بر طبق یک ضرب داخلی در فضای ویژگی است. بنابراین از محاسبه $\varphi(x)$ خودداری کرده و به جای آن از تابع کرنل استفاده می‌شود. انتخاب کرنل‌های غیرخطی به ما اجازه‌ی ساخت جداکننده‌های خطی در فضای ویژگی را می‌دهد در صورتی که آن‌ها در فضای اصلی غیرخطی هستند. از مزایای تابع کرنل می‌توان به حل مسائل محاسباتی که دارای ابعاد زیادی هستند، امکان استفاده از ابعاد نامتناهی و کارآمدبودن از نظر زمانی و حافظه اشاره کرد.

با توجه به اینکه الگوریتم خطیSVM فقط به ضرب نقطه‌ای $\langle \phi(x_i), \phi(x_j) \rangle$ بستگی دارد. با جایگزینی توابع هسته به جای این ضرب‌های نقطه‌ای، تابع رگرسیون خطی در فضای جدید به‌صورت زیر نوشته می‌شود:

$$\begin{aligned} &f(x) = \langle \omega, \phi(x) \rangle + b = \\ &\sum_{i=1}^{n}(a_i - a_i^*)\langle \phi(x_i), \phi(x) \rangle + b \\ &\rightarrow f(x) = \sum_{i=1}^{n}(a_i - a_i^*)k(x_i, x) + b \end{aligned} \qquad (۹)$$

بنابراین این روش، کلیدی برای حل مسائل غیرخطی فراهم می‌کند. هر تابع مثبت متقارن $k(x, x')$ که در شرط زیر صدق کند (رابطه ۱۰)، متناظر با یک ضرب نقطه‌ای است و می‌تواند به عنوان یک تابع هسته به کار رود (صادقی‌فر و همکاران، ۱۳۸۹).

$$\int k(x, x')g(x)g(x')dx\,dx' \succ 0 \quad \forall g \in L_2 \qquad (۱۰)$$

متداول‌ترین توابع هسته، توابع چند جمله‌ای، توابع پایه شعاعی گوسی، توابع پایه شعاع نمایی و توابع سیگموئید هستند.

نزدیک‌ترین K-همسایگی (KNN)

یکی دیگر از روش‌های مدل‌سازی در داده‌کاوی، الگوریتم نزدیکترین K-همسایگی است. این الگوریتم جزء روش‌های یادگیری نظارت شده است که هم در طبقه‌بندی و هم در پیش‌بینی استفاده می‌شود. نحوه عملکرد این الگوریتم براساس مشاهدات و نمونه‌ها است. براساس این الگوریتم می‌توان یک نمونه جدید را براساس اکثریت K گروه و دسته که نزدیک‌ترین همسایگی‌ها را با آن نمونه داشته باشند، تقسیم‌بندی کرد. به عبارت دیگر می‌توان گفت این روش K تعداد از الگوهای مشابه را پیدا کرده و براساس آن‌ها ارزش نمونه مورد مطالعه را پیش‌بینی می‌کند. در

روش نزدیکترین K-همسایگی نمونه ناشناخته S در یک کلاس از پیش تعریف شده C_i متعلق به C براساس نمونه‌ها کلاس‌بندی شده قبلی (داده‌های آموزش) طبقه‌بندی می‌شود. زمانی که نمونه S کلاس‌بندی می‌شود، KNN فاصله آن را با همه نمونه‌های موجود در داده‌های آموزشی اندازه‌گیری می‌کند. فاصله اقلیدسی از متداول‌ترین معیارهای اندازه‌گیری فاصله است. سپس تمام مقادیر فاصله مرتب می‌شوند به طوری که $d_i \le d_{i+1},\ i = 1,2,3,...,n$ است. K نمونه با کمترین فاصله به نمونه جدید به عنوان نزدیکترین K-همسایه شناخته می‌شوند و برای کلاس‌بندی نمونه جدید S به کلاس موجود $c_i \varepsilon C,\ 1 \prec i \le n$ استفاده می‌شود. تصمیم طبقه‌بندی به طبیعت داده‌ها بستگی دارد. این الگوریتم جزء روش‌های تنبل به حساب می‌آید به این دلیل که مرحله آموزش را همان زمان که نمونه جدید باید طبقه‌بندی شود اجراء می‌کند برخلاف الگوریتم‌های یادگیری، نقطه مقابل آن که داده‌های آموزشی را قبل از ورود نمونه جدید طبقه‌بندی می‌کنند. بر این اساس KNN نسبت به دیگر الگوریتم‌های یادگیری محاسبات بیشتری را لازم دارد. KNN برای داده‌های پویا و داده‌هایی که سریع تغییر می‌کنند و به روز می‌شوند مناسب است (زهور و همکاران، ۲۰۰۹).

ارزیابی مدل‌ها

بعد از مرتب‌سازی داده‌ها و تشکیل جدول اطلاعاتی، متغیر بارش روزانه در ایستگاه کرمانشاه به عنوان متغیر وابسته و داده‌های مدل گردش عمومی NCEP به عنوان متغیر مستقل به مدل‌های داده‌کاوی معرفی شدند. دوره آماری ۱۹۶۱-۱۹۹۱، به عنوان دوره آموزش و دوره ۲۰۰۱-۱۹۹۲ به عنوان دوره آزمون در نظر گرفته شدند. با در نظر گرفتن مقادیر مشاهده شده ($Z(x_i)$) و مقادیر پیش‌بینی شده ($\hat{Z}(x_i)$) توسط هر یک از مدل‌ها بعد از اجرای آن‌ها، از دو معیار ریشه میانگین مربعات خطا (RMSE[1]) و میانگین خطای اریبی (MBE[2]) برای ارزیابی مدل‌ها استفاده شد. معیار RMSE بزرگی خطا و معیار MBE میزان انحراف از خط نیمساز را نشان می‌دهد.

$$RMSE = \left(\sum_{i=1}^{n} (\hat{Z}(x_i) - Z(x_i))^2 / n \right)^{1/2} \qquad (۱۱)$$

$$MBE = \sum_{i=1}^{n} (\hat{Z}(x_i) - Z(x_i)) / n \qquad (۱۲)$$

ابتدا مدل‌سازی با داده‌های روزانه انجام شد و تحلیل نتایج با مقادیر روزانه و ماهانه مستخرج شده از نتایج داده‌های روزانه انجام شد.

نتایج و بحث

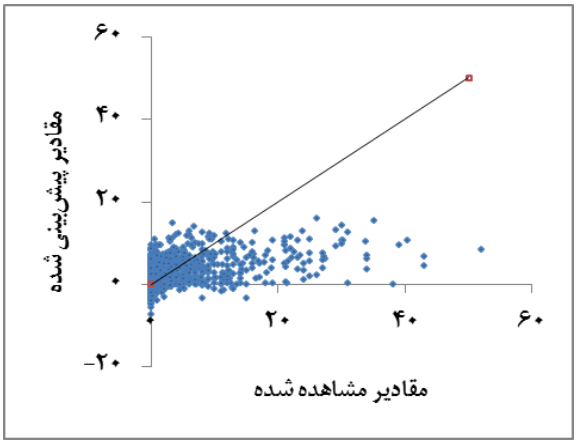
در مدل‌سازی داده‌ها ابتدا نتایج روزانه تحلیل شدند و مقادیر پیش‌بینی شده بارش روزانه با هر یک از مدل‌های مورد مطالعه در مقابل مقادیر مشاهده ترسیم شدند (شکل ۱). از خط نیمساز نیز برای میزان انحراف مقادیر پیش‌بینی شده از مقادیر مشاهده شده استفاده شد. نتایج نشان می‌دهد تمامی مدل‌های مورد مطالعه در این پژوهش در پیش‌بینی بارش‌های شدید روزانه ضعیف عمل می‌کنند و برآورد کمی را نشان می‌دهند (جدول ۱). با مشاهده شکل ۱ مشخص می‌شود که مدل KNN حداکثر تا ۲۱ میلی‌متر، مدل M5 حداکثر تا ۱۶ میلی‌متر و مدل SVM حداکثر تا ۱۱ میلی‌متر را برای بارش روزانه پیش‌بینی می‌کنند در حالی‌که تا ۵۷ میلی‌متر بارش روزانه در ایستگاه مورد مطالعه گزارش شده است. همچنین بجز مدل KNN خروجی بقیه مدل‌ها مقادیر منفی را نیز برای بارش روزانه ارائه می‌کنند که این می‌تواند به دلیل خاصیت برون‌یابی و تعمیم در رگرسیون باشد که در روش ماشین بردار پشتیبان و مدل درختی M5 که از رگرسیون استفاده می‌کنند دیده شده است.

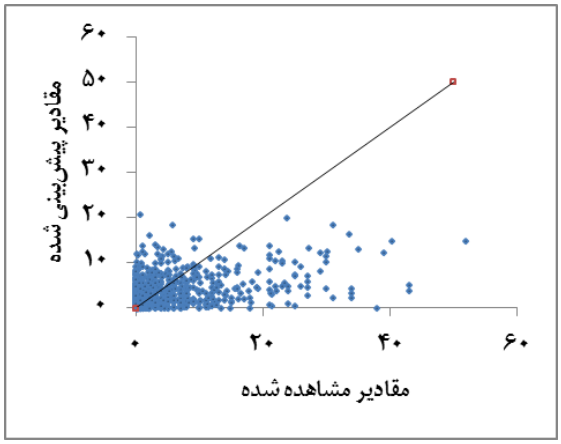
نتایج تحلیل‌های آماری مقادیر ماهانه بارش (شکل ۲ و جدول ۲) نیز نشان از کم برآورد تمامی مدل‌های مورد مطالعه دارد و با توجه به میانگین خطای اریبی مشخص می‌شود مدل ماشین بردار پشتیبان نسبت به دو مدل دیگر برآورد کمتری را نشان می‌دهد و مدل KNN نسبت به بقیه مدل‌ها با کمترین خطا نتایج بهتری را به دست می‌دهد.

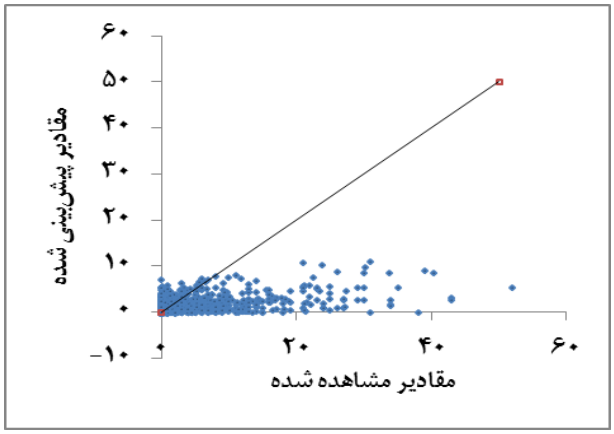
میانگین و انحراف معیار ماهانه بارش به تفکیک ماه‌های سال ترسیم شدند (شکل‌های ۳ و ۴). نتایج نشان می‌دهد که در تمام ماه‌های سال نتایج پیش‌بینی بارش توسط مدل‌ها دارای میانگین و انحراف معیار کمتری نسبت به مقادیر مشاهده شده (P) است. از کم بودن انحراف معیار

1- Root Mean Square Error
2- Mean Bias Error

مقادیر پیش‌بینی شده نسبت به مقادیر مشاهده شده همچنین با توجه به اینکه حد پایین بارش صفر است می‌توان نتیجه گرفت که مقادیر حدی بیشینه بارش با این مدل‌ها به خوبی پیش‌بینی نشده است. پایین بودن میانگین و انحراف معیار بارش روزانه شبیه‌سازی شده به کمک مدل‌های داده‌کاوی مورد استفاده در این پژوهش با نتایج مدرسی (۱۳۸۸) که از مدل کوچک مقیاس‌سازی K-NN در بررسی اثر تغییر اقلیم بر دبی حداکثر در حوضه آبریز گرگان‌رود استفاده کرده بود و نتایج آکسورن سینگچای و سرینیلتا (۲۰۱۱) هماهنگی دارد.



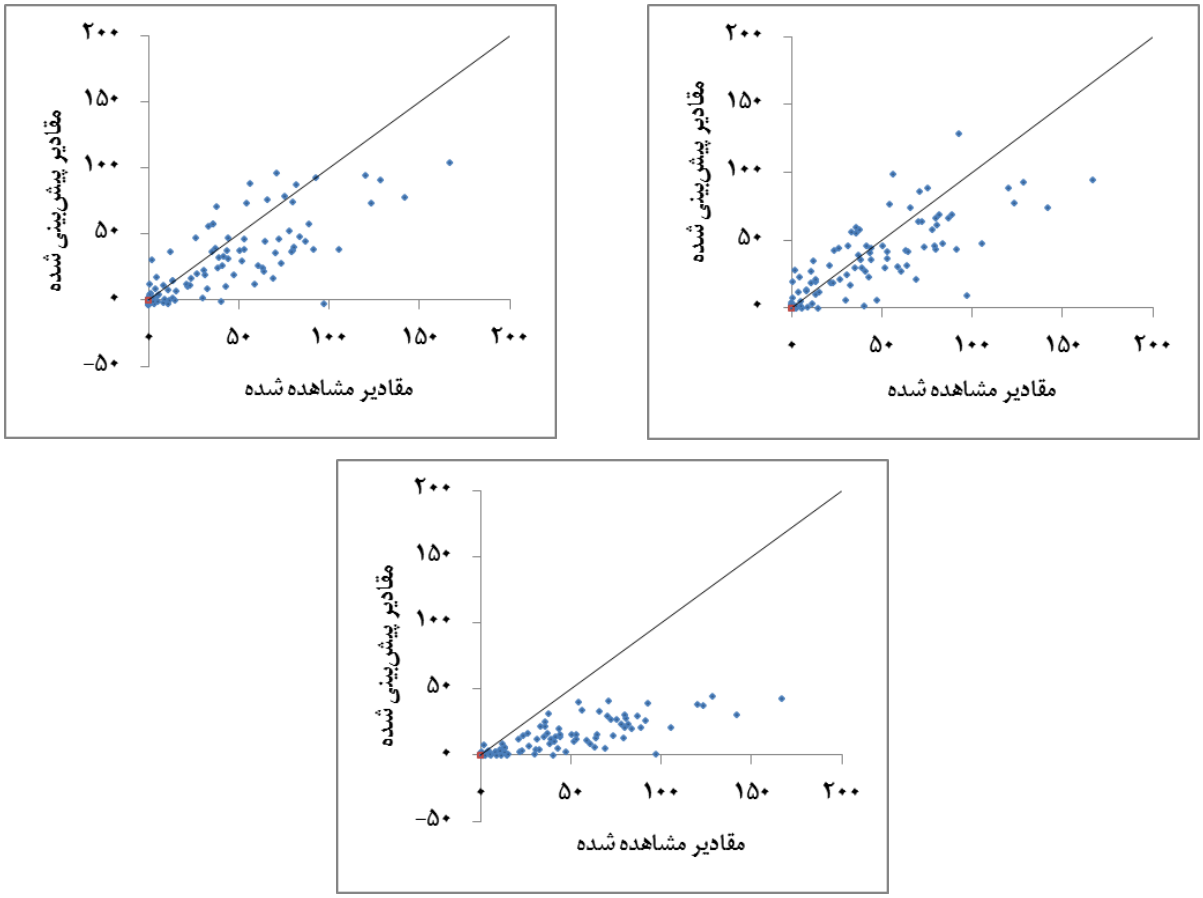




شکل ۱- مقادیر مشاهده شده و پیش‌بینی شده بارش روزانه (برحسب میلی‌متر) به روش KNN (بالا سمت راست)، روش M5 (بالا سمت چپ) و روش SVM (پایین) در مقابل خط نیمساز

جدول ۱- نتایج تحلیل آماری مدل‌های پیش‌بینی بارش روزانه

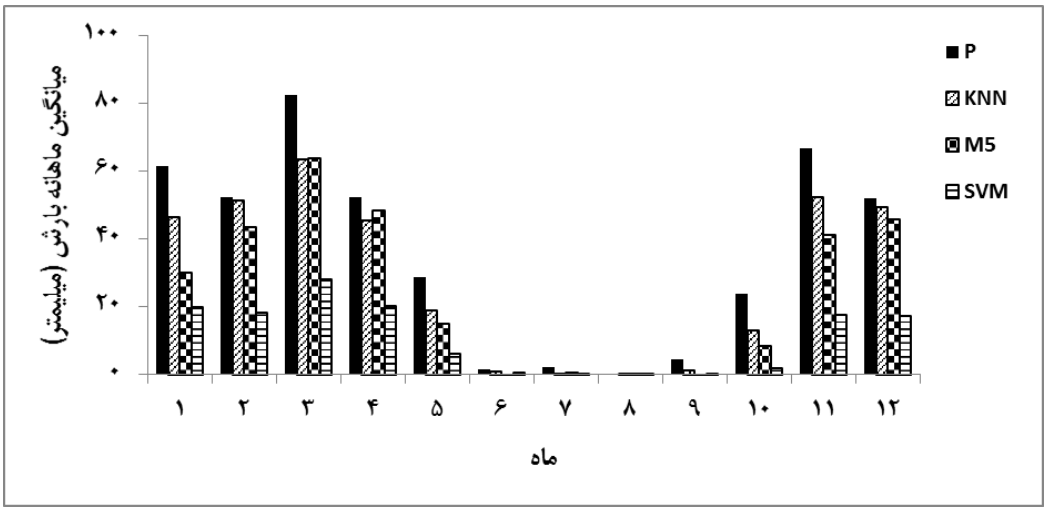
	نزدیکترین K-همسایگی	مدل درختیM5	ماشین بردار پشتیبان
ریشه میانگین مربعات خطا (RMSE)	۳/۴۳	۳/۴۲	۳/۶۸
میانگین خطای اریب (MBE)	-۰/۲۲	-۰/۳۵	-۰/۸۰
ضریب تبیین (R^2)	۰/۳۱	۰/۳۲	۰/۳۶



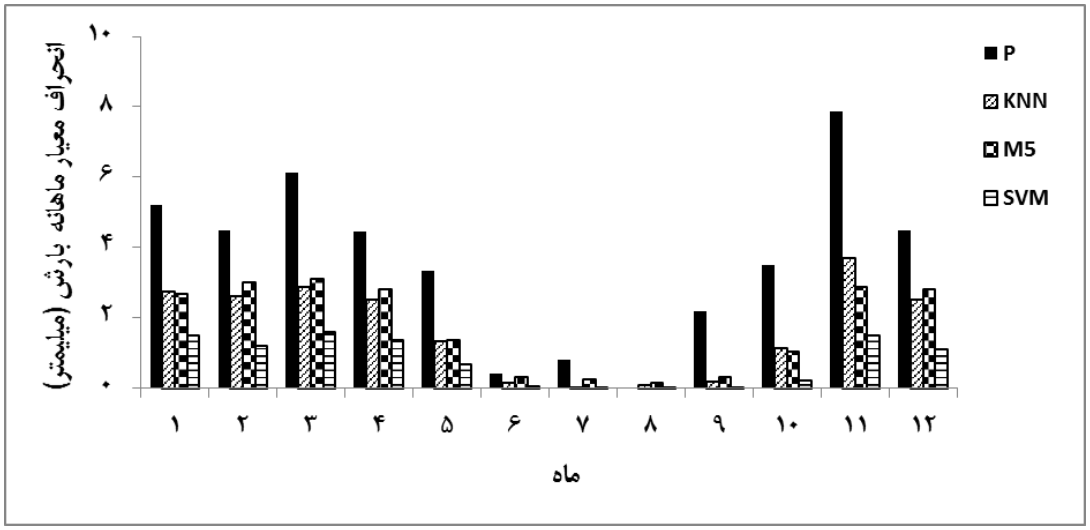
شکل ۲ - مقادیر مشاهده شده و پیش‌بینی شده میانگین بارش ماهانه (برحسب میلی‌متر) به روش KNN (بالا سمت راست)، روش M5 (بالا سمت چپ) و روش SVM (پایین) در مقابل خط نیمساز

جدول ۲ - نتایج تحلیل آماری مدل‌های پیش‌بینی بارش ماهانه

	نزدیکترین K-همسایگی	مدل درختیM5	ماشین بردار پشتیبان
ریشه میانگین مربعات خطا (RMSE)	۲۱/۷	۲۳/۹	۳۶/۱
میانگین خطای اریب (MBE)	-۶/۱	-۹/۸	-۲۳/۲
ضریب تبیین (R2)	۰/۶۸	۰/۶۵	۰/۶۹



شکل ۳- میانگین بارش ماهانه مشاهده شده (P) و پیش‌بینی شده توسط مدل در دوره ۲۰۰۱- ۱۹۹۲



شکل۴ - انحراف معیار بارش ماهانه مشاهده شده (P) و پیش‌بینی شده توسط مدل در دوره ۲۰۰۱- ۱۹۹۲

نتیجه‌گیری

در این پژوهش کارایی سه مدل KNN، M5 و SVM در مدل‌سازی رابطه بین بارش روزانه ایستگاه هواشناسی کرمانشاه با داده‌های خروجی مدل NCEP ارزیابی شد که نشان از ضعف این مدل‌ها در ریز مقیاس نمایی بارش روزانه دارد. تمام این مدل‌ها مقادیر بارش را کمتر از مقادیر واقعی برآورد می‌کنند به طوری که میانگین و انحراف معیار بارش ماهانه پیش‌بینی شده با هر یک از این مدل‌ها از مقادیر مشاهده شده در سطح ایستگاه پایین‌تر است. با این وجود در بین مدل‌های مورد مطاله مدل KNN نسبت به مدل‌های M5 و SVM در مقیاس ماهانه نتایج بهتری را به دست می‌دهد.

منابع

۱. قمقامی م. ۱۳۸۹. ارزیابی و مقایسه تحلیلی مدل‌های ناپارامتری و پارامتری تولید داده‌های هواشناسی. پایان نامه کارشناسی ارشد. دانشگاه تهران. ۱۲۸ ص.

۲. مدرسی ف. ۱۳۸۸. بررسی اثر تغییر اقلیم بر دبی حداکثر. پایان‌نامه کارشناسی ارشد. دانشگاه تهران. ۱۷۶ ص.

۳. صادقی‌فر م. بزرگ‌نیا ا. و زهره‌وند ن. ۱۳۸۹. کاربرد مدل آمیخته ARIMA-SVM در پیش‌بینی کوتاه مدت قیمت نفت، چهارمین کنفرانس داده‌کاوی ایران، دانشگاه صنعتی شریف. ۷ ص.

4. Aksornsingchai P. and Srinilta Ch. 2011. Statistical downscaling for rainfall and temperature prediction in Thailand. Proceeding of the international multi conference of engineers and computer scientists, 6 p.
5. Basak D. Pal S. and Patranabis D. CH. 2007. Support Vector Regression. Neural Information Processing – Letters and Reviews. 11(10):203-224.
6. Chen H. Yu Xu C. and Guo Sh. 2012. Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. Journal of hydrology, 434-435: 36-45.
7. Crowford T. Betts N. and Mortlock D. F. 2007. GCM grid-box choice and predictor selection associated with statistical downscaling of daily precipitation over Northern Ireland. Journal of Climate Research. 34:145-160.
8. Deepashree R. and Mujumdar P. 2011. A comparison of three methods for downscaling daily precipitation in the Punjab region. Hydrological Processes. 25(23):3575–3589.
9. Dibike B. Y. and Coulibaly P. 2006. Temporal neural networks for downscaling climate variability and extremes. Journal Neural Networks. 19:135-144.
10. Hewitson B. C. and Crane R. G. 1996. Climate downscaling: Techniques and applications. Climate Research. 7:85-95.
11. Quinlan J. R. 1992. Learning with continuous classes. Proceedings of Fifth Australian joint conference on artificial intelligence, Singapore. 343-348.
12. Sajjad Khan M. Coulibaly P. and Dibike Y. 2006. Uncertainty analysis of statistical downscaling methods. Journal of hydrology. 319(4):357-382.
13. Tripathi S. Srinivas V. and Nanjundiah R. S. 2006. Downscaling of precipitation for climate change scenarios: A support vector machine approach. Journal of Hydrology. 621-640.
14. Two Crows Corporation. 1999. Introduction to datamining and knowledge discovery, third edition. Available at: www.twocrows.com. 36 p.

15. Two Crows Corporation. 1999. Introduction to datamining and knowledge discovery, third edition. Available at: www.twocrows.com. 36 p.
16. Vapnik V. 1982. Estimation of dependences based on empirical data. Springer-Verlag, 400 p.
17. Vapnik V. and Chervonenkis A. 1974. Theory of pattern recognition. Nauka, Moscow. 353 p.
18. Witten I. H. and Frank E. 2005. Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann:San Francisco. 664 p.
19. Xu C.y. 1999. From GCMs to river flow: A review of downscaling methods and hydrologic modeling approaches. Progress in physical Geography. 23(3):229-249.
20. Zahoor J. Abrar M. Bashir Sh. And Mirza A. 2009. Seasonal to inter-annual climate prediction using data mining KNN technique. Wireless Networks, Information Processing and Systems, Communications in Computer and Information Science. 20:40-51.