نشریه آبیاری و زهکشی ایران
شماره۳، جلد ۱۲، مرداد - شهریور ۱۳۹۷، ص. ۵۸۷-۵۷۹

Iranian Journal of Irrigation and Drainage
No. 3, Vol. 12, Agu.-Sep. 2018, p. 579-587

بررسی کارایی مدل KNN و درخت M5 در مدل‌سازی جریان رودخانه: مطالعه موردی ایستگاه سرمو

آتنا خلیلی نفت چالی[1*]، حسین خزیمه نژاد[2]

تاریخ دریافت: ۱۳۹۶/۹/۲۰ تاریخ پذیرش: ۱۳۹۶/۱۰/۲۵

چکیده

پیش‌بینی دقیق دبی جریان نکته کلیدی در برنامه‌ریزی و مدیریت بهینه منابع آب به شمار می‌آید. حوضه‌ی گرگانرود، یکی از حوضه‌های بخش شمالی کشور و واقع در استان گلستان است .تاثیر خشکسالی و ترسالی بر نوسانات پایه و جریان کل رودخانه، نقش عمده‌ای را در برنامه‌ریزی بهره‌برداری از منابع آب حوضه دارد. در این تحقیق از مدل KNN و درخت تصمیم M5 به‌عنوان یکی از شیوه‌های داده‌کاوی برای برآورد دبی جریان رودخانه گرگانرود واقع در ایستگاه سرمو بهره گرفته‌شد. بدین منظور از داده‌های بارندگی و دبی جریان ایستگاه سرمو واقع در محمدآباد تحت پنج سناریوی مختلف و با اعمال توابع انتقال بر روی داده‌ها، بهره گرفته‌شد. نتایج نشان داد که مدل درخت تصمیم M5 در اکثر مواقع بر مدل KNN برتری دارد و پیش‌بینی دقیق‌تری را حاصل می‌نماید. همچنین در میان سناریوهای تعریف شده، مدل b و c که به ترتیب شامل تمامی داده‌ها و داده‌های بارندگی روزانه، بارندگی روز قبل و بارندگی دو روز قبل می‌باشند، تحت تابع انتقال میانگین متحرک پنج روزه با داشتن معیار R^2 برابر ۰/۹۹۹ و معیار MAE و RMSE برابر ۰/۰۰۱ دقیق‌ترین برآورد را نتیجه می‌دهد.

واژه‌های کلیدی: توابع انتقال، داده‌کاوی، دبی، درخت تصمیم، سناریو

مقدمه

در زمینه‌ی مهندسی آب به‌ویژه در مواردی که مسئله‌ی برداشت آب از رودخانه مطرح است، آگاهی از کمیت آب اهمیت ویژه‌ای دارد. اگرچه بعضی زیر حوضه‌ها دارای واحدهای اندازه‌گیری به‌منظور ثبت پیوسته جریان هستند، ولی بعضا مهندسین با حوضه‌هایی مواجه می‌شوند که فاقد اطلاعات مورد نیاز همچون آبدهی هستند و یا این اطلاعات دارای خلأها و خدشه‌های آماری و محدودیت‌های دوره‌های آماری می‌باشد. پیش‌بینی دبی جریان در یک مقطع از رودخانه از دیرباز موردتوجه هیدرولوژیست‌ها و کارشناسان مهندسی رودخانه بوده و روش‌های متعددی برای این کار تاکنون استفاده شده‌است (کریمی و همکاران، ۱۳۸۸).

در علم آمار روش‌های مختلفی برای دسته‌بندی، شناخت الگوها و پیش‌بینی و مدل‌سازی داده‌ها وجود دارد که در یک نگاه کلی می‌توان این روش‌ها را به دو دسته پارامتری و ناپارامتری تقسیم‌بندی نمود. در مدل‌های آماری پارامتری مانند مدل‌های رگرسیون‌های خطی و غیرخطی، پارامترهای مدل توسط روش‌های مختلف در مرحله‌ی واسنجی، تخمین زده می‌شوند. در مدل‌های ناپارامتری، مرحله‌ی تخمین پارامترها وجود ندارد (Lall and sharma., 1996). روش‌های داده‌کاوی، روش‌های مدل کردن رابطه‌ی نهفته در داده‌ها هستند که به‌صورت خودکار به دسته‌بندی مجموعه داده‌ها (معمولا مجموعه‌های بزرگ) و کشف ارتباط نهفته در بین آن‌ها به‌منظور قابل فهم شدن و در نتیجه سودمند شدن آن‌ها می‌پردازند (Hand et al., 2001). استفاده از تکنیک‌های غیر پارامتریک کارایی چشمگیری در بهبود تخمین‌های صورت‌گرفته خواهدداشت (Twarakavi et al., 2009). رویکرد k–نزدیک‌ترین همسایه، یکی از مهم‌ترین و توسعه یافته‌ترین رویکردهای غیرپارامتریک می‌باشد که در بسیاری از پژوهش‌های نوین جهت تشخیص الگو و کلاسه‌بندی‌های آماری به کار گرفته شده‌است (شریف آذری و عراقی‌نژاد، ۱۳۹۲). نقدی و همکاران (۱۳۹۲) در مقاله‌ای با استفاده از روش غیرپارامتریک K- نزدیک‌ترین همسایگی (KNN) ضمن پیش‌بینی جریان ورودی به مخزن سد زاینده‌رود، روش‌های تخمین K و همچنین افق زمانی مدل پیش‌بینی در روش KNN تحلیلی نموده‌اند.

۱- دانشجوی دکتری مهندسی علوم آب، دانشگاه بیرجند
۲- استادیار گروه مهندسی آب، دانشکده کشاورزی، دانشگاه بیرجند
(* - نویسنده مسئول: (Atenakhalili_2014@yahoo.com

نتایج نشان می‌دهند که با بهره‌گیری از روش وزن‌دهی معکوس نمایی فاصله و نیز انتخاب افق زمانی ۶ ماهه، می‌توان به حداقل درصد حجم خطا در پیش‌بینی ماهانه جریان دست‌یافت. عزمی و عراقی نژاد (۱۳۹۱) از روش KNN به‌منظور پیش‌بینی جریان رودخانه در حوضه بالادست سد زاینده‌رود استفاده کردند و این روش را برای سری‌های تاریخی بلندمدت مناسب دانستند.

درختان تصمیم ابزاری است که قابلیت پاسخ‌گویی به مسائل پیچیده و غیرخطی را دارد (طالبی و اکبری، ۱۳۹۲). لاندهه و دیکسیت مدل درختی M5 را برای پیش‌بینی جریان رودخانه در یک روز قبل در دو ایستگاه، رودخانه نارمادا و دیگری در حوضه رودخانه کریشنا در هند به کار بردند (Londhe and Dixit., 2011). ستاری و همکاران توانایی مدل‌های درختی M5 و بردارهای ماشین در پیش‌بینی جریان رودخانه سوهو در ترکیه را مورد مطالعه قرار داده و نشان دادند که پیش‌بینی‌های مدل درختی M5 تطابق بهتری با داده‌های مشاهداتی داشته است (Sattari et al., 2013).

در این مطالعه ابتدا عملکرد دو مدل درخت تصمیم M5 و مدل KNN در پیش‌بینی دبی رودخانه تحت سناریوهای مختلف در منطقه مورد مطالعه بررسی گردید. سپس از یکسری توابع انتقال بر روی‌داده‌ها استفاده شد و در نهایت بهترین سناریو برای پیش‌بینی دبی رودخانه پیشنهادگردید.

مواد و روش‌ها

معرفی روش مورد استفاده

درخت تصمیم M5: درخت‌های تصمیم روشی برای نمایش یک سری از قوانین هستند که منتهی به یک رده یا مقدار می‌شوند . درخت‌های تصمیم به کمک جداسازی متوالی داده‌ها به یک سری گروه مجزا تشکیل‌شده و سعی می‌شود در فرآیند جداسازی، فاصله‌ی بین گروه‌ها افزایش یابد (ظهیری و قربانی،۱۳۹۲). اولین مرحله برای ایجاد یک مدل درختی، استفاده از یک معیار انشعاب است. معیار انشعاب برای الگوریتم M5 بر اساس عملکرد انحراف استاندارد مقادیر هر کلاس و یا طبقه است که در هر گره به‌دست آمده است. معیار انشعاب بیانگر میزان خطا در آن گره می‌باشد و مدل حداقل خطای مورد انتظار را به‌عنوان نتیجه‌ی آزمایش هر صفت در آن گره محاسبه می‌کند. خطای مدل عموما با اندازه‌گیری دقت پیش‌بینی مقادیر هدف موارد دیده نشده سنجش می‌شود (ستاری و نهرین، ۱۳۹۲).

کاهش انحراف معیار(SDR) از رابطه ۱ به دست می‌آید (Quinlan., 1992).

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

که T بیانگر یک سری نمونه‌هایی است که به گره می‌رسد، Ti بیانگر نمونه‌هایی است که i امین خروجی سری پتانسیلی را دارند و sd بیانگر انحراف معیار است.

روش نزدیک‌ترین همسایه KNN: در این روش به این شرح است که با مشاهده متغیرهای مستقل در زمان واقعی، مدل به جستجوی الگوهای مشابه شرایط فعلی در سری تاریخی می‌پردازد. وقایعی که در سری تاریخی در این الگوها پیش آمده‌اند می‌توانند به عنوان گزینه‌های محتمل در شرایط فعلی در نظر گرفته شوند. احتمال وقوع هر یک از این حالت‌ها در شرایط حاضر، بستگی به شباهت بردار متغیرهای مستقل فعلی با بردار متغیرهای مستقل مشاهداتی در سری تاریخی دارد (عزمی و عراقی نژاد، ۱۳۹۱).

برای انتخاب نمونه‌های مشابه از فاصله اقلیدسی استفاده می‌شود، بدین‌صورت که هر نمونه از بانک داده که با نمونه هدف (مجهول) کم‌ترین فاصله یا به‌عبارت‌دیگر بیش‌ترین تشابه از نظر ویژگی را داشته باشد انتخاب و وزن دهی می‌شود. جاگتاپ و همکاران رابطه‌ی فیثاغورث را در محاسبه‌ی فاصله‌ی اقلیدسی توصیه‌کردند (رابطه ۲) (Jagtap et al., 2004).

$$D(X,Y) = \sqrt{\sum_{i-1}^{nf} (X_i - Y_i)^2} \quad (۲)$$

که در آن X نماینده‌ی نمونه‌ای از داده‌ها با چند پارامتر مشخص (x_1 تا x_n) در بانک مرجع و Y نمونه داده هدف با همان تعداد پارامتر (y_1تا y_n) می‌باشد.

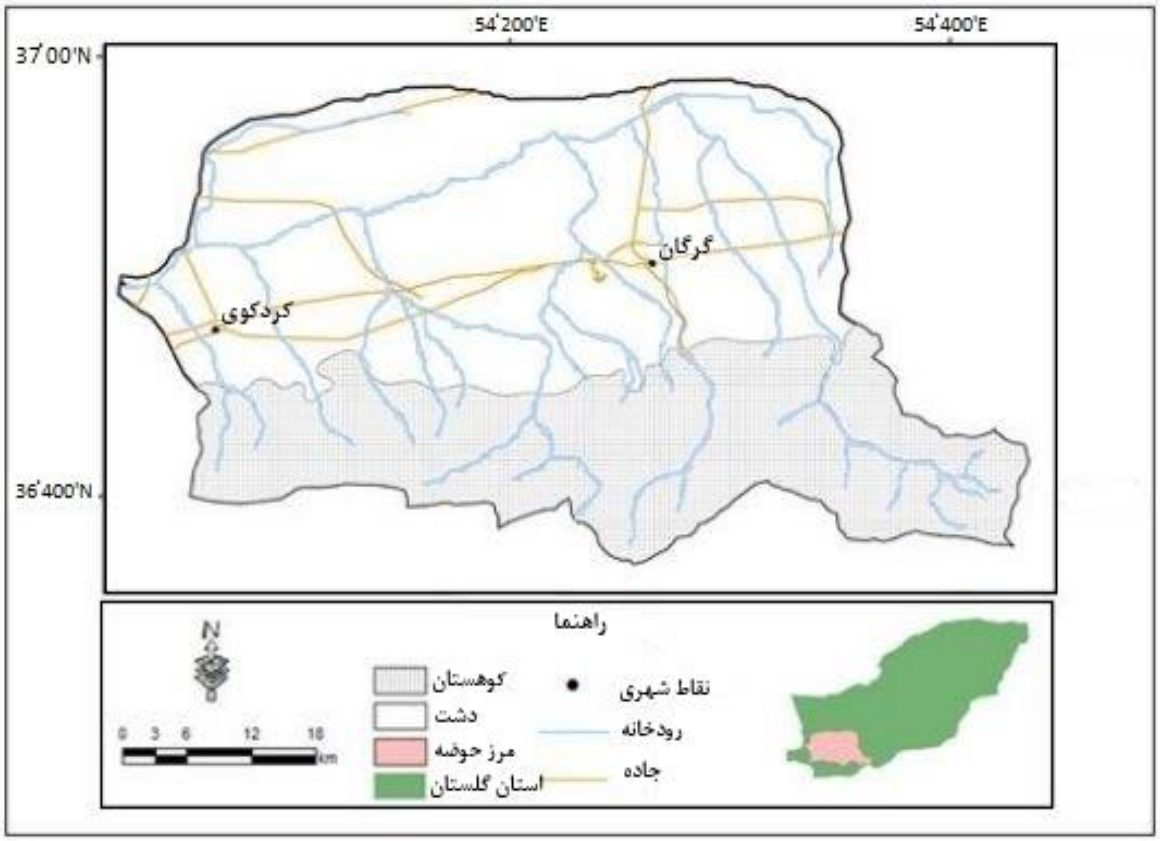
$$X = (x_1, x_2, x_3, \dots, x_n) \quad (۳)$$

$$X = (y_1, y_2, y_3, \dots, y_n) \quad (۴)$$

همسایه‌های نزدیک‌تر به هدف، سهم بیش‌تری نسبت به همسایه‌های دورتر به هدف دارند. روش وزن‌دهی به هر همسایه به صورت 1/d است که d فاصله هدف تا همسایه می‌باشد.

منطقه مورد مطالعه

حوضه‌ی آبریز گرگانرود در شمال شرق کشور به موقعیت جغرافیایی ۵۴ درجه تا ۵۶ درجه و ۲۹دقیقه طول شرقی و ۳۶ درجه و ۳۶ دقیقه تا ۳۷ درجه و ۴۷ دقیقه واقع شده است (شکل ۱). در این تحقیق از اطلاعات ایستگاه سرمو واقع در موقعیت جغرافیایی ۵۴ درجه و ۴۹ دقیقه طول شرقی و ۳۶ درجه و ۴۹ دقیقه می‌باشد و براساس روش کوپن دارای اقلیم مدیترانه‌ای مرطوب جنب‌حاره‌ای می‌باشد.



شکل ۱- حوضه آبریز محدوده مورد مطالعه (حسین‌زاده و همکاران، ۱۳۹٤)

جمع‌آوری داده‌ها و تجزیه‌وتحلیل آن‌ها

در این مطالعه از داده‌های بارندگی و دبی روزانه سال ۱۳۵۴ تا ۱۳۸۶ ایستگاه سرمو برای پیش‌بینی دبی استفاده شده‌است. برای استفاده از داده‌های مذکور در مدل KNN و درخت تصمیم M5، ابتدا داده‌ها به دو دسته تقسیم شدند. ۷۰ درصد داده‌ها برای آموزش و ۳۰ درصد باقیمانده برای آزمون به‌کار گرفته شد. خصوصیات آماری پارامترهای مورد استفاده در این تحقیق در جدول ۱ آمده‌است.

جدول ۱- خصوصیات آماری پارمترهای مورد استفاده

پارامتر	کمینه	بیشینه	میانگین	انحراف از معیار
بارندگی روزانه (میلی‌متر)	۰	۱۵۸	۲/۰۷	۱/۱۹
دبی روزانه (مترمکعب بر ثانیه)	۰	۵۲	۶/۸۹	۱/۵۵

به‌منظور مقایسه نتایج حاصل از الگوریتم‌های تنبل KNN و مدل درختی M5 در پیش‌بینی میزان دبی، پنج سناریو با توجه به پارامترهای تاثیرگذار بر دبی آب تعریف شد. بهترین سناریو موردنظر برای پیش‌بینی دبی آب برای تمامی ماه‌های سال با استفاده از روش آزمون‌وخطا انجام شد. در جدول ۲ پارامترهای ورودی در هر سناریو ارائه گردید. برای بهره‌گیری از الگوریتم KNN و مدل درختی M5 از نرم‌افزار Weka 3.7 استفاده شده است. این نرم‌افزار مجموعه‌ای از بروزترین الگوریتم‌های ماشینی و ابزارهایی برای پیش‌پردازش داده‌ها می‌باشد. با توجه به این‌که کلیه امکانات Weka در قالب واسط‌های کاربری می‌باشند، کاربران می‌توانند متدهای مختلف را بر روی داده-های خود پیاده‌سازی کرده و بهترین الگوریتم را انتخاب نمایند.

این نرم‌افزار پشتیبانی ارزشمندی را برای کل فرآیند داده‌کاوی-های تجربی فراهم می‌نماید. این پشتیبانی‌ها، آماده‌سازی داده‌های ورودی، ارزیابی آماری چارچوب‌های یادگیری و نمایش گرافیکی داده‌های ورودی و نتایج یادگیری را دربر می‌گیرند. همچنین، هماهنگ با دامنه وسیع الگوریتم‌های یادگیری، این نرم‌افزار شامل ابزارهای متنوع پیش پردازش داده‌ها است. این جعبه ابزار متنوع و جامع، از طریق یک واسط متداول در دسترس است، به نحوی که کاربر می‌تواند روش‌های متفاوت را در آن با یکدیگر مقایسه کند و روش‌هایی را که برای مسایل مدنظر مناسب‌تر هستند، تشخیص دهد.

محیط این نرم‌افزار، شامل روش‌هایی برای همه مسایل استاندارد داده‌کاوی مانند رگرسیون، رده‌بندی، خوشه‌بندی، کاوش قواعد انجمنی و انتخاب ویژگی می‌باشد.

جدول۲- پارامترهای ورودی در سناریوهای مختلف

سناریو	پارامترهای ورودی
a	P_t, Q_t
b	$P_{t-2}, P_{t-1}, P_t, Q_{t-2}, Q_{t-1}, Q_t$
c	$P_{t-2}, P_{t-1}, P_t, Q_t$
d	$P_t, Q_{t-2}, Q_{t-1}, Q_t$
e	$P_{t-1}, P_t, Q_{t-1}, Q_t$

Pt-1: بارندگی در روز قبل: میلی‌متر (mm) , Pt-2 : بارندگی در دو روز قبل : میلی‌متر (mm) , Pt: بارندگی روزانه: میلی‌متر (mm)، Qt-1: دبی در روز قبل: مترمکعب بر ثانیه (m3/s) , Qt-2 : دبی در دو روز قبل: مترمکعب بر ثانیه (m^3/s) , Qt : دبی روزانه: مترمکعب بر ثانیه (m^3/s)

پیش‌پردازش اطلاعات متغیرهای پیش‌بینی کننده (تابع انتقال)

الف) استفاده از اصل داده‌ها: در این حالت بر روی داده‌ها هیچ‌گونه عملیات تغییرمقیاس انجام نشد. ورودی و خروجی همان داده‌های خام بودند که به مدل معرفی‌شدند.

ب) استفاده از داده‌های نرمال: در این روش با توجه به میانگین و انحراف معیار داده‌ها، مقادیر با یک مقیاس متناسب کوچک می‌شوند. این رابطه را می‌توان همان رابطه‌ی معروف نرمال‌سازی نیز نامید که به‌صورت زیر تعریف می‌شود:

$$X_i = \left(\frac{x_i - m}{\delta_i}\right) \quad (۵)$$

m میانگین متغیر در طول دوره‌ی آماری، δ_i انحراف معیار متغیر در طول دوره‌ی آماری و x_i مقدار خام متغیر در طول دوره‌ی آماری است.

ج) حداکثر مقیاس: در طول دوره‌ی آماری در این روش با توجه به حداکثر مقدار متغیر مقادیر متغیرها کوچک می‌شوند. رابطه این انتقال به شرح زیر است:

$$X_i = \left(\frac{x_i}{U_i}\right) \quad (۶)$$

U_i حداکثر مقدار متغیر در طول دوره‌ی آماری و x_i مقدار خام متغیر در طول دوره‌ی آماری است.

د) استفاده از میانگین متحرک سه روزه‌ی داده‌های خام: در این حالت به‌جای داده‌های روزانه از میانگین متحرک سه روزه استفاده گردید و با در نظرگرفتن این پارامترها به‌عنوان متغیر مستقل، میانگین متحرک سه روزه‌ی بارش برای یک سال بعد پیش‌بینی گردید.

ه) استفاده از میانگین متحرک پنج روزه‌ی داده‌های خام: در این حالت میانگین متحرک پنج روزه برای کل پارامترهای ورودی مدل محاسبه شد و با در نظر گرفتن این پارامترها به‌عنوان متغیر مستقل، میانگین متحرک پنج روزه‌ی بارش برای یک سال بعد پیش‌بینی گردید.

معیارهای ارزیابی مدل: عملکرد الگوریتم KNN و درخت تصمیم M5 توسط آماره‌های ریشه‌ی متوسط خطای مربعات (RMSE)، ضریب همبستگی (R^2) و متوسط قدر مطلق خطا (MAE) ارزیابی و از طریق رتبه‌بندی آماره‌ها و نزولی کردن آن‌ها در نرم‌افزار اکسل، بهترین رتبه انتخاب گردید.

$$RMSE = \sqrt{\frac{(\sum(E_{si} - E_{oi})^2)}{n-1}} \quad (۷)$$

$$R^2 = \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}))^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2 (Y_i - \bar{Y})^2)} \quad (۸)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}(E_{si} - E_{oi}) \quad (۹)$$

N برابر با تعداد کل داده‌ها ، E_{si} سطح آب تخمین زده‌شده، E_{oi} داده مشاهده‌ای با $\bar{x}$ و $\bar{Y}$ متوسط مقادیر x_i و Y_i هستند.

نتایج و بحث

الگوریتم تنبل KNN و درخت تصمیم M5 تحت ۵ سناریو مورد بحث، در نرم‌افزار Weka اجرا گردید و شاخص‌های آماری به‌دست‌آمده از پنج تابع انتقال مذکور در جدول ۳ تا ۷ نمایش داده‌شد. مقدار بالای R^2 و پایین بودن مقادیر MAE و RMSE نشان‌گر قدرت بالای مدل‌ها در پیش‌بینی دبی در منطقه مورد مطالعه می‌باشد.

مقایسه نتایج نهایی مدل‌های KNN و درخت تصمیم M5 با تاثیر و بدون تاثیر تابع انتقال نرمال کردن و حداکثر مقیاس نشان می‌دهد که تابع انتقال تاثیر چندانی در افزایش R^2 نداشته است و باعث کاهش اندکی در میزان MAE و RMSE می‌شود و در نتیجه نرمال کردن با کاهش فضای موردنیاز برای ذخیره پایگاه داده و همچنین اطمینان از ذخیره‌ی منطقی اطلاعات و تابع انتقال حداکثر مقیاس با کاهش مقادیر متغیرها نتایج مطلوب‌تری را حاصل می‌گردانند. همچنین مشاهده می‌گردد که درخت تصمیم M5 با تاثیر و بدون تاثیر تابع انتقال نرمال کردن و حداکثر مقیاس، در تمامی سناریوها به استثنا سناریو E در تابع انتقال حداکثر مقیاس، نسبت به مدل KNN پیش‌بینی دقیق‌تری را نتیجه می‌دهد. در حالتی که اصل داده‌ها وارد مدل شود، سناریوی e که شامل داده‌های بارندگی روزانه و روز قبل و دبی روزانه و روز قبل می‌باشد، پیش‌بینی بهتری را حاصل می‌کند.

جدول ۳- مقادیر معیار ارزیابی سناریوهای مختلف برای اصل داده‌ها

مدل	سناریو	R^2	MAE	RMSE (متر کعب بر ثانیه)
KNN	a	۰/۰۵	۰/۹۲	۱/۱۵
	b	۰/۵۲	۰/۲۳	۰/۷۶
	c	۰/۰۵	۰/۹۵	۱/۵۶
	d	۰/۴۵	۰/۲۲	۰/۸۶
	e	۰/۳۴	۰/۲۳	۱/۱۷
M5	a	۰/۰۱	۰/۹۲	۱/۱۳
	b	۰/۷۲	۰/۱۷	۰/۵۳
	c	۰/۲۶	۰/۹۲	۱/۱۵
	d	۰/۷۴	۰/۱۶	۰/۵۳
	e	۰/۷۴	۰/۱۶	۰/۵۲

جدول ٤- مقادیر معیار ارزیابی سناریوهای مختلف برای داده‌های نرمال

مدل	سناریو	R^2	MAE	RMSE (متر کعب بر ثانیه)
KNN	a	۰/۰۱	۰/۶۰	۰/۷۴
	b	۰/۵۲	۰/۱۵	۰/۵۰
	c	۰/۱۴	۰/۶۱	۰/۹۹
	d	۰/۴۴	۰/۱۵	۰/۵۶
	e	۰/۶	۰/۱۴	۰/۴۵
M5	a	۰/۰۱	۰/۵۹	۰/۷۳
	b	۰/۷۴	۰/۱۱	۰/۳۴
	c	۰/۱۶	۰/۶	۰/۷۴
	d	۰/۷۴	۰/۱۱	۰/۳۴
	e	۰/۷۴	۰/۱۱	۰/۳۴

جدول ٥- مقادیر تعیین معیار ارزیابی سناریوهای مختلف برای داده‌های حداکثر مقیاس

مدل	سناریو	R^2	MAE	RMSE (متر کعب بر ثانیه)
KNN	a	۰/۰۰۱	۰/۰۲	۰/۰۲
	b	۰/۵۲	۰/۰۰۱	۰/۰۱
	c	۰/۱۴	۰/۰۰۱	۰/۰۲
	d	۰/۵۹	۰/۰۰۱	۰/۰۱
	e	۰/۹	۰/۱۳	۰/۲۹
M5	a	۰/۰۱	۰/۰۲	۱/۰۲
	b	۰/۷۴	۰/۰۰۱	۰/۰۱
	c	۰/۱۶	۰/۰۲	۰/۰۲
	d	۰/۷۴	۰/۰۰۱	۰/۰۱
	e	۰/۷۴	۰/۰۰۱	۰/۰۱

مقایسه نتایج نهایی مدل‌های KNN و درخت تصمیم M5 با تاثیر و بدون تاثیر تابع انتقال میانگین متحرک سه روزه و پنج روزه نشان می‌دهد که تابع انتقال میانگین متحرک پنج روزه تاثیر زیادی در بهبود پیش‌بینی مدل‌ها داشته و کارآمد می‌باشد. دلیل این امر آن است که تابع میانگین متحرک با به‌کارگیری متوسط داده‌ها در یک دوره معین و تشکیل سری زمانی جدید، نوسانات موجود در داده‌ها را کاهش و یا به عبارتی دیگر، این نوسانات را هموار می‌کند. بنابراین با حذف و یا کاهش نوسانات روزانه، روند تغییرات درازمدت بارز می‌گردد. در این حالت نیز مشاهده شد که مدل درخت تصمیم M5 نسبت به مدل KNN قدرت پیش‌بینی بیش‌تری دارد. در مورد داده‌های میانگین متحرک سه روزه‌ی سناریوی d با داشتن داده‌های بارندگی روزانه، دبی روز قبل و دو روز قبل و دبی روزانه، دبی دقیق‌تری را پیش‌بینی می‌کند.

در مورد داده‌های میانگین متحرک پنج روزه سناریوی b با داشتن تمامی داده‌ها و سناریوی c با داشتن داده‌های بارندگی روزانه، بارندگی روز قبل، بارندگی دو روز قبل و دبی روزانه، دبی دقیق‌تری را پیش‌بینی می‌کند. به‌طوری‌که که معیار R^2 برابر ۰/۹۹۹ و معیار MAE و RMSE برابر ۰/۰۰۱ می‌باشند.

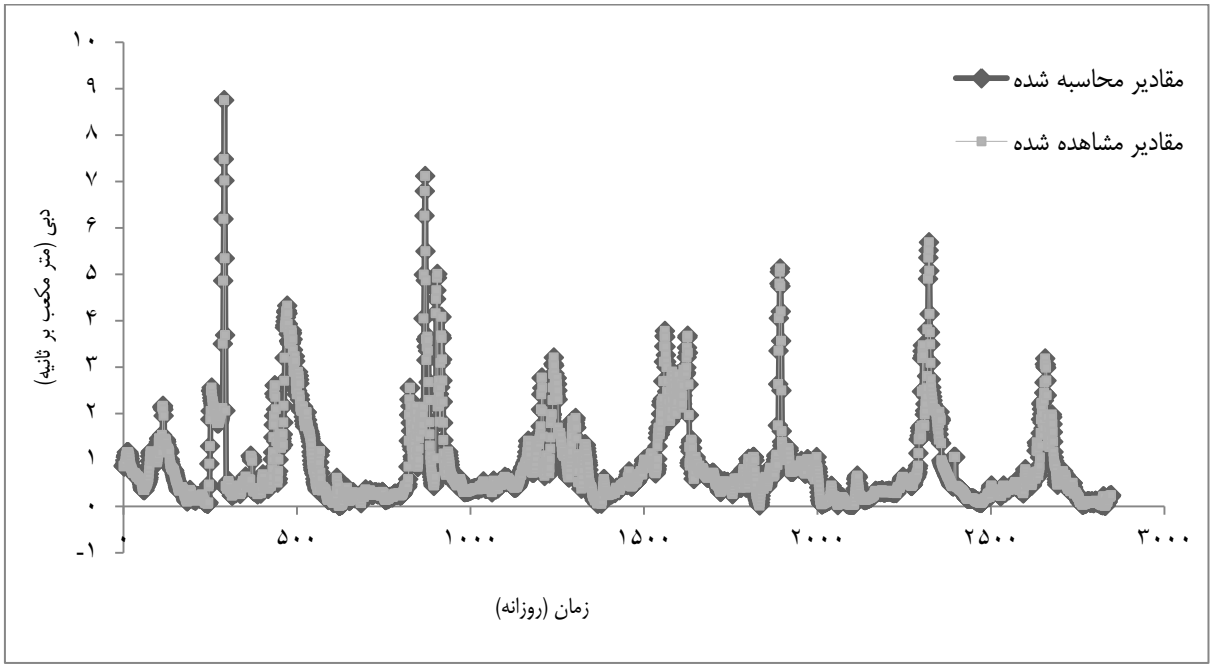
جدول ٦- مقادیر معیار ارزیابی سناریوهای مختلف برای داده‌های میانگین متحرک سه روزه

مدل	سناریو	R^2	MAE	RMSE (متر کعب بر ثانیه)
KNN	a	۰/۰۰۱	۰/۸۹	۱/۰۶
	b	۰/۸۱	۰/۱۶	۰/۴
	c	۰/۰۰۱	۰/۹۹	۱/۴۲
	d	۰/۸۸	۰/۱۲	۰/۳۳
	e	۰/۸۱	۰/۱۶	۰/۴۱
M5	a	۰/۰۰۱	۰/۸۸	۱/۰۵
	b	۰/۹۲	۰/۰۱	۰/۲۴
	c	۰/۱۷	۰/۹	۱/۰۷
	d	۰/۹۴	۰/۰۹	۰/۲۴
	e	۰/۹۲	۰/۱	۰/۲۷

بنابراین مدل درخت تصمیم M5 با تابع انتقال میانگین متحرک پنج ساله با داشتن متوسط R^2 متوسط ۰/۸۱۶ و RMSE متوسط ۰/۲ و MAE متوسط ۰/۲۷۲ دقیق‌ترین پیش‌بینی را ارائه می‌دهد. در شکل ۲ مقادیر مشاهده شده و شبیه‌سازی‌شده دبی رودخانه به‌عنوان نمونه برای سناریو b نشان داده شده‌است و مدل درختی M5 با تابع انتقال میانگین متحرک پنج روزه تحت سناریو c رابطه‌ی ۱۰ تا ۱۲ را نتیجه داده است.

جدول ۷- مقادیر معیار ارزیابی سناریوهای مختلف برای داده‌های میانگین متحرک پنج روزه

مدل	سناریو	R^2	MAE	RMSE (متر کعب بر ثانیه)
KNN	a	۰/۰۰۱	۰/۹۲	۱/۱
	b	۰/۹۷	۰/۰۶	۰/۱۴
	c	۰/۹۹	۰/۰۵	۰/۱۳
	d	۰/۹۲	۰/۰۱	۰/۲۵
	e	۰/۹	۰/۱۳	۰/۲۹
M5	a	۰/۱۴	۰/۸۸	۱/۰۳
	b	۰/۹۹۹	۰/۰۰۱	۰/۰۰۱
	c	۰/۹۹۹	۰/۰۰۱	۰/۰۰۱
	d	۰/۹۸	۰/۰۵	۰/۱۵
	e	۰/۹۶	۰/۰۷	۰/۱۸



شکل ۲- مقادیر مشاهداتی و شبیه‌سازی‌شده جریان با مدل M5 تحت سناریوی b

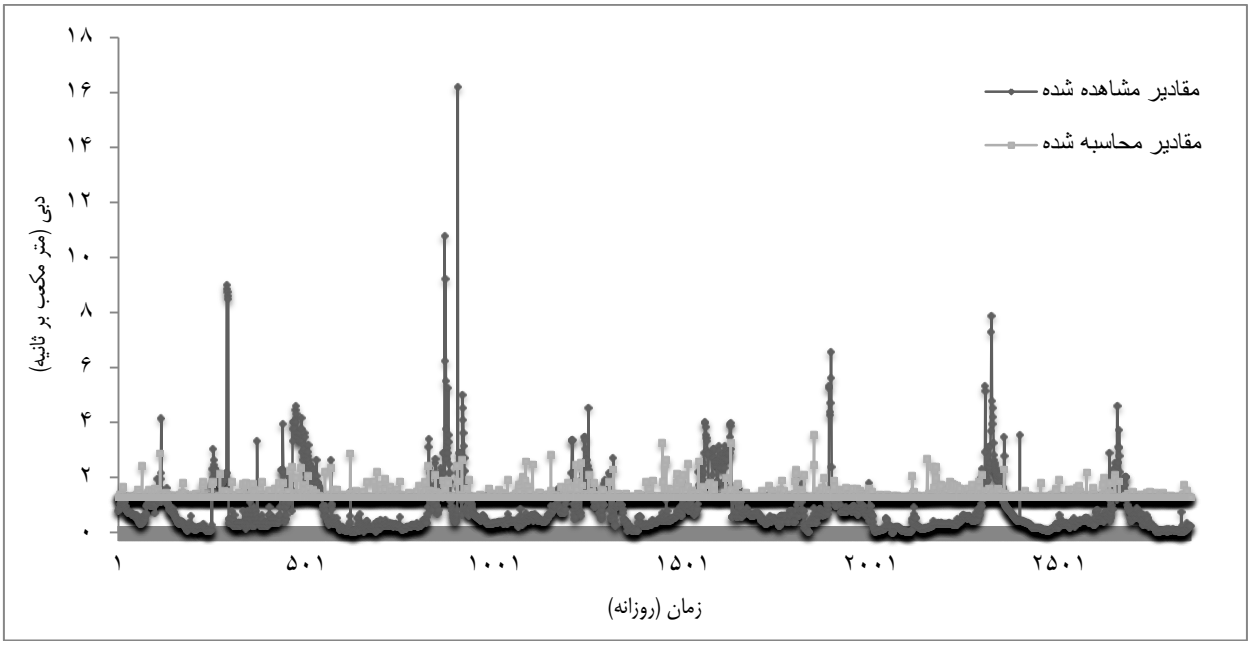
انجام می‌گردد که در هر دو مدل M5 و KNN در تمامی توابع نتیجه نامطلوبی می‌دهد و بدترین حالت آن برای مدل درختی M5 با داده‌ها بدون تاثیر از تابع انتقال می‌باشد که نتیجه‌ی آن در شکل ۳ نشان داده شده است.

$$P(t-2) \le 0.455: Q = 0.9795 * P(t-2) + 0.0147 * Q + 0.0019 \quad (۱۰)$$

$$P(t-2) > 0.455: Q = 0.9999 * P(t-2) \quad (۱۱)$$

$$P(t-2) > 0.748: Q = P(t-2) \quad (۱۲)$$

در سناریو a پیش‌بینی دبی فقط بر اساس بارندگی و دبی روزانه



شکل ۳- مقادیر مشاهده‌شده و محاسبه‌شده برای مدل M5 تحت سناریو a

نتیجه‌گیری

روش KNN و درخت تصمیم M5 به دلیل سادگی یکی از بهترین گزینه‌ها برای انجام پیش‌بینی به‌خصوص در زمینه‌ی هیدرولوژی می‌باشد. در این مطالعه توانایی الگوریتم تنبل KNN و مدل درخت تصمیم M5 در برآورد دبی جریان در یکی از سرشاخه‌های رودخانه گرگانرود در محل ایستگاه سرمو مورد ارزیابی قرار گرفت. نتایج به‌دست‌آمده حاکی از آن است که در تمامی سناریوها به استثناء سناریوی E در تابع انتقال حداکثر مقیاس، مدل درخت تصمیم M5 بر الگوریتم KNN برتری دارد. همچنین سناریوی b که شامل داده‌های بارندگی و دبی روزانه، یک روز قبل و دو روز قبل بود در اکثر حالات نتایج خوبی را برای شبیه‌سازی دبی حاصل می‌کرد و بهترین حالت آن نیز در مدل درخت تصمیم M5 با تابع انتقال میانگین متحرک پنج ساله بوده است. تابع میانگین متحرک نیز بهترین نتایج را در بین تمامی توابع انتقال مورد آزمون در این مطالعه را حاصل می‌نمود.

منابع

حسین‌زاده،م،م.، عمادالدین،س.، نامجو،ف. ۱۳۹۴. پهنه‌بندی و واکاوی فرایندهای هوازدگی در حوضه قره‌سو گرگان. فصل‌نامه جغرافیای طبیعی. ۸. ۲۹: ۱۸-۱.

ستاری،م،ت.، نهرین،ف. ۱۳۹۲. پیش‌بینی مقادیر حداکثر بارش روزانه با استفاده از سیستم‌های هوشمند و مقایسه آن با مدل درختی M5؛ مطالعه موردی ایستگاه‌های اهر و جلفا. فصلنامه علمی پژوهشی مهندسی آبیاری و آب. ۴. ۱۴ : ۸۳- ۹۸.

شریف‌آذری،س.، عراقی‌نژاد،ش. ۱۳۹۲. توسعه مدل ناپارامتری شبیه‌ساز داده‌های ماهانه هیدرولوژیکی. مجله مدیریت آب و آبیاری. ۳. ۱: ۸۳-۹۵.

طالبی،ع.، اکبری،ز. ۱۳۹۲. بررسی کارایی مدل درختان تصمیم گیری در برآورد رسوبات معلق رودخانه‌ای (مطالعه موردی: حوضه سد ایلام). مجله علوم و فنون کشاورزی و منابع طبیعی. علوم آب و خاک. ۱۷. ۶۳: ۱۰۹-۱۲۱.

ظهیری،ع.، قربانی،خ. ۱۳۹۲. شبیه‌سازی دبی جریان در مقاطع مرکب به کمک مدل درخت تصمیم M5. مجله پژوهش‌های حفاظت آب و خاک. ۲۰. ۳: ۱۱۳-۱۳۲.

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

کریمی ماسوله،ح.، احمدوند،م.، معاضد،ه. ۱۳۸۸. کاربرد شبکه‌های عصبی در پیش‌بینی دبی رودخانه کارون بر اساس داده‌های آماری بارندگی شش ماه گذشته ایستگاه‌های بالادست، هشتمین سمینار بین المللی مهندسی رودخانه، اهواز، دانشگاه شهید چمران.

نقدی بانسوله،ک.، موسوی،ج. ۱۳۹۲. پیش‌بینی جریان ورودی به مخزن سد زاینده‌رود با استفاده از الگوریتم پیش‌بینی K- نزدیک ترین همسایگی (KNN). پنجمین کنفرانس مدیریت منابع آب

ایران. تهران. انجمن علوم و مهندسی منـابع آب ایـران. دانشـگاه شهید بهشتی

Hand.D., Heikki,M and Padhraic,S. 2001. Principles of Data Mining. A Bradford book. The MIT press. Cambridge, Massachusetts, London, England.

Jagtap,S.S., Lall,U., Jones,J.W., Gijsman,A.J., Ritchie.J.T. 2004. Dynamic nearest-neighbor method for estimating soil water parameters. Transactions of the ASAE. 47:1437–1444.

Lall,U., Sharma,A. 1996. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resources Research. 32.3: 679-694.

Londhe,S,N., Dixit,P,R. 2011. Forecasting Stream Flow Using Model Trees. International Journal of Earth Sciences and Engineering. 4.6: 282-285.

Quinlan,J.R. 1992. Learning with continuous classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence (AI'92). Singapore: World Scientific. pp. 343–348.

SattariM.T., Pal,M., Apaydin,H., Ozturk,F. 2013. M5 model tree application in daily river flow forecasting in Sohu stream, Turkey. Water Resources. 40.3: 233-242.

Twarakavi,N.K.C., Šimůnek,J and Schaap,M.G. 2009. Development of Pedotransfer Functions for Estimation of Soil Hydraulic Parameters using Support Vector Machines. Soil oil Science Society of America Journal. 73:1443-1452.

Iranian Journal of Irrigation and Drainage
No. 3, Vol. 12, Agu.-Sep. 2018, p. 579-587

نشریه آبیاری و زهکشی ایران
شماره۳، جلد ۱۲، مرداد - شهریور ۱۳۹۷، ص. ۵۸۷-۵۷۹

Investigation of Ability of KNN and M5 Trees Models in River Flow Modeling: Case Study Sarmo Stations

A.khalili naft chali[1]*, H Khozeymehnezhad [2]
Recived: Dec.11, 2017 Accepted: Jan.15, 2017

Abstract

Prediction of the flow discharge accurately is a key point in the optimum planning and management of water resources. The Ghorghanrood watershed is one of the watersheds in the north of Iran that it is located in Gholestan Province. The impact of drought and rain on basic fluctuations and overall flow of the river plays a major role in planning of implementation of watershed resources. In this research, the KNN model and M5 decision tree were used as one of the methods of data-mining for estimating the flow discharge of Ghorghanrood River that it is located in Sarmo Station. In this regard, the raining and flow discharge data of Sarmo Station (in Mohammadabad) under five various scenarios by applying transfer functions on the data. The results showed that M5 decision tree has mostly superiority over KNN model and reaches more accurate prediction. Also, in the defined scenarios, the b and c models which respectively include all data and daily raining data, last day raining and two days before raining, result the most accurate estimation. These two scenarios are under the transfer function, the five-day moving mean and have R^2 standard of 0.999 and MAE and RMSE standards of 0.001.

Keywords: transfer functions, data-mining, discharge, decision tree, scenario

1 - PhD student of water Science Engineering, University of Birjand
2 - Associate professor, Department of water Science Engineering, University of Birjand
(* - Corresponding Author Email: Atenakhalili_2014@yahoo.com)