ارزیابی مدل ماشین بردار پشتیبان (SVM) در پیش‌بینی جریان ماهانه رودخانه باراندوزچای ارومیه

سید امین غیبی [۱]، میرعلی محمدی [۲]، هیراد عبقری [۳]

۱- * دانشجوی کارشناسی ارشد عمران-آب با گرایش آب و سازه های هیدرولیکی (amin.hesari.b.e.72@gmail.com)
۲- دانشیار مهندسی عمران دانشکده فنی, دانشگاه ارومیه,(m.mohammadi@urmia.ac.ir)
۳- دانشیار دانشکده منابع طبیعی, دانشگاه ارومیه، (h.abghari@urmia.ac.ir)

چکیده

بعد از نرمال سازی و ایستایی داده ها با استفاده آزمون چولگی و ADF. نتایج نشان داد که سری ماهانه پس از نرمال و استاندارد شدن ایستا می‌باشد. برای پیش‌بینی سری زمانی جریان ماهانه رودخانه باراندوزچای با استفاده از مدل SVM، الگوهای ورودی در دو حالت الف) بدون در نظر گرفتن بارش و ب) با در نظر گرفتن بارش مورد بررسی قرار گرفت. نتایج نشان داد که در حالت اول (بدون در نظر گرفتن اثر بارش در پیش‌بینی جریان) رفته رفته دقت در مدل‌سازی افزایش یافته اما از تاخیر چهارم به بعد عملکرد این مدل کاهش می‌یابد. با در نظر گرفتن اثر بارش نیز بهترین عملکرد این مدل مربوط به الگوی B8 بود. با توجه به نتایج به دست آمده مدل SVM در الگوی B8 و با مقادیر NS = 0.95، R = 0.98 و RMSE = 2.36 (m3/s) بالاترین دقت را در مدل‌سازی داشته است.

واژه‌های کلیدی:

ماشین بردارپشتیبان، مدل‌سازی و پیش‌بینی، چولگی،ADF

مقدمه

با توجه به بحران آب دریاچه ارومیه و روند رو به خشک آن و همچنین اهمیت و حساسیت موضوع مهارآب های سطحی و استفاده بهینه از آن بخصوص در ایران که کمبود آب در پهنه وسیعی از کشور وجود دارد، نیاز به شناسایی و به مدل درآوردن رفتار و عملکرد منابع آبی، جهت برنامه‌ریزی بلندمدت و کوتاه‌مدت و استفاده بهینه از پتانسیل‌های آن‌ها بسیار حساس و مورد نیاز می‌باشد.

ماشین بردار پشتیبانی (Support vector machines- SVM) یکی از روش‌های یادگیری بانظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی از جمله شبکه‌های عصبی پرسپترون نشان داده است. مبنای کار دسته‌بندی کنندهٔ SVMدسته‌بندی خطی داده‌ها است. در تقسیم خطی داده‌ها خطی انتخاب می‌شود که حاشیه اطمینان بیشتری داشته باشد.

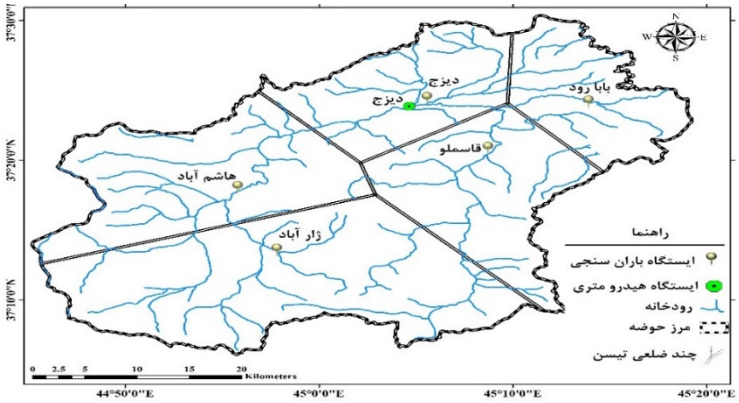
مواد و روش‌ها

منطقه مورد مطالعه

این رودخانه از دو رشته ارتفاعات مرزی ایران و ترکیه و ایران و عراق به نام های جمال الدین و ارتفاعات ککوداغ الوق سرچشمه می گیرد و بعد از طی مسافتی وارد جلگه شده دشت بیل را آبیاری می کند و بعد هم از آبادی زیوه گذشته و به روستای هفتوان باراندوز می رسد و از این ناحیه به بعد به نام باراندوزچای نامیده می شود. رودخانه باراندوز در مسیر خود با آبیاری کردن زمین های متعدد روستایی در منطقه داروغه شعبه کوچکی به آن پیوسته که شعبه اصلی آن باباروداست به دریاچه اورمیه می ریزد.

داده های مورد استفاده

دراین تحقیق از داده های بارندگی و جریان رودخانه ای باراندوزچای ارومیه دردوره های آماری ۱۳۵۵تا ۱۳۹۵ به صورت روزانه ای و ماهانه ای به کار برده شد



شکل ۱- نقشه باراندوزچای ارومیه

شکل (۱) ایستگاه های بارانسنجی حوزه باراندوزچای ارومیه را نشان می دهد که ایستگاه دیزج به دلیل طبیعی بودن رژیم جریان و ورودی بالادست آن به عنوان ایستگاه هیدرومتری انتخاب گردید.

ماشین بردار پشتیبان

در مدل رگرسیونی SVM، تابعی مرتبط با متغیر وابسته Y که خود تابعی از چند متغیر مستقل X، است برآورد می‌شود. مشابه سایر مسائل رگرسیونی، فرض می‌شود که رابطه‌ی میان متغیرهای مستقل و وابسته با تابع جبری مانند $f(x)=w^T.\phi(x)+b$ به علاوه مقداری اغتشاش مشخص شود $(y=f(x)+\varepsilon)$. قابل ذکر است که در کتب مرجع مرتبط با

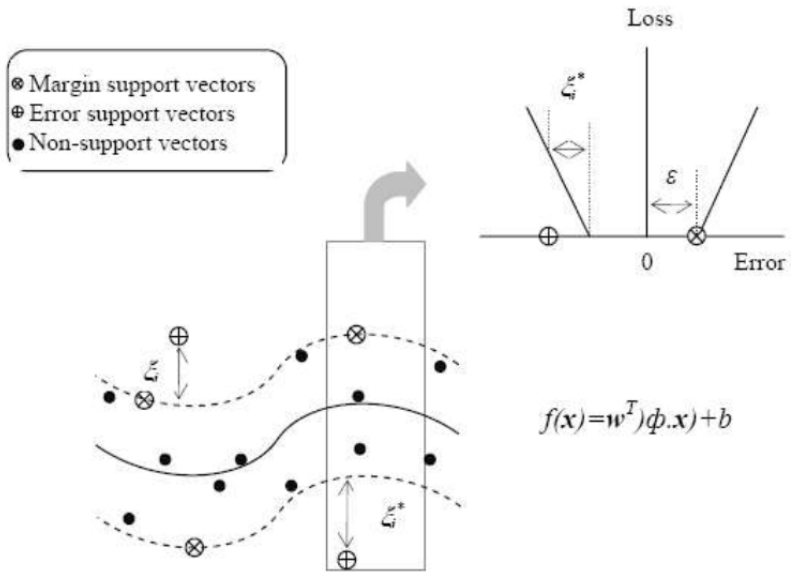
SVM اغتشاش به عنوان خطای مجاز ε تعریف شده است. چنانچه W (بردار ضرایب) و b (ثابت) مشخصه‌های تابع رگرسیونی و ϕ نیز تابع کرنل باشد، آنگاه هدف پیدا کردن فرم تابعی برای $f(x)$ است. این مهم با آموزش مدل SVM توسط مجموعه‌ای از نمونه‌ها (مجموعه آموزش) محقق می شود. این روند شامل بهینه سازی متوالی تابع خطا ست. بسته به تعریف، این تابع خطا دو نوع مدل SVM تعریف می‌شود: الف) SVM رگرسیونی نوع یک یا $SVM-\varepsilon$ و ب) SVM رگرسیونی نوع دو یا $SVM-\upsilon$.(نیکبخت و شهنازی,۱۳۸۸)

در این تحقیق از مدل ε-SVM رگرسیونی، به دلیل کاربرد گسترده آن در مطالعات رگرسیونی، برای پیش‌بینی جریان رودخانه باراندوزچای استفاده شد. بنابراین برای محاسبه W و b لازم است تابع خطا (معادله۱) در مدل ε-SVM با درنظرگرفتن شرایط مندرج در معادله‌ی(۱) بهینه شود.

$$\frac{1}{2}W^T W + C\sum_{i=1}^{N}\xi_i + C\sum_{i=1}^{N}\xi_i^* \qquad (۱)$$
$$\text{Subject to:}\begin{cases} W^T.\phi(x_i)+b-y_i \le \varepsilon+\xi_i^* \\ y_i - W^T.\phi(x_i)-b \le \varepsilon+\xi_i \\ \xi_i^*,\xi_i \ge 0, i=1,...,N \end{cases}$$

در معادلات بالا C عددی صحیح مثبت است که عامل تعیین جریمه در هنگام رخ دادن خطای آموزش مدل است. ϕ تابع کرنل[۱]، N تعداد نمونه‌ها و دو مشخصه ξ_i^*, ξ_i متغیرهای کمبود هستند که حد بالا و پایین خطای آموزش بوده و خطای مجاز ε را مشخص می‌کنند. در مسائل پیش‌بینی می شود که داده‌ها، درون بازه مرزی ε قرار گیرند (شکل ۱). حال اگر داده‌ای خارج از بازه ε قرار گرفت آنگاه یک خطا معادل با ξ_i^*, ξ_i وجود خواهد داشت. ذکر این نکته نیز لازم است که مدل SVM مشکلات ناشی از کم تخمینی و فوق برازشی را با کمینه کردن همزمان عبارت $C\sum_{i=1}^{N}\left(\xi_i+\xi_i^*\right)$ در معادله ۱ حل می‌کند.(احمدی و خلیلی,۱۳۹۱)

بنابراین با معرفی دو ضریب لاگرانژ a_i^*, a_i مسئاله ای بهینه‌سازی با حداکثرسازی عددی تابع درجه دوم رابطه (۲)با ششرایط مشخص حل خواهد شد.



شکل ۲- مدل SVM رگرسیونی

$$Max: Z=\sum_{i=1}^{N} y_i\left(\alpha_i-\alpha_i^*\right)-\varepsilon\sum_{i=1}^{N}\left(\alpha_i-\alpha_i^*\right)-$$
$$0.5\sum_{i,j=1}^{N}\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)\phi\left(x_i\right)^T.\phi\left(x_j\right) \quad (۲)$$
$$\text{Subject to:}\begin{cases}\sum_{i=1}^{1}\left(a_i-a_i^*\right)=0\\ 0\le a_i\le C, i=1,2,\ldots 1\\ 0\le a_i^*\le C, i=1,2,\ldots 1\end{cases}$$

تابع هدف بالا در معادله‌ی $\Delta w_i(n)=\mu y(n)x_i(n)$ تابع محدب ا ست و بنابراین جواب معادله‌ی مذکور یکتا و بهینه خواهد بود. پس از تعریف ضرایب لاگرانژ در این معادله‌ی مشخصه‌های W و b در مدل SVM رگرسیونی با استفاده از شرایط تئوری کراش- کوهن-تاکر محاسبه می‌شوند (فلچر، ۲۰۰۹).

که در آن $W=\sum_{i=1}^{N}\left(\alpha_i-\alpha_i^*\right)\phi\left(x_i\right)$ است. در نتیجه برای مدل SVM رگرسیونی خواهیم داشت:

$$W=\sum_{i=1}^{N}\left(\alpha_i-\alpha_i^*\right)\phi\left(x_i\right)^T.\phi\left(x\right)+b \quad (۳)$$

باید توجه دا شت که ترم‌های لاگرانژ $\left(\alpha_i-\alpha_i^*\right)$ می‌تواند صفر، و یا غیر صفر با شند. بنابراین فقط مجموعه داده‌هایی که ضرایب $\overline{\alpha_i}$ آنها غیرصفر است در معادلهٔ رگرسیون نهایی وارد می‌شوند و این مجموعه داده ها به عنوان بردارهای پشتیبان شناخته می‌شوند .به طور ساده، بردارهای پشتیبان آن داده‌هایی هستند که به ساخته شدن تابع رگرسیونی کمک می‌کنند. در میان بردارهای مذکور آن‌هایی که مقدار $\left|\overline{\alpha_i}\right|$ آنها کمتر از C با شد بردارهای پ شتیبان حا شیه‌ای نامیده می شوند. هنگامی که مقدار $\left|\overline{\alpha_i}\right|$ بردارهای پ شتیبان برابر مقدار C با شد، به عنوان بردار پشتیبان خطا، یا بردار پشتیبان کراندار شناخته می شود .بردارهای پشتیبان حا شیه ای در حا شیه مرز غیرحساس یافت می شوند، در حالی که بردارهای پشتیبان خطا خارج از بازه هستند (شکل ۱). در نهایت تابع SVM رگر سیونی را می توان به فرم معادله‌ی (۴-1) بازنویسی کرد (نیکبخت وشهنازی,۱۳۸۸)

$$f\left(x\right)=\sum_{i=1}^{N}\overline{\alpha_i}.\phi\left(x\right)+b \quad (۴)$$

در معادله‌ی (۴) محاسبه $\phi\left(x\right)$ در فضای مشخصهٔ آن ممکن است بسیار پیچیده با شد. برای حل این مشکل روند معمول در مدل SVM رگر سیونی انتخاب یک تابع کرنل به صورت $K\left(x_i,x\right)=\phi\left(x_i\right)^T\sqrt{b^2-4ac}$ ا ست. می‌توان از توابع مختلف کرنل برای ساخت انواع مختلف مدل $SVM-\varepsilon$ ا ستفاده کرد. انواع رایج توابع کرنل قابل ا ستفاده در مدل SVM رگر سیونی در جدول ۱ ارائه شده است.(هامل,۲۰۰۹)

جدول ۱- توابع کرنل رایج در ماشین های بردار پشتیبان (هامل,۲۰۰۹)

نوع تابع	تابع کرنل
خطی	$K(x_i,x_j)=x_i^T.x_j$
چند جمله ای	$K(x_i,x_j)=(\gamma x_i^T.x_j+C)^d$
تانژانت هیپربولیک	$K(x_i,x_j)=\tanh(\gamma x_i^T.x_j+C)$
RBF	$K(x_i,x_j)=\exp\left(-\gamma\left\|x_i-x_j\right\|^2\right)$

۲-۱-ارزیابی مدل SVM برای پیش بینی دبی جریان ماهانه رودخانه مورد مطالعه

به منظور ارزیابی عملکرد مدل‌های به کار گرفته شده از سه معیار ضریب همبستگی(R)، جذر میانگین مربعات خطا (RMSE) و نش- ساتکلیف (NS) به شرح زیر استفاده گردید:

$$RMSE = \left[\frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{n}\right]^{0.5} \quad (۵)$$

$$R = \sqrt{\frac{\left[\sum_{i=1}^{n}(Q_i - \overline{Q})(\hat{Q}_i - \overline{Q})\right]^2}{\sum_{i=1}^{n}(Q_i - \overline{Q})^2 \sum_{i=1}^{n}(\hat{Q}_i - \overline{Q})^2}} \quad (۶)$$

$$NS = 1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \overline{Q})^2} \quad (۷)$$

که در روابط فوق Q_i، $\hat{Q}_i$، $\overline{Q}$ به ترتیب دبی جریان مشـــاهداتی، دبی جریان محاســـباتی، میانگین دبی جریان و n تعداد داده‌ها می‌با شد. مدلی که حاوی کمترین مقدار RMSE و یا بیشترین مقدار ضریب همبستگی (R) و ضریب نش- ساتکلیف با شد به عنوان مدل مطلوب شناخته می‌شود.

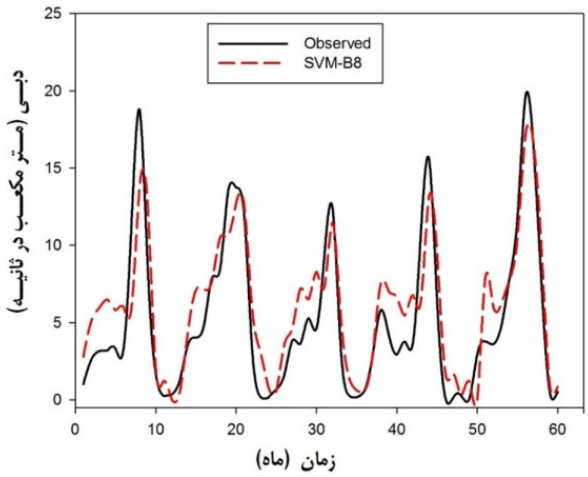
مدل سازی دبی ماهانه جریان رودخانه باراندوزچای با استفاده از مدل SVM

در این مرحله به منظور کاهش دامنه تغییرات داده‌های جریان ماهانه رودخانه باراندوزچای و همســـان‌ســـازی اطلاعات ورودی و خروجی، داده‌ها ا ستاندارد سازی شد. در مرحله بعد مقادیر بهینه مشخصه‌های مدل SVM شامل ε و C تعیین می‌گردد. همچنین در این مطالعه تابع کرنل مورد اســـتفاده، تابع RBF انتخاب شـــد چرا که از دقت بهتری در برآورد جریان ماهانه رودخانه باراندوزچای برخوردار بود. در تابع RBF نیز مشـــخصـــه γ بایســـتی تعیین گردد. بنابراین در حالت کلی برای پیش‌بینی جریان ماهانه رودخانه باراندوزچای توسط مدل SVM، لازم است که مقادیر بهینه سه مشخصه مذکور به‌دست آید که بدین منظور دو مشخصه ε و C توسط الگوریتم بهینه‌سازی جستجوی شبکه و متغیر γ نیز به صورت آزمون و خطا محاسبه شد. البته قابل ذکر است که الگوریتم بهینه‌سازی جستجوی شبکه بسیار کند عمل می‌کند و زمان محاسباتی زیادی را به خود اختصاص می‌دهد. برای حل این مشکل در تحقیق مذکور طبق توصیه (چن و یو,۲۰۰۶) از برنامه اصلاح شده الگوریتم جستجوی شبکه که به نام الگوریتم جستجوی شبکه دو مرحله‌ای معروف است به همراه اعتبارسنجی متقاطع استفاده شد. برای این منظور ابتدا با انتخاب شبکه‌هایی با ابعاد بزرگ محدوده مشخصه‌های ε و C به ازای مقدار ثابت مشخصه γ تعیین شد. سپس با مشخص شدن محدوده مذکور و تقسیم آن به شبکه‌هایی با ابعاد ریزتر مقادیر دقیق دو مشخصه ε و C مشخص شدند. روند مذکور برای دیگر مقادیر γ نیز تکرار شد و بدین‌طریق مدل‌های متفاوتی با تغییر در مقدار γ حاصل شدند. حال می‌توان از بین مدل‌های توسعه داده شده مدل با کمترین خطا را تعیین کرده و مشخصه‌های آن را به عنوان مقادیر بهینه ε، C و γ انتخاب نمود.

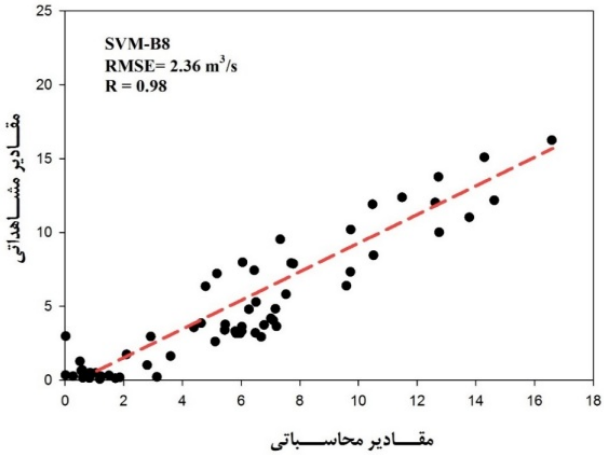
برای مدل سازی جریان ماهانه رودخانه باراندوزچای با ا ستفاده از مدل SVM، داده‌های ۳۵ سال (از مهر ۱۳۵۵ تا شهریور ۱۳۸۹) برای آموزش و پنج سال (از مهر ۱۳۹۰ تا شهریور ۱۳۹۵) به عنوان داده‌های تست انتخاب شدند. به طور کلی ۸۵ در صد داده‌ها برای آموزش و ۱۵ در صد نیز برای تست در نظر گرفته شد. برای هر الگوی ورودی با تغییر ε، C و γ شبکه‌های مختلفی ساخته و آموزش داده شد و در نهایت ساختاری که دارای کمترین خطا بود، به عنوان مناسب‌ترین الگو انتخاب شد. جدول ۱ شاخص‌های آماری مربوط به نتایج حاصل از کاربرد مدل SVM و مناسب‌ترین مقادیر ε، C و γ را برای هر الگوی ورودی نشان می‌دهد.

جدول ۲-تحلیل‌های آماری نتایج حاصل از مدل ماشین بردار پشتیبان برای الگوهای ورودی جریان ماهانه رودخانه باراندوزچای

الگوی ورودی	آموزش			تست			ضرایب مدل SVM		
	RMSE (m^3/s)	R	NS	RMSE (m^3/s)	R	NS	C	ε	γ
B1	۵/۹۲	۰/۷۰	۰/۵۲	۳/۶۴	۰/۶۹	۰/۶۹	۴	۰/۱	۰/۲
B2	۵/۱۷	۰/۷۹	۰/۶۹	۳/۲۱	۰/۸۱	۰/۷۳	۱۰	۰/۱	۰/۵
B3	۴/۱۴	۰/۸۳	۰/۸۶	۲/۴۲	۰/۹۵	۰/۹۰	۸	۰/۱	۰/۳۳
B4	۴/۹۹	۰/۸۲	۰/۸۳	۳/۱۰	۰/۸۴	۰/۸۰	۱۰	۰/۱	۰/۲۵
B5	۵/۱۰	۰/۷۲	۰/۶۸	۳/۳۱	۰/۸۱	۰/۷۵	۹	۰/۱	۰/۲
B6	۵/۴۱	۰/۷۶	۰/۷۳	۳/۱۴	۰/۸۱	۰/۷۵	۹	۰/۱	۰/۵
B7	۵/۰۳	۰/۸۲	۰/۷۸	۳/۰۶	۰/۸۴	۰/۸۱	۱۰	۰/۱	۰/۲۵
B8	۴/۴۰	۰/۸۷	۰/۸۲	۲/۳۶	۰/۹۸	۰/۹۵	۱۰	۰/۱	۰/۱۷
B9	۴/۶۳	۰/۸۴	۰/۷۹	۲/۸۱	۰/۸۷	۰/۸۲	۱۰	۰/۱	۰/۱۳
B10	۵/۶۰	۰/۸۳	۰/۷۷	۲/۹۵	۰/۸۴	۰/۷۹	۱۰	۰/۱	۰/۱



شکل ۳- مقادیر مشاهداتی و مقادیر پیش‌بینی شده جریان ماهانه حاصل از الگوی B8 مدل SVM در مرحله تست



شکل ۴- پراکندگی مقادیر مشاهداتی و مقادیر پیش‌بینی شده جریان ماهانه حاصل از الگوی B8 مدل SVM در مرحله تست

نتیجه‌گیری و پیشنهادها

در مطالعه حاضر از روش هوش محاسباتی برای مدل‌سازی و پیش‌بینی جریان ماهانه رودخانه باراندوزچای در محل ایستگاه هیدرومتری دیزج در دو حالت الف) بدون در نظر گرفتن اثر بارندگی و ب) با در نظر گرفتن اثر بارندگی در پیش‌بینی جریان استفاده گردید.

در ابتدا الگوهای ورودی در دو حالت برای معرفی به مدل مورد استفاده آماده گردید. در حالت اول فقط از داده‌های جریان رودخانه به منظور پیش‌بینی و مدل سازی استفاده شد. در حالت دوم داده‌های بارندگی نیز برای مدل سازی به کار گرفته شد. در مجموع ۱۰ الگوی مختلف برای پیش‌بینی جریان ماهانه انتخاب و به مدل‌ها معرفی گردید.

در حالت اول (بدون در نظر گرفتن اثر بارش در پیش‌بینی جریان) رفته رفته دقت در مدل‌سازی افزایش یافته اما از تاخیر چهارم به بعد عملکرد هردو مدل کاهش می‌یابد. با در نظر گرفتن اثر بارش نیز بهترین عملکرد مدل مربوط به الگوی B8 بوده و به طور کلی تاثیر دادن بارش در مدل سازی موجب افزایش دقت و کارایی مدل SVM شده است. با توجه به نتایج به دست آمده مدل SVM در الگوی B8 و با مقادیر NS = 0.95، R = 0.98 و RMSE = 2.36 (m3/s) بهترین عملکرد را در پیش‌بینی جریان ماهانه رودخانه باراندوزچای داشته است.

۱. پیشنهاد می شود که تاثیر سایر پارامترهای هیدرولوژیکی از قبیل تبخیر، دما و غیره (علاوه بر داده‌های بارش و دبی جریان) در مدل‌سازی جریان رودخانه با استفاده از برنامه‌ریزی بیان ژن و ماشین بردار پشتیبان مورد بررسی قرار گیرد.
۲. مطالعه مشابه برای رودخانه مورد مطالعه با افزایش داده‌ها هر پنج سال یکبار تکرار و نتایج مورد بحث و بررسی قرار گیرد.

منابع

- احمدی ف. خلیلی ک. ۱۳۹۱. تعیین زمان تغییر روند جریان سالانه رودخانه‌های حوضه دریاچه ارومیه. کنفرانس بین المللی دریاچه ارومیه، ۱۸ تا ۲۰ آذر ، دانشگاه ارومیه، ارومیه.
- نیک بخت شهبازی ع.ر. ۱۳۸۸. کاربرد ماشین بردار پشتیبان در پیش بینی جریان رودخانه. هشتمین کنفرانس هیدرولیک ایران. دانشکده فنی دانشگاه تهران.
- Chen, H.L. and Rao, A.R., 2006. Linearity analysis on stationarity segments of hydrologic time series. J. Hydro., 277: 89-99.
- Hamel, L. H. (2009). Knowledge discovery with support vector machines (Vol. 3). John Wiley & Sons.
- Fealcher, C. 2009. Automaticlly defined functions in gene expression programming. In: Nedjah, N., Mourelle, L.M., Abraham, A. (Eds.), Genetic Systems Programming: Theory and Experiences, Studies in Computational Intelligence, Springer-Verlag. 13: 21-56.