محیط شناسی، سال سی و هشتم، شمارهٔ ۶۱، بهار ۹۱، صفحهٔ ۷۱-۸۲

توسعهٔ مدلی مناسب بر مبنای شبکهٔ عصبی مصنوعی و ماشین بردار پشتیبان برای پیش‌بینی بهنگام اکسیژن‌خواهی بیوشیمیایی ۵ روزه

علی اسکندری[*1]، روح‌اله نوری[2]، حامد معراجی[3]، امین کیاقادی[4]

۱- مربی گروه مهندسی عمران و محیط زیست، دانشکدهٔ فنی، دانشگاه آزاد اسلامی بوشهر

۲- دانشجوی دکترای گروه مهندسی عمران و محیط زیست، دانشکدهٔ فنی، دانشگاه آزاد اسلامی بوشهر rooholahnoori@gmail.com

۳- دانشجوی دکترای گروه مهندسی عمران و محیط زیست، دانشکدهٔ فنی، دانشگاه آزاد اسلامی بوشهر hamedmeraji@gmail.com

٤- دانشجوی کارشناسی ارشد مهندسی محیط زیست دانشکدهٔ فنی، دانشگاه آزاد اسلامی بوشهر amin.kiaghadi@gmail.com

تاریخ دریافت : ۸۹/۱۲/۱۲ تاریخ پذیرش: ۹۰/٤/۹

چکیده

محدودیت سنسورهای سخت‌افزاری برای اندازه‌گیری برخی مشخصه‌های کیفی آب مانند اکسیژن‌خواهی بیوشیمیایی ۵ روزه (BOD_5) که از لحاظ زمانی هزینه‌بر هستند، تلاش‌ها را به سمت استفاده از سنسورهای نرم‌افزای برای پیش‌بینی بهنگام BOD5 سوق داده است. هدف اصلی مقاله مذکور نیز توسعهٔ سنسور نرم‌افزاری مناسب بر مبنای مدل‌های هوشمند شبکهٔ عصبی مصنوعی (ANN) و ماشین بردار پشتیبان (SVM) برای تخمین بهنگام BOD_5 در رودخانهٔ سفیدرود است. برای این منظور با قرار دادن BOD_5 به عنوان تابعی از دیگر متغیرهای کیفیت آب، مدل‌های مناسبی برای این موضوع با استفاده از دو مدل ANN و SVM توسعه داده شد. در توسعهٔ مدل ANN نقش توابع آموزش لونبرگ–مارکویت (LM)، پس انتشار ارتجاعی (RP) و گرادیان مزدوج مقیاس‌دار (SCG) در بهینه کردن مشخصه‌های ANN ارزیابی شد. همچنین برای بهینه کردن مشخصه‌های مدل SVM از الگوریتم بهینه‌سازی جستجوی شبکهٔ دو مرحله‌ای استفاده شد. نتایج این تحقیق مبین عملکرد برتر مدل ANN با الگوریتم LM (مدل (ANN (LM) نسبت به دو الگوریتم دیگر بود. همچنین مدل SVM نیز از عملکرد مناسبی در تخمین BOD_5 برخوردار بود، به طوری‌که مقدار ضریب همبستگی پیرسون برای این مدل در مرحله تست معادل ۰/۹۵ به‌دست آمد. در نهایت نیز بررسی‌های بیشتر برای ارزیابی یکی از دو مدل منتخب بر مبنای آماره نسبت تفاوت توسعه داده شده انجام پذیرفت که نتایج به‌دست آمده از این آماره حاکی از عملکرد برتر مدل SVM نسبت به ANN (LM) بود.

کلید واژه

شبکهٔ عصبی مصنوعی، ماشین بردار پشتیبان، رودخانهٔ سفیدرود، اکسیژن خواهی بیوشیمیایی ۵ روزه

سرآغاز

اکسیژن خواهی بیوشیمیایی ۵ روزه[1] (BOD_5) به عنوان یکی از مهمترین مشخصه‌های کیفی آب و شاخص آلودگی با آلاینده‌های دارای منشاء مواد آلی نقش مهمی در ارزیابی شرایط شیمیایی و بیولوژیکی سیستم‌های منابع آب ایفا می‌کند. روش مرسوم برای اندازه‌گیری این مشخصه مهم معمولاً با دشواری‌ها و خطاهای فراوانی همراه است (Beltran, et al., 1998; Einax, et al., 1999). وجود عوامل پیچیده مختلف از قبیل اکسیژن مورد نیاز جلبکی در نمونه، تحت تأثیر قرارگیری فعالیت میکربی در نمونه به دلیل حضور مواد سمی، متفاوت بودن شرایط آزمایشگاهی با سیستم‌های منابع آب مربوط و مهمتر از همه نیاز به ۵ روز زمان تحت شرایط ثابت دما (۲۰ درجه سلسیوس) تخمین این مشخصه شاخص را با عدم قطعیت و سوالات زیادی همراه می‌سازد (Singh, et al., 2009). در این میان مشکل زمان‌بر بودن تعیین BOD_5 عملاً استفاده از آن را در مطالعات کیفی منابع آب برای اتخاذ تصمیمات مدیریتی مناسب بسیار محدود کرده، به‌طوری‌که این موضوع لزوم انجام مطالعات گسترده برای تخمین بهنگام این مشخصه شاخص را مشخص می‌کند. با مرور مطالعات انجام شده می‌توان تلاش‌های صورت گرفته برای تخمین بهنگام BOD_5 را به دو دسته شامل استفاده از تست‌های میکربی (Sohn, et al., 1995; Rastogi, et al., 2003) و مدل‌های کیفی (Oliveira-Esquerre et al., 2004a; Zhao and Chi, 2005; Honggui and Junfei,

نویسنده مسئول: تلفن: ۰۹۱۷۷۷۱۰۰۳۹ E-Mail: ali_eskandari225@yahoo.com

2009) تقسیم کرد که هدف اصلی این مقاله نیز توسعه مدلی مناسب برای دستیابی به این مهم است.

به هر حال شایان ذکر است که هر دو دسته مدل‌های آماری و عددی به‌منظور مدل‌سازی کیفی سیستم‌های منابع آب مورد استفاده قرار گرفته‌اند؛ اما به دلیل نیاز به اطلاعات گسترده هیدرولوژیکی و کیفیت آب و همچنین انعطاف‌پذیری کمتر مدل‌های عددی نسبت به مدل‌های آماری، در این تحقیق از مدل‌های آماری برای تخمین بهنگام BOD_5 استفاده شده است.

همچنین از بین مدل‌های آماری نیز مدل‌های کلاسیک مانند رگرسیون خطی به دلیل تأثیرپذیری BOD_5 از مشخصه‌های متعدد که دارای رفتار پیچیده و غیرخطی نیز هستند، در بسیاری از موارد با عملکرد مناسبی همراه نبوده (Oliveira-Esquerre, et al., 2004b; Dogan, et al., 2008) و بنابراین لزوم استفاده از روشهای آماری پیشرفته مانند روشهای مبتنی بر هوش مصنوعی گزینه‌ای مناسب برای دستیابی به این مهم است.

در این راستا (Oliveira-Esquerre, et al. ,2004b) با استفاده از شبکه عصبی مصنوعی[۲] (ANN) پرسپترون چندلایه، مدلی برای پیش‌بینی بهنگام غلظت BOD_5 ورودی و خروجی به لاگونی هوادهی شده در برزیل ارائه کردند. در تحقیقی در تصفیه‌خانه‌ای ارتباط بین بو و BOD_5 در فاضلابی واقع در ترکیه با استفاده از ANN مورد تجزیه و تحلیل قرار گرفته و مدلی مناسب برای این منظور گزارش شد (Onkal-Engin, et al., 2005).

Zhao and Chi (2005) با استفاده از مدل ترکیبی شبکه عصبی تأخیر زمانی[۳]–تجزیه‌وتحلیل مؤلفه اصلی[۴] اقدام به ارائه مدلی برای تخمین بهنگام BOD_5 در خروجی تصفیه‌خانهٔ فاضلاب شنیانگ چین کردند.

Dogan و همکاران (2005) با استفاده از ANN مدلی مناسب برای پیش‌بینی BOD_5 ورودی به تصفیه‌خانه‌ای در ترکیه ارائه کردند و نتایج را با مدل رگرسیون خطی چندمتغیره مقایسه کردند. برای این منظور از اطلاعات اکسیژن‌خواهی شیمیایی، دبی ورودی، مواد معلق، نیتروژن کل و فسفر کل استفاده شد که در نهایت برتری ANN نسبت به مدل رگرسیونی گزارش شد. Singh et al., (2009) با قرار دادن BOD_5 به عنوان تابعی از ۱۳ مشخصه کیفی آب، مدلی مناسب با استفاده ANN برای تخمین بهنگام این مشخصه در رودخانه گمتی واقع در هند ارائه کردند. نتایج تحقیق مذکور حاکی از عملکرد مناسب شبکهٔ عصبی در پیش‌بینی BOD_5 بوده و ضریب تعیین معادل ۷۷/. در مرحله تست برای مدل مذکور گزارش شده است.

یکی دیگر از مدل‌های هوشمند که دارای عملکرد قابل قبولی در پیش‌بینی مشخصه‌های پیچیده و غیرخطی است، مدل ماشین بردار پشتیبان[۶] (SVM) است که ایده اصلی آن در دههٔ ۱۹۶۰ توسط Vapnik (1995)، ریاضیدان روسی، مطرح شد.

اگرچه استفاده از SVM در تخمین بهنگام BOD_5 تا به حال گزارش نشده، اما شایان ذکر است که این مدل در دیگر زمینه‌های مدیریت منابع آب با عملکرد موفقیت‌آمیزی همراه بوده است (Babovic, et al., 2000; Bray and Han, 2004; Chen and Yu, 2007a; Lin, et al., 2009a).

Liong and Sivapragasam (2002) عملکرد موفقت‌آمیز SVM برای تخمین سیلاب و مدل‌سازی بارش–رواناب را گزارش کردند. Lin و همکاران (2009b) با استفاده از مدل SVM اقدام به پیش‌بینی جریان ساعتی ورودی به مخزن سدی در شمال تایوان کردند.

Nouri و همکاران (2009a) با مروری انتقادی بر مطالعات انجام گرفته جهت تخمین ضریب انتشار طولی در رودخانه‌های طبیعی، مدلی برتر با استفاده از SVM نسبت به مدل‌های کلاسیک برای تخمین این مشخصه مهم ارائه کردند. در تحقیقی دیگر نیز مدلی مناسب برای تخمین جریان ماهانه با مدل ترکیبی ماشین بردار پشتیبان–تجزیه و تحلیل مؤلفه اصلی توسط Noori, et al. (2011a) توسعه داده شد.

افزون بر مطالعات ذکر شده تحقیقاتی زیادی نیز از برتری نسبی مدل SVM در مقایسه با مدل ANN حکایت دارند که برای نمونه می‌توان به مطالعات Bray and Han (2004)، Liu, et al. (2008))، Behzad, et al. (2010) و Noori, et al. (2011a) اشاره کرد. بنابراین با توجه به مطالب مذکور اهداف اصلی تحقیق حاضر عبارتند از:

(۱) توسعهٔ مدلی مناسب برای پیش‌بینی بهنگام BOD_5 در رودخانه سفیدرود با استفاده از مدل ANN؛

(۲) بررسی عملکرد توابع مختلف آموزش برای بهینه کردن مشخصه‌های شبکه عصبی به منظور تخمین بهنگام BOD_5؛

(۳) امکان‌سنجی استفاده از SVM در تخمین بهنگام BOD_5؛ و در نهایت؛

(۴) مقایسهٔ بین مدل‌های توسعه یافته توسط ANN و SVM و انتخاب مدل برتر.

محدودهٔ مورد مطالعه و اطلاعات مسئله

حوضهٔ رودخانه سفیدرود واقع در شمال و شمال غرب کشور با مساحت ۵۹۱۹۶ کیلومترمربع بین رشته کوههای البرز و زاگرس واقع شده است. مهمترین زهکش این حوضهٔ رودخانه سفیدرود بوده که از رشته کوههای زاگرس سرچشمه گرفته و پس از عبور از شهرها و مناطق مسکونی متعدد و همچنین زمین‌های کشاورزی و صنعتی در نهایت به دریای خزر می‌ریزد.

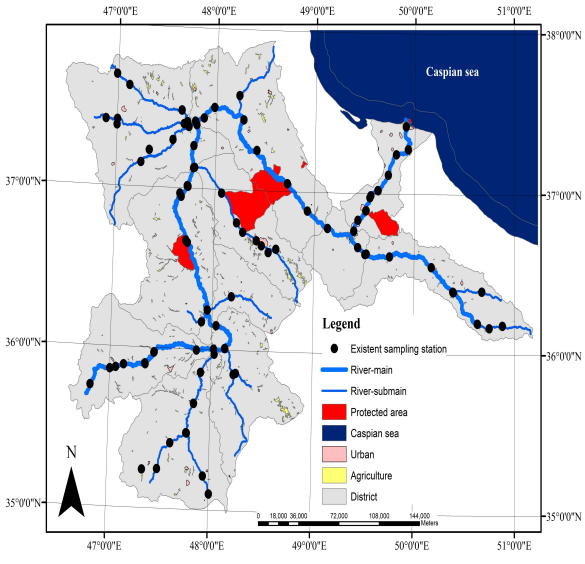
این رودخانه هم اکنون به دلیل ورود آلاینده‌های متعدد شهری، کشاورزی و صنعتی از وضعیت مناسبی برخوردار نبوده و لزوم مطالعات کیفی کارآمد و اتخاذ تدابیر مدیریتی مناسب برای کاهش ورود آلاینده به آن و در نهایت بهبود وضعیت کیفی این رودخانه انکارناپذیر است.

در این تحقیق از اطلاعات برداشت شده در ۹۴ ایستگاه کیفی واقع بر رودخانه سفیدرود و سرشاخه‌های آن به منظور ارائهٔ مدل بهنگام BOD_5 استفاده شده است (شکل شمارهٔ ۱).

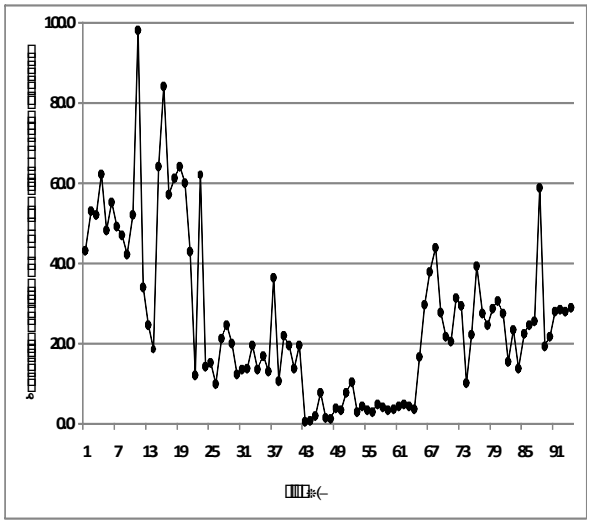
در شکل شمارهٔ (۲) مقادیر BOD_5 بیشینه اندازه‌گیری شده در ایستگاههای مختلف رودخانه سفیدرود نشان داده شده است. با توجه به اطلاعات در دسترس، بردارهای ورودی به مدل‌های ANN و SVM برای تخمین بهنگام BOD_5 به ترتیب مقادیر کمینه، میانگین و بیشینه اکسیژن محلول و همچنین مقادیر میانگین و بیشینهٔ مشخصه‌های هدایت الکتریکی، نیترات و فسفات اندازه‌گیری شده در ۹۴ ایستگاه انتخاب شدند.

مشخصه‌های اکسیژن محلول، نیترات و فسفات به طور مستقیم بر BOD_5 تأثیرگذارند. البته باید توجه کرد در صورتی‌که عامل افزایش هدایت الکتریکی در رودخانه ناشی از منابع آلاینده انسان‌ساخت مانند ورود فاضلاب‌های بهداشتی منطقه باشد، رابطه مشخصی بین هدایت الکتریکی و BOD_5 پیش‌بینی می‌شود.

همچنین با توجه به اثرپذیری مقدار BOD_5 از موقعیت مکانی ایستگاهها در اثر نزدیکی، یا دوری به مراکز تولید مواد آلاینده شهری، صنعتی و کشاورزی؛ مختصات جغرافیایی هر ایستگاه یعنی طول و عرض جغرافیایی نیز به عنوان ورودی به مدل‌های ANN و SVM انتخاب شدند. بنابراین در مجموع اطلاعات پیش‌گفت تعداد ۱۱ ورودی به مدل‌های ANN و SVM برای تخمین بهنگام BOD_5 را تشکیل می‌دهند.



شکل شمارهٔ (۱): محدودهٔ مورد مطالعه حوضهٔ آبریز رودخانه سفیدرود و ایستگاههای نمونه‌برداری (Noori, et al., 2011b)



شکل شمارهٔ (۲): مقادیر بیشینه اندازه‌گیری شده در ایستگاههای مختلف رودخانه سفیدرود

مواد و روشها

شبکهٔ عصبی مصنوعی

با توجه به مراجع کافی برای ANN (Galant, 1994; Haykin, 1999)، در این تحقیق با توضیحاتی اندک در مورد نوع شبکه مورد استفاده و تقسیم‌بندی متغیرهای مسئله برای ورود به مدل‌های مذکور اشاره می‌شود. در این تحقیق از شبکه عصبی پیش‌خور با یک لایهٔ پنهان استفاده شده است. تابع انتقال نیز در

لایه پنهان و خروجی شبکه به ترتیب تابع سیگموئیدی و خطی انتخاب شدند.

همچنین برای بهینه کردن مشخصه‌های شبکه از الگوریتم‌های لونبرگ-مارکویت[۷] (LM)، پس انتشار ارتجاعی[۸] (RP) و گرادیان مزدوج مقیاس‌دار[۹] (SCG) استفاده شد. همچنین برای جلوگیری از مشکل فوق برازشی و کم برازشی شبکه از الگوریتم توقف آموزش[۱۰] (STA) استفاده شد.

خوانندگان محترم برای اطلاعات بیشتر در مورد هر یک از الگوریتم‌های LM، RP و SCG و همچنین روش STA می‌توانند به (Noori, et al., 2011c ; Noori, et al.,2010a) مراجعه نمایند. با فرایند استاندارد سازی، اطلاعات ورودی به شبکهٔ عصبی نیز به بازه ۱- و ۱ محدود شدند.

ماشین بردار پشتیبان

با توجه به لزوم اختصار در مقاله مذکور و همچنین وجود کتابها و مقالات متعدد در ارتباط با نظریهٔ SVM، در ادامه تنها توضیح مختصری در ارتباط با تکنیک ماشین بردار پشتیبان رگرسیونی که در این مقاله از آن استفاده شده است، ارائه خواهد شد. در مدل رگرسیونی SVM، تابعی مرتبط با متغیر وابسته **Y** که خود تابعی از چند متغیر مستقل **x** است براورد می‌شود. مشابه سایر مسائل رگرسیونی، فرض می‌شود که رابطهٔ میان متغیرهای مستقل و وابسته با تابع جبری مانند ($f(\mathbf{x})=\mathbf{w}^T.\phi(\mathbf{x})+b$) $\mathrm{f}(\mathbf{x})$ به علاوه مقداری اغتشاش[۱۱] مشخص شود ($\mathbf{y}=f(\mathbf{x})+noise$).

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$\mathbf{f}(\mathbf{x})$ است.

این مهم با آموزش مدل SVM توسط مجموعه‌ای از نمونه‌ها (مجموعه آموزش) محقق می‌شود.

این روند شامل بهینه‌سازی متوالی تابع خطاست. بسته به تعریف این تابع خطا دو نوع مدل SVM تعریف می‌شود:

SVM رگرسیونی نوع ۱ (به عنوان ε-SVM رگرسیونی نیز شناخته می‌شود)؛

SVM رگرسیونی نوع ۲ (به عنوان υ-SVM رگرسیونی شناخته می‌شود).

شایان ذکر است که در این تحقیق از مدل ε-SVM رگرسیونی، به دلیل کاربرد گسترده آن در مطالعات رگرسیونی، برای تخمین بهنگام BOD_5 استفاده شده است.

بنابراین برای محاسبه **w** و b لازمست تابع خطا (معادله ۱) در مدل ε-SVM با درنظرگرفتن شرایط مندرج در معادلهٔ ۲ بهینه شود.

$$\frac{1}{2}\mathbf{w}^T.\mathbf{w}+C\sum_{i=1}^{N}\xi_i+C\sum_{i=1}^{N}\xi_i^{\bullet} \qquad (۱)$$

$$\begin{aligned}&\mathbf{w}^T.\phi\left(\mathbf{x}_i\right)+b-\mathbf{y}_i\le\varepsilon+\xi_i^{\bullet} \qquad (۲)\\&\mathbf{y}_i-\mathbf{w}^T.\phi\left(\mathbf{x}_i\right)-b\le\varepsilon+\xi_i\\&\xi_i,\xi_i^{\bullet}\ge 0\quad,\quad i=1,\ldots,N\end{aligned}$$

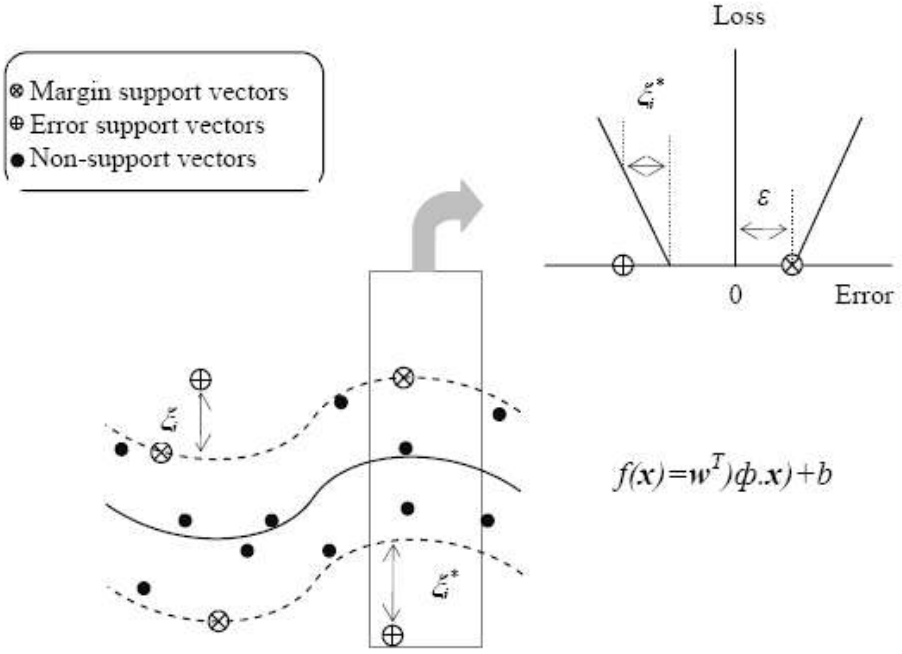
در معادلات بالا C عددی صحیح مثبت است که عامل تعیین جریمه در هنگام رخ دادن خطای آموزش مدل است.

ϕ تابع کرنل[۱۲]، N تعداد نمونه‌ها و دو مشخصه ξ_i و $\xi_i^{\bullet}$ متغیرهای کمبود[۱۳] هستند که حد بالا و پایین خطای آموزش مرتبط با مقدار خطای مجاز ε را مشخص می‌کنند.

در مسائل پیش‌بینی می‌شود که داده‌ها، درون بازه مرزی ε قرار گیرند (شکل شمارهٔ ۳). حال اگر داده‌ای خارج از بازه ε قرار گرفت آنگاه یک خطا معادل ξ_i و $\xi_i^{\bullet}$ وجود خواهد داشت.

ذکر این نکته نیز لازم است که مدل SVM مشکلات ناشی از کم‌تخمینی[۱۴] و فوق‌برازشی[۱۵] را با کمینه کردن همزمان دو ترم $\mathbf{w}^T.\mathbf{w}/2$ و خطای آموزشی، یعنی $C\sum_{i=1}^{N}(\xi_i+\xi_i^{\bullet})$ را در معادلهٔ ۱ حل می‌کند.

بنابراین با معرفی ۲ ضریب لاگرانژ a_i و a_i^* مسئلهٔ بهینه‌سازی با حداکثرسازی عددی تابع درجه دوم زیر (معادله ۳) با شرایط معادله (۴) حل خواهد شد.



شکل شمارهٔ (۳): مدل SVM رگرسیونی

$$\sum_{i=1}^{N} y_i(\alpha_i-\alpha_i^*)-\varepsilon\sum_{i=1}^{N}(\alpha_i+\alpha_i^*)-0.5\sum_{i,j=1}^{N}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\phi(x_i)^T\phi(x_j) \quad (۳)$$

$$\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)=0$$
$$0\le\alpha_i\le C \quad (۴)$$
$$0\le\alpha_i^*\le C \quad , \quad i=1,2,...,N$$

تابع هدف بالا در معادلهٔ (۳) تابع محدب است و بنابراین جواب معادلهٔ (۳) یکتا و بهینه خواهد بود. پس از تعریف ضرایب لاگرانژ در معادلهٔ (۳) مشخصه‌های **w** و b در مدل SVM رگرسیونی با استفاده از شرایط تئوری کراش-کوهن-تاکر[۱۶] محاسبه می‌شوند (Fletcher, 1987) که در آن $\mathbf{W}=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)\phi(\mathbf{x}_i)$ است. در نتیجه برای مدل SVM رگرسیونی خواهیم داشت:

$$\mathbf{W}=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)\phi(\mathbf{x}_i)^T\phi(\mathbf{x})+b \quad (۵)$$

باید توجه داشت که ترم‌های لاگرانژ $(\alpha_i-\alpha_i^*)$ می‌توانند صفر، و یا غیرصفر باشند. بنابراین فقط مجموعه داده‌هایی که ضرایب $\bar{\alpha}_i$ آنها غیرصفر است در معادلهٔ رگرسیون نهایی وارد می‌شوند و این مجموعه داده‌ها 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 $|\bar{\alpha}_i|$ آنها کمتر از C باشد بردارهای پشتیبان حاشیه‌ای[۱۷] نامیده می‌شوند. هنگامی که مقدار $|\bar{\alpha}_i|$ بردارهای پشتیبان برابر مقدار C باشد، به عنوان بردار پشتیبان خطا[۱۸] ، یا بردار پشتیبان کراندار شناخته می‌شود. بردارهای پشتیبان حاشیه‌ای در حاشیه مرز غیرحساس یافت می‌شوند، در حالی‌که بردارهای پشتیبان خطا خارج از بازه هستند (شکل شمارهٔ ۳). در نهایت تابع SVM رگرسیونی را می‌توان به فرم زیر بازنویسی کرد:

$$f(\mathbf{x})=\sum_{i=1}^{N}\bar{\alpha}_i\phi(\mathbf{x}_i)^T\phi(\mathbf{x})+b \quad (۶)$$

در معادلهٔ (۶) محاسبه $\phi(\mathbf{x})$در فضای مشخصهٔ آن ممکن است بسیار پیچیده باشد. برای حل این مشکل روند معمول در مدل SVM رگرسیونی انتخاب یک تابع کرنل به صورت

$K(\mathbf{x}_i,\mathbf{x})=\phi(\mathbf{x}_i)^T\sqrt{b^2-4ac}$ است. می‌توان از توابع مختلف کرنل برای ساخت انواع مختلف مدل ε-SVM استفاده کرد. انواع رایج توابع کرنل قابل استفاده در مدل SVM رگرسیونی عبارتند از: کرنل چندجمله‌ای با ۳ مشخصه هدف، کرنل سیگموئیدی[۱۹] شامل ۲ مشخصه هدف و کرنل توابع پایه شعاعی[۲۰] (RBF) با یک مشخصهٔ هدف. در این تحقیق با توجه به عملکرد مطلوب آن در مطالعات قبلی (Lin, et al., 2009b; Noori, et al., 2011a)، تابع کرنل RBF به همراه γ که در معادله (۷) داده شده است به عنوان انتخاب برتر مورد استفاده قرار گرفته است. شایان ذکر است که در معادلهٔ (۷) γ کنترل‌کننده میزان نوسان تابع گوسی و همچنین کنترل‌کننده نتایج و تعمیم‌دهندهٔ مدل SVM است.

$$K(x_i,x)=\exp\left(-\gamma\left|x_i-x\right|^2\right) \qquad (۷)$$

بحث و نتایج

توسعهٔ مدل شبکه عصبی

در این تحقیق برای ورود اطلاعات به شبکهٔ عصبی به دلیل استفاده از تابع محرک[۲۱] تانژانت سیگموئید در لایه پنهان شبکه و با توجه به این واقعیت که تابع مذکور محدود به بازهٔ (۱+,۱−) است، اطلاعات ورودی و مشخصه هدف به بازهٔ مذکور استانداردسازی شدند. همچنین تابع محرک در لایهٔ خروجی شبکه نیز تابع خطی انتخاب گردید.

برای دستیابی به بهترین مدل ANN برای تخمین بهنگام BOD_5 می‌باید بهترین معماری شبکه با استفاده از سعی و خطا تعیین شود. ذکر این نکته لازم است که برخی از مراجع به لحاظ تئوریک و عملی تعداد یک لایهٔ پنهان در شبکه را برای تخمین هر مشخصه غیرخطی و پیچیده کافی دانسته‌اند (Haykin, 1999; Noori, et al., 2009b)، اما تعداد نرون‌ها که به عنوان واحدهای محاسباتی ANN هستند در لایهٔ پنهان شبکه لازم است با روند سعی و خطا به‌دست آیند. البته در انتخاب نرون‌ها باید دقت کرد که تعداد زیاد آنها در لایه پنهان منجر به ناپایداری شبکه شده و استفاده از آنرا در مراحل بعدی برای تخمین بهنگام مشخصه هدف که در این تحقیق BOD_5 است با مشکل مواجه می‌کند. به‌طور کلی تعداد نرون‌ها در لایهٔ پنهان با حجم اطلاعات ورودی به شبکه متناسب است.

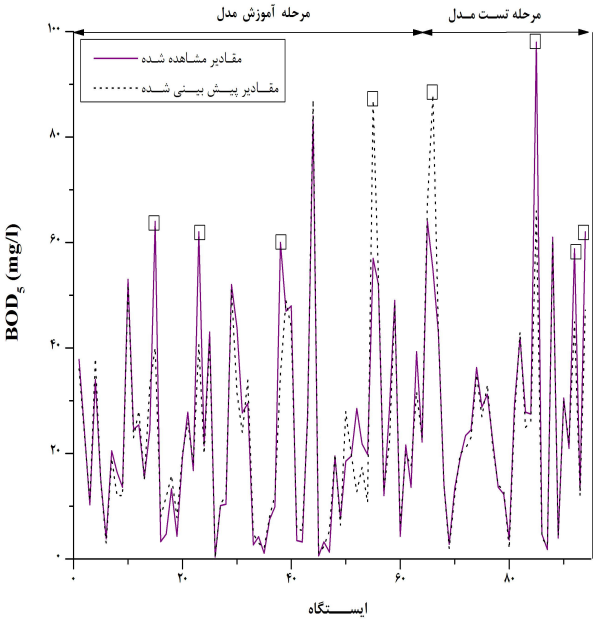
بنابراین با توضیحات مذکور و با توجه به حجم اطلاعات ورودی به ANN در این تحقیق، تعداد نرون بین ۲ تا ۱۰، برای توسعهٔ مدل مناسب شبکه عصبی انتخاب شد. در مرحلهٔ بعد لازمست که مشخصه‌های شبکه یعنی وزن‌ها و بایاس در مدل ANN بهینه شوند.

برای این منظور همانطور که ذکر شد از الگوریتم‌های LM و SCG به دلیل کارآمدی مناسب و RP به دلیل سرعت بالا در بهینه کردن مشخصه‌های شبکه استفاده گردید که نتایج به‌دست آمده برای هر الگوریتم به همراه ارزیابی دقت مدل‌ها بر مبنای دو شاخص آماری ضریب همبستگی پیرسون[۲۲] (R) و مقدار میانگین مجذور مربعات خطا[۲۳] (RMSE) در جدول شمارهٔ (۱) ارائه شده است. بر مبنای این جدول می‌توان مشاهده کرد که هر سه مدل توسعه یافته از مقادیر مطلوب R و همچنین RMSE در مرحله آموزش و تست برخوردار بوده و نتایج هر سه مدل تقریباً نزدیک به یکدیگر است. به هر حال با بررسی بیشتر مشخص می‌شود که عملکرد مدل با الگوریتم LM (یعنی مدل ANN (LM)) اندکی از بقیه بهتر بوده و مدل‌های دربرگیرنده الگوریتم RP و SCG در مراحل بعدی از دقت قرار دارند.

همچنین در ادامه برای قضاوت بهتر در مورد دقت مدل منتخب در این مرحله، مقادیر مشاهداتی BOD_5 در مقابل مقادیر پیش‌بینی شده توسط این مدل در شکل شمارهٔ (۴) نشان داده شده است.

جدول شمارهٔ (۱): نتایج مراحل آموزش و تست مدل‌های ANN و SVM

مدل	الگوریتم بهینه‌سازی	مرحله آموزش		مرحله تست	
		RMES	R	RMSE	R
ANN	LM	۷/۵	۰/۹۳	۱۱/۶	۰/۹۰
	RP	۷/۹	۰/۹۰	۱۲/۱	۰/۸۷
	SCG	۸/۲	۰/۸۸	۱۲/۸	۰/۸۵
SVM	جستجوی شبکه دو مرحله‌ای	۴/۶	۰/۹۷	۹/۱۶	۰/۹۵



شکل شمارهٔ (۴): نتایج مراحل آموزش و تست مدل ANN (LM) (بیانگر نقاط حدی بالا هستند که مدل در این نقاط از دقت مطلوبی برخوردار نبوده است)

با دقت بیشتر در این شکل مشخص است که مدل منتخب اگرچه از دقت مناسبی برای پیش‌بینی بهنگام BOD_5 برخوردار است، اما در تخمین مقادیر حدی و بویژه مقادیر بیشینهٔ این مشخصه کیفی نسبت به دیگر مقادیر از عملکرد ضعیف‌تری برخوردار است. واقعیت مذکور در نقاطی حدی بیشینه در شکل شمارهٔ (۴) توسط علائمی مربعی برجسته شده است.

از این شکل مشخص است که برای نمونه تفاضل بین مقادیر پیش‌بینی شده و الگوهای مشاهده‌ای با مدل منتخب در ۴ مورد از مرحله آموزش و همچنین در سه مورد از مرحله تست مدل با خطای زیادتری نسبت به بقیه الگوها برخوردار است.

با مروری بر دیگر مطالعات انجام شده در زمینهٔ پیش‌بینی BOD_5 با مدل ANN مشخص می‌شود که مدل ارائه شده در این تحقیق نسبت به مدل‌های معرفی شده توسط دیگر محققان از دقت قابل قبولی برخوردار و در اکثر موارد نیز دارای دقت بالاتری است.

به عنوان مثال مقدار R در مرحلهٔ تست برای پیش‌بینی BOD_5 در مطالعات (Oliveira-Esquerre, et al.,2002، Dogan et al.,2008) و (Singh, et al.,2009) به ترتیب معادل ۰/۶۰، ۰/۹۲ و ۰/۸۸ گزارش شده است، در حالی‌که این عامل در تحقیق حاضر برای مدل ANN با تابع آموزش LM در مرحلهٔ تست معادل ۰/۹۰ به‌دست آمده است.

توسعهٔ مدل ماشین بردار پشتیبان

در این مرحله از تحقیق نیز قبل از ورود اطلاعات به مدل SVM رگرسیونی به دلیل تفاوت زیاد در دامنهٔ تغییرات هر مشخصه نسبت به دیگر مشخصه‌ها و برای همسان‌سازی اطلاعات ورودی و خروجی به مدل لازمست استانداردسازی اطلاعات صورت گیرد. بنابراین مشابه مرحله قبل استانداردسازی اطلاعات در بازه (۱+،۱−) انجام پذیرفت. در مرحلهٔ بعد بهینه‌سازی مشخصه‌های مدل SVM-ε یعنی مقادیر ε و C انجام می‌گیرد.

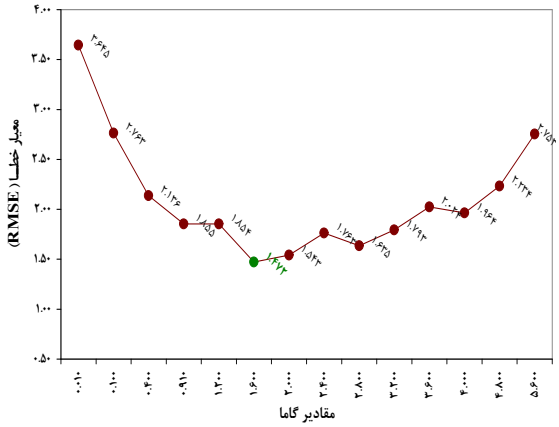
همان‌طور که ذکر شد تابع کرنل مورد استفاده در این تحقیق تابع RBF انتخاب شد که در این تابع نیز باید مشخصه γ بهینه شود. بنابراین در حالت کلی برای تخمین بهنگام BOD_5 در رودخانه سفیدرود توسط مدل SVM رگرسیونی، لازمست که مقادیر بهینه سه مشخصه مذکور به‌دست آیند. در این تحقیق برای رسیدن به این هدف، دو مشخصه ε و C توسط الگوریتم بهینه‌سازی جستجوی شبکه[۲۴] و مشخصه γ نیز به صورت سعی و خطا بهینه 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 Chen and Yu (2007b) از برنامه اصلاح شده الگوریتم جستجوی شبکه که به نام الگوریتم جستجوی شبکه دو مرحله‌ای[۲۵] معروف است به همراه اعتبارسنجی متقاطع[۲۶] 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.

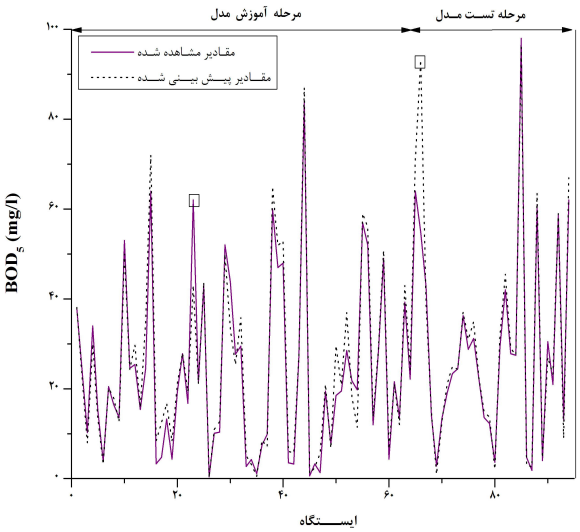
روند مذکور برای دیگر مقادیر γ نیز تکرار شد و بدینطریق مدل‌های متفاوتی با تغییر در مقدار γ حاصل شدند. حال می‌توان از بین مدل‌های توسعه داده شده برای بهینه کردن مشخصه‌های ε، C و γ مدل با کمترین خطا را تعیین کرده و مشخصه‌های آنرا به عنوان مقادیر بهینه ε، C و γ انتخاب کرد.

شکل شماره (۵) مقادیر خطای هر مدل (در قالب معیار RMSE) به ازای مقادیر مختلف γ را نشان می‌دهد. از شکل شمارهٔ (۵) مشخص است که مدل توسعه داده شده به ازای γ ،معادل ۱/۴۷۲ دارای کمترین خطاست. بنابراین با انتخاب مدل با مقدار γ، معادل ۱/۴۷۲ می‌توان مقادیر متناظر ε و C با این γ را نیز تعیین

کرد که در این تحقیق به ترتیب معادل ۰۳۷/. و ۱۳ به‌دست آمده است. در ادامه نیز با لحاظ مقادیر بهینه محاسبه شده سه مشخصه ε، C و γ در مدل SVM رگرسیونی، مدل پیش‌بینی بهنگام BOD_5 در رودخانهٔ سفیدرود توسعه داده شد. نتایج مراحل آموزش و تست این مدل در جدول شمارهٔ (۱) نشان داده شده است. همچنین برای قضاوت بهتر در مورد عملکرد مدل در مراحل آموزش و تست، کردار مربوط به مقادیر مشاهداتی BOD_5 در مقابل مقادیر پیش‌بینی شده با این مدل در شکل شمارهٔ (۶) نشان داده شده است.



شکل شمارهٔ (۵): مقادیر خطای هر مدل به ازای مقادیر مختلف γ برای مدل SVM



شکل شمارهٔ (۶): نتایج مراحل آموزش و تست مدل SVM رگرسیونی (بیانگر نقاط حدی بالا هستند که مدل در این نقاط از دقت مطلوبی برخوردار نبوده است)

در شکل شمارهٔ (۶) مشخص است که مدل SVM رگرسیونی از دقت مناسبی برای پیش‌بینی بهنگام BOD_5 برخوردار است و تقریباً در تمامی موارد، اختلاف بین مقادیر پیش‌بینی شده BOD_5 با مدل با مقادیر مشاهداتی بسیار اندک است. واقعیت مذکور از نتایج مندرج در جدول شمارهٔ (۱) نیز قابل استنتاج است، به نحوی که در این جدول مقدار آمارهٔ R در مراحل آموزش و تست مدل SVM تقریباً نزدیک به ۱ است. شایان ذکر است که مقدار ایده‌ال دو آماره R و RMSE برای ارزیابی مدل‌ها به ترتیب برابر یک و صفر است. همچنین با توجه به مقادیر حدی و بویژه مقادیر بیشینه این مشخصه کیفی مشخص می‌شود که مدل به استثنای دو مورد که با علائمی مربعی در شکل شمارهٔ (۶) برجسته شده است، در بقیهٔ موارد از دقت قابل قبولی برخوردار است.

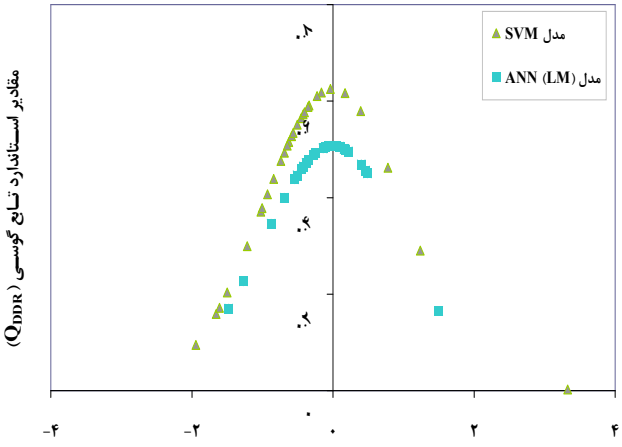
ارزیابی و مقایسه عملکرد مدل‌های توسعه داده شده جهت تخمین بهنگام BOD_5

در مراحل قبل برای تخمین بهنگام BOD_5 در رودخانه سفیدرود مدل‌های مختلفی با ANN با تغییر در الگوریتم آموزش آن و همچنین مدل SVM رگرسیونی توسعه داده شد. نتایج توسعه مدل شبکه عصبی حاکی از برتری عملکرد مدل (LM) ANN بود. نتایج تحقیقات مشابه نیز عملکرد برتر الگوریتم آموزش LM نسبت به دیگر الگوریتم‌های آموزشی را مشخص می‌کند (Noori, et al., 2010a; Noori, et al., 2011c). همچنین نتایج به‌دست آمده از توسعه دو مدل ANN و SVM بر مبنای دو آماره R و RMSE حاکی از برتری نسبی مدل SVM نسبت به ANN بود. به هر حال باید توجه داشت که قضاوت فقط بر مبنای این دو آماره در برخی از موارد نیاز به بررسی بیشتری دارد زیرا آنها فقط نشان‌دهندهٔ متوسطی از خطا در مدل بوده و علاوه بر این مقدار این دو آماره متأثر از تعداد الگوهای انتخابی برای آموزش و تست مدل است. بنابراین در ادامه برای قضاوت بهتر در مورد عملکرد دو مدل توسعه داده شده (LM) ANN و SVM از آمارهٔ نسبت تفاوت توسعه داده شده[۲۷] (DDR) با (Noori, et al.,2010b) در مراحل تست هر مدل استفاده شده است (معادلهٔ ۸). جزئیات بیشتر در مورد این آماره در مراجع مربوط در دسترس است.

$$DDR = \left(\frac{\text{Predicted Value}}{\text{Observed Value}} \right) - 1 \qquad (۸)$$

مطابق معادلهٔ (۸) اگر DDR=0، آنگاه مقادیر پیش‌بینی شده معادل مقادیر اندازه‌گیری شده است. اگر DDR>0، مقدار پیش‌بینی شده از مقدار مشاهداتی بیشتر و اگر DDR<0، مقدار پیش‌بینی شده از مقدار مشاهداتی کمتر است. برای قضاوت بهتر در این مورد و دید

بهتر، می‌توان تابع گوسی مقادیر DDR را محاسبه و به صورت توزیع نرمال استاندارد رسم کرد. برای این منظور ابتدا لازم است مقادیر DDR استاندارد شده و سپس با استفاده از تابع گوسی، مقادیر نرمال شده DDR (Q_{DDR}) محاسبه شود. توزیع نرمال استاندارد برای هر یک از مدل‌های منتخب ANN و SVM در شکل شمارهٔ (۷) به نمایش گذاشته شده است.



شکل شمارهٔ (۷): نمودار توزیع نرمال استاندارد شده مقادیر DDR برای مدل‌های ANN (LM) و SVM

در شکل شمارهٔ (۷) تمایل بیشتر گراف به خط مرکزی و بزرگتر بودن مقدار ماکزیمم Q_{DDR} برای هر مدل نشان‌دهندهٔ دقت بیشتر مدل است. ماکزیمم مقدار Q_{DDR} برای مدل‌های ANN (LM) و SVM به ترتیب ۰/۵۱ و ۰/۶۲ است. با توجه به نتایج به دست آمده از این شکل واضح است که مدل SVM از عملکرد بهتری نسبت به مدل ANN (LM) برخوردار است.

از مقایسه نتایج به‌دست آمده از آمارهٔ DDR با آماره‌های R و RMSE مشخص می‌شود که نتایج این سه آماره در این تحقیق با یکدیگر متناسب بوده و هر سه آماره حاکی از برتری مدل SVM در پیش‌بینی بهنگام BOD_5 هستند. بنابراین با توجه به نتایج این تحقیق مدل SVM به عنوان بهترین مدل برای تخمین بهنگام BOD_5 در رودخانه سفیدرود توصیه می‌شود.

نتیجه‌گیری

هدف اصلی مقالهٔ مذکور توسعهٔ مدلی مناسب برای حذف محدودیت زمانی در ارتباط با اندازه‌گیری یکی از شاخص‌های مهم کیفیت آب، یعنی BOD_5 قرار داده شد. برای این منظور از مدل‌های ANN و SVM استفاده شد و برای اجرای این مدل‌ها اطلاعات رودخانه سفیدرود مورد استفاده قرار گرفت. یافته‌های این تحقیق را می‌توان در بندهای زیر خلاصه کرد:

- نتایج بررسی‌های مربوط به ارزیابی عملکرد الگوریتم‌های بهینه‌سازی مختلف بر دقت خروجی ANN نشان‌دهندهٔ عملکرد برتر الگوریتم آموزش LM بود.
- مدل منتخب ANN (LM) از دقت مناسبی در تخمین بهنگام BOD_5 برخوردار بود. به هر حال این مدل در پیش‌بینی مقادیر حدی بیشینه از عملکرد ضعیف‌تری برخوردار است.
- با استفاده از الگوریتم بهینه‌سازی جست‌وجوی شبکهٔ دو مرحله‌ای، مقادیر بهینهٔ مشخصه‌های مدل SVM یعنی ε، C و γ به ترتیب معادل ۰/۰۳۷، ۱۳ و ۱/۴۷۲ به‌دست آمد.
- ارزیابی عملکرد مدل SVM بر مبنای دو آماره R و RMSE مبین دقت بیشتر این مدل نسبت به مدل ANN (LM) بود.
- بررسی دقیق‌تر عملکرد دو مدل ANN (LM) و SVM بر مبنای آماره نسبت تفاوت توسعه یافته نیز حاکی از برتری مدل SVM در این تحقیق بود.

در نهایت با توجه به نتایج به‌دست آمده در این تحقیق مدل SVM برای پیش‌بینی بهنگام BOD_5 در رودخانهٔ سفیدرود توصیه شد.

یادداشت‌ها

1- 5-Days Biochemical Oxygen Demand (BOD_5)
2- Artificial Neural Network (ANN)
3- Time-Delay Neural Network
4- Principal Component Analysis
5- Proper Orthogonal Decomposition
6- Support Vector Machine (SVM)
7- Levenberg-Marquardt (LM)
8- Resilient back-Propagation (RP)
9- Scaled Conjugate Gradient (SCG)
10- Stop Training Algorithm (STA)
11- Noise
12- Kernel Function
13- Slack Variable
14- Under Fitting
15- Over Fitting
16- Karush-Kuhn-Tucker
17- Margin Support Vector

18- Error Support Vector
19- Sigmoid Kernel
20- Radial Basis Function (RBF)
21- Activated Function
22- Pearson Correlation Coefficient (R)
23- Root Mean Square Error (RMSE)
24- Grid Search
25- Two-Steps Grid Search
26- Cross-Validation
27- Developed Discrepancy Ratio (DDR)

منابع مورد استفاده

Babovic,V., et al .2000. From global to local modelling: a case study in error correction of deterministic models. In: Proceeding of fourth international conference on hydroinformatics, Iowa City, USA: CD-ROM, IAHR.

Behzad,M., et al .2010. Comparative study of SVMs and ANNs in aquifer water level prediction. Journal of Computing in Civil Engineering, 24(408), 1943-5487.

Beltran,J.L., et al .1998. Multivariate calibration of polycyclic aromatic hydrocarbon mixtures from excitation-emission fluorescence spectra. Anal. Chim. Acta 373, 311-319.

Bray,M., D.,Han .2004. Identification of support vector machines for runoff modeling. Journal of Hydroinformatics 6, 265-280.

Chen,S.T., P.S.,Yu .2007a. Pruning of support vector networks on flood forecasting. Journal of Hydrology 347, 67-78.

Chen,S.T., P.S.,Yu .2007b. Real-time probabilistic forecasting of flood stages. Journal of Hydrology 340, 63-77.

Dogan,E., et al. 2008. Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand. Environmental Progress & Sustainable Energy 27, 439-446.

Einax,J.W., et al. 1999. Quantitative description of element concentrations in longitudinal river profiles by multiway PLS models. Fres. J. Anal. Chem. 363, 655-661.

Fletcher,R. 1987. Practical Methods of Optimization. Wiley, New York.

Gallant,S.I. 1993. Neural Network Learning and Expert Systems. MIT Press, Cambridge.

Haykin,S. 1999. Neural Networks: A Comprehensive Foundation. Prentice Hall, New Jersey.

Honggui,H., Q.,Junfei .2009. Biological oxygen demand (BOD) soft measuring based on dynamic neural network (DNN): A simulation study. Proceedings of the 7th Asian Control Conference, Hong Kong, China, August 27-29.

Lin,G.F., et al. 2009a. Effective forecasting of hourly typhoon rainfall using support vector machines. Water Resources Research DOI: 10.1029/2009WR007911.

Lin,G.F., et al. 2009b. Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods. Journal of Hydrology. 372, 17-29.

Liong,S.Y., C.,Sivapragasam .2002. Flood stage forecasting with SVM. Journal of the American Water Resources Association. 38, 173-186.

Liu,H., et al .2008. Soil water content forecasting by ANN and SVM hybrid architecture. Environmental Monitoring and Assessment. 13 (1-3), 257-262.

Nouri,R., et al. 2009a. Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. Environmental Engineering Science 26, 1503-1510.

Nouri,R., et al .2009b. Results uncertainty of solid waste generation forecasting by hybrid of wavelet transform-ANFIS and wavelet transform-neural network. Expert Systems with Applications 36, 9991-9999.

Nouri,R., et al. 2010a. Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic. Expert Systems with Applications 37, 5856-5862.

Nouri,R., et al. 2010b. Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration. Atmospheric Environment 44, 476-482.

Nouri,R., et al. 2011a. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. Journal of Hydrology 401, 177-189.

Nouri,R., et al .2011b. Development and application of reduced-order neural network model based on proper orthogonal decomposition for BOD_5 monitoring: active and online prediction. Environmental Progress & Sustainable Energy DOI 10.1002/ep.10611.

Nouri,R., et al .2011c. A framework development for predicting the longitudinal dispersion coefficient in natural streams using artificial neural network. Environmental Progress & Sustainable Energy DOI 10.1002/ep.10478.

Oliveira-Esquerre,K.P., et al. 2004a. Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part I. Linear approaches. Chemical Engineering Journal 104, 73-81.

Oliveira-Esquerre,K.P., et al. 2004b. Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part II. Nonlinear approaches. Chemical Engineering Journal 105, 61-69.

Oliveira-Esquerre,K.P., et al .2002. Simulation of an industrial wastewater treatment plant using artificial neural networks and principal component analysis. Brazilian Journal of Chemical Engineering 19, 365-370.

Onkal-Engin,G., et al .2005. Determination of the relationship between sewage odourand BOD by neural networks. Environmental Modelling & Software 20, 843-850.

Rastogi,S., et al. 2003. Development and characterization of a novel immobilized microbial membrane for rapid determination of biochemical oxygen demand load in industrial waste-waters, Biosensors and Bioelectronics 18, 23-29.

Singh,K.P., et al. 2009. Artificial neural network modeling of the river water quality-A case study. Ecological Modelling 220, 888-895.

Sohn,M.J., et al. 1995. Rapid estimation of biochemical oxygen demand using a microbial multi-staged bioreactor, Analytica Chimica Acta 313, 221-227.

Vapnik,V. 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Zhao,L., T.,Chai .2005. Wastewater BOD forecasting model for optimal operation using robust time-delay neural network. LNCS 3498, 1028-1033.