نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۰، شماره ۳، مرداد – شهریور ۱۳۹۵، ص. ۷۵۷-۷۴۳

Journal of Water and Soil
Vol. 30, No. 3, Jul.-Agu. 2016, p. 743-757

الگوبندی متغیرهای کیفی آب با روش داده مبنا در رودخانه سفیدرود

شیما سلیمانی[۱]- امید بزرگ حداد[۲]- مجتبی مروج[۳*]

تاریخ دریافت: ۱۳۹۳/۱۱/۰۶
تاریخ پذیرش: ۱۳۹۴/۰۷/۱۹

چکیده

افزایش برداشت از منابع آب سطحی، به عنوان در دسترس‌ترین منبع آب و افزایش تخلیه پساب‌ها به این منابع منجر به کاهش کیفیت آب‌های سطحی شده است. لذا پایش و الگوبندی کیفیت منابع سطحی بیش از پیش احساس می‌شود. از آنجایی که روش‌های داده‌مبنا توانایی بالایی در الگوبندی دارند، در این تحقیق ابتدا با دو روش پیش‌پردازش ضریب همبستگی و تجزیه مولفه اصلی (PCA)، ورودی‌های روش رگرسیون بردار پشتیبان حداقل مربعات (LSSVR) تعیین و سپس الگوریتم ژنتیک-رگرسیون بردار پشتیبان حداقل مربعات (GA-LSSVR) توسعه داده شد که توانایی تنظیم خودکار و بهینه ضرایب روش LSSVR را دارد. الگوریتم GA-LSSVR برای الگوبندی متغیرهای کیفی Na^+، K^+، Mg^{2+}، So_4^{2-}، Cl^-، pH، هدایت الکتریکی (EC) و مجموع باقی‌مانده خشک (TDS) رودخانه سفیدرود برای طول دوره آماری ۲۰ سال (۱۳۶۴-۱۳۸۴) به کار گرفته شد. نتایج الگوبندی با الگوریتم GA-LSSVR و روش‌های پیش‌پردازش ضریب همبستگی و PCA نشان داد که مقادیر ضریب تشخیص (R^2) متغیرهای کیفی EC، TDS، Cl^- و Na^+ به ترتیب ۰/۹۸، ۰/۹۸، ۰/۹۷ و ۰/۹۴ می‌باشد که الگوبندی این متغیرها نسبت به دیگر متغیرهای کیفی از دقت بیش‌تری برخوردار است. در مجموع با توجه به مثبت بودن مقادیر نش-سایتکلیف (*NS*) هر دو روش از قابلیت بالایی برای انتخاب ورودی‌های الگو برخوردار هستند.

واژه‌های کلیدی: الگوریتم GA-LSSVR، ضریب همبستگی پیرسون، روش PCA

مقدمه

اهمیت کیفیت آب سطحی، به عنوان یکی از مهم‌ترین و آسیب‌پذیرترین منابع تامین آب، امری کاملاً بدیهی است. متاسفانه ورود آلاینده‌ها به پیکره‌های آبی در سال‌های اخیر به دلیل رشد شهرنشینی، توسعه اقتصادی و افزایش تولیدات صنعتی، افزایش یافته است. پارامترهای کیفی آب بیان‌کننده خصوصیات فیزیکی، شیمیایی و بیولوژیکی آب هستند. لذا اهمیت پایش پارامترهای کیفی آب رودخانه‌ها بیش از پیش احساس می‌شود. هرکدام از مصارف گوناگون آب مانند کشاورزی، شرب، پرورش ماهی، صنعت، آب‌زیان و غیره نیازمند آب با یک کیفیت مشخص هستند که با نمونه‌برداری های مکرر، آزمایش و تحلیل نتایج مشخص می شود. اما هزینه نمونه‌برداری از آب‌های سطحی، اندازه‌گیری پارامترهای کیفی در محیط آزمایشگاه، خطاهای انسانی و غیره از جمله مشکلات موجود در تخمین غلظت پارامترهای کیفی هستند. به همین منظور برای الگوبندی پارامترهای کیفی آب روش‌های مختلفی وجود دارند که در این بین روش‌های داده‌مبنا در ده‌های اخیر مورد توجه پژوهش‌گران قرار گرفته است.

استفاده از روش رگرسیون بردار پشتیبان[۴] (SVR) در دهه اخیر در هیدرولوژی افزایش یافته است. راگهواندرا و دکا (۱۶) کاربرد SVR در هیدرولوژی را مرور کردند. آن‌ها روش SVR را در توسعه الگو بارش-رواناب، پیش‌بینی جریان رودخانه، تبخیر، تراز آب دریاچه و مخزن، سیلاب و خشک‌سالی به کار بردند.

روش LSSVR[۵] در پیش‌بینی و الگوبندی پارامترهای کیفی مختلف آب به‌کار گرفته شده است. یورانگ و لیانگژون (۲۷) با روش LSSVR به مطالعه پیش‌بینی الگو کیفیت آب در رودخانه لیوگزی[۶] در چین پرداختند. آن‌ها پارامترهای اکسیژن مورد نیاز شیمیایی[۷] (COD) و اکسیژن محلول[۸] (DO) را با یک الگوریتم

۱، ۲ و ۳- به ترتیب دانشجوی کارشناسی ارشد، دانشیار و دانشجوی دکتری گروه مهندسی آبیاری و آبادانی، دانشکده کشاورزی و منابع طبیعی، دانشگاه تهران
(*-نویسنده مسئول: Email: mojtaba.moravej@gmail.com)

4- Support Vector Regression
5- Least Squares Support Vector Regression
6- Liuxi
7- Chemical Oxygen Demand
8- Dissolved Oxygen

تلفیقی متشکل از روش LSSVR و الگوریتم بهینه‌بندی مجموعه ذرات[1] (PSO) پیش‌بینی کردند. آن‌ها در این تحقیق نشان داده شد که روش LSSVR توانایی غلبه بر ضعف‌های روش پرسپترون چند لایه[2] (MLP) را دارد. سینگ و همکاران (۱۸) از روش خوشه‌بندی بردار پشتیبان[3] (SVC & SVR) برای بهینه‌بندی برنامه‌ریزی پایش داده‌های کیفیت آب سطحی استفاده کردند. در این تحقیق ۱/۵۰۰ نمونه داده کیفی از ۱۰۰ نقطه پایش در طی ۱۵ سال در شهر لوکنو[4] در هند بررسی شد. هدف این تحقیق کاهش نقاط و تعداد اندازه‌گیری مقادیر کیفی و پیش‌بینی مقادیر تقاضای اکسیژن بیوشیمایی[5] (BOD) بود. آن‌ها الگوی جدید برای برنامه‌ریزی پایش در آینده به‌دست آورند که ۹۲/۵ درصد نقاط اندازه‌گیری داده‌های کیفی را کاهش می‌دهد و همچنین ابزاری برای پیش‌بینی BOD با همبستگی بالا ارائه دادند. تان و همکاران (۲۲) در چین، به پیش‌بینی مقادیر فسفر با روش LSSVR پرداختند و هچنین کارآیی روش LSSVR را با روش‌های شبکه عصبی تابع پایه شعاعی[6] (RBF) و پس‌انتشار[7] (BP) مقایسه کردند. مقایسه نتایج، برتری روش LSSVR را نشان داد. لیو و همکاران (۱۰) با به‌کارگیری روش برنامه‌ریزی ژنتیکی[8] (GP) و داده‌های حقیقی-رگرسیون بردار پشتیبان[9] (RGA-SVR) به حل مسئله پیش‌پیش‌بینی پارامترهای کیفی آب برای پرورش ماهی پرداختند. آنها از الگوریتم GA به منظور تنظیم ضرایب روش SVR استفاده کردند. مقایسه نتایج مجذور میانگین مربعات خطا[10] (*RMSE*) حاصل از روش‌های SVR، BP و الگوریتم RGA-SVR، برتری الگوریتم RGA-SVR را نشان داد.

تاکنون تحقیقات متنوعی در نحوه انتخاب متغیرهای ورودی به الگوهای داده‌مبنا و تعیین ساختار این متغیرها انجام شده است. یون و همکاران (۲۶) از همبستگی متقاطع[11] برای تعیین ساختار متغیرهای ورودی به الگو داده‌مبنا ماشین بردار پشتیبان[12] (SVM) استفاده کردند و سطح آب زیرمینی را الگوبندی کردند. نوری و همکاران (۱۲) از روش تجزیه مولفه اصلی[13] (PCA)، برای انتخاب ساختار متغیرهای ورودی که بیش‌ترین تاثیر را بر خروجی الگو داده‌مبنا شبکه عصبی مصنوعی[14] (ANN) دارند، استفاده کردند و در نتیجه غلظت روزانه پارامترکربن مونوکسید در هوا را داده‌کاوی کردند. نوری و همکاران (۱۳) ورودی‌های الگو داده‌مبنا ANN با آزمون گاما[15] را تعیین کردند و پارامتر ماده زاید جامد[16] را الگوبندی کردند. فلاح مهدی‌پور و همکاران (۶) از الگوهای پیش‌فرض برای انتخاب ساختار ورودی متغیرها به GP استفاده کردند و سطح آب زیرزمینی را الگوبندی کردند. قویدل و منتصری (۷) از رگرسیون گام به گام برای تعیین ساختار متغیرهای ورودی به الگو داده‌مبنا شبکه عصبی استفاده کردند و پارامتر کیفی TDS را الگوبندی کردند.

مرور منابع نشان می‌دهد که روش‌های داده‌مبنا در بسیاری از زمینه‌های هیدرولوژی و مدیریت منابع آب به کار گرفته شده‌اند (۱۶). روش LSSVR روش نسبتا جدیدی در میان روش‌های داده‌مبنا است که قابلیت آن در پیش‌بینی و الگوبندی در رشته‌های مختلف اثبات شده است. اما بزرگ‌ترین نقطه ضعف این روش حساسیت بالای آن به ضرایب موازنه بین خطا و حاشیه[17] (γ) و ضریب تابع کرنل[18] (σ) است، این تحقیق از GA که یک الگوریتم تکاملی بر اساس نظریه داروین است و در حل مسائل بهینه‌بندی بسیار کارا می‌باشد، به منظور تنظیم خودکار و بهینه ضرایب این روش استفاده شد. به طوری که یک الگوریتم تلفیقی متشکل از روش LSSVR و GA به‌دست آمد. در این تحقیق از الگوریتم GA-LSSVR برای الگوبندی پارامترهای مختلف کیفی رودخانه استفاده می‌شود. اغلب تحقیقات بر روی الگوبندی پارامترهای BOD، COD و DO انجام گرفته است. ولی در تحقیق حاضر به منظور کاهش هزینه نمونه‌برداری سطحی، اندازه‌گیری پارامترهای کیفی در محیط آزمایشگاه، خطاهای انسانی و غیره به الگوبندی پارامترهای Na^{+}، K^{+}، Mg^{2+}، So_4^{2-}، Cl^{-}، pH، EC[19] و TDS[20] با بیش‌ترین قابلیت تقریب در رودخانه سفید رود واقع در رشت پرداخته می‌شود. همچنین نتایج حاصل از به کارگیری دو روش پیش‌پردازش داده ضریب همبستگی و PCA در انتخاب متغیرهای ورودی به الگوریتم GA-LSSVR مقایسه شده است.

مواد و روش‌ها

روش LSSVR

روش LSSVR توسط سویکنز و همکاران (۲۵) توسعه داده شد که اصلاحاتی نسبت به روش اصلی توسط وپنیک[21] (۲۴) در آن لحاظ

1- Particle Swarm Optimization
2- Multilayer Perceptron
3- Support Vector Clustering
4- Lucknow
5- Biochemical Oxygen Demand
6- Radial Basis Function
7- Back Propagation
8- Genetic Programming
9- Real-Value Genetic Algorithm SVR
10- Root Mean Square Error
11- Cross Correlation
12- Support Vector Machin
13- Principle Component Analysis
14- Artificial Neural Network
15- Gamma Test
16- Solid Waste
17- Tradeoff parameter between error and margine
18- Width of the Gaussian basis function
19- Electric Conductivity
20-Total Dissolved Solid
21- Vapnik

شده است. در ادامه روش LSSVR شرح داده می‌شود.

در این روش فرض می‌شود که ارتباط بین داده‌های ورودی و خروجی غیرخطی است ولی با یک رابطه نگاشت[1] غیرخطی می‌توان فضایی به نام فضای ویژگی[2] ایجاد کرد که در آن فضا ارتباط بین داده‌های ورودی و خروجی به صورت خطی باشد. رابطه خطی بین ورودی‌ها و خروجی‌ها در فضای ویژگی توسط رابطه (۱) قابل بیان است.

$$y^O(x)=\langle \boldsymbol{\omega},\boldsymbol{\varphi}(\boldsymbol{x})\rangle + b \qquad (۱)$$

که در آن، $\boldsymbol{x}$ و $y^O(x)$ = به ترتیب ورودی و خروجی داده‌های مشاهداتی دسته آموزش؛ $\boldsymbol{\omega}$ = بردار وزن؛ $\boldsymbol{\varphi}(\boldsymbol{x})$ = یک تابع غیر خطی که داده‌ها را از فضای اصلی به فضای ویژگی نگاشت می‌کند. به طوری که $\langle .,. \rangle$ مشخص کننده ضرب داخلی دو بردار در فضای هیلبرت[3] (فضایی که به نام فضای داخلی یا نقطه‌ای نیز خوانده می‌شود و در آن ضرب داخلی دو بردار عددی حقیقی می‌باشد) و b = بایاس[4] می‌باشند. مقادیر b و $\boldsymbol{\omega}$ را می‌توان با حل مسئله بهینه‌بندی زیر به‌دست آورد.

$$Min\ J(\boldsymbol{\omega},e)=\frac{1}{2}\left\|\boldsymbol{\omega}^2\right\|+\frac{1}{2}\gamma\sum_{t=1}^{T}e_t^2$$

$$Subject\ \ to \qquad (۲)$$

$$y^O{}_t=\langle \omega.\phi(x_t)\rangle + b + e_t \quad t=1,2,...,T$$

که در آن، $J(\omega,e)$ = تابع خسارت[5] که در واقع خطای آموزش را محاسبه می‌کند؛ $\left\|\boldsymbol{\omega}^2\right\|$ = نرم اقلیدسی[6]؛ e_t = متغیر خطای اتفاق افتاده برای ورودی t ام که $(t = 1, ..,T)$ ؛ γ = یک ثابت قابل تنظیم بوده که به نوعی مقدار $\left\|\boldsymbol{\omega}^2\right\|$ را با توجه به پیچیدگی تابع خطا تعیین می‌کند؛ t = شمارنده تعداد داده‌های ورودی که از ۱ تا T متغیر است و T = تعداد کل داده‌های ورودی هستند. ایده کلیدی در حل مسئله بهینه‌سازی (۲) ساختن شکل لاگرانژی تابع هدف مبنا و قیود مربوط، با دسته متغیرهای دوگان می‌باشد. می‌توان نشان داد که این شکل از تابع هدف دارای نقاط ایستا[7] نسبت به متغیرهای اصلی و دوگان است. لاگرانژین تابع به صورت زیر تعریف می‌شود.

$$L(\boldsymbol{\omega},b,e_t,l_t)=J(\boldsymbol{\omega},e_t)-\sum_{t=1}^{T}l_t\left(\langle \boldsymbol{\omega},\boldsymbol{\varphi}(\boldsymbol{x}_t)\rangle + b + e_t - y^O{}_t\right) \qquad (۳)$$

$$l_t \geq 0$$

که در آن $L(\omega, b, e_t, l_t)$ = تابع لاگرانژین و l_t = ضریب لاگرانژ برای ورودی t ام هستند. با توجه به نقاط ایستا می‌بایست، مشتقات جزئی L نسبت به متغیرهای اصل تابع $\boldsymbol{\omega}$، b، e_t و l_t برابر صفر قرار گیرند. در نتیجه با جایگزینی روابط (۴) در رابطه (۳)، مقدار بیشینه L محاسبه می‌شود.

$$\frac{\partial L}{\partial \omega}=0\rightarrow \boldsymbol{\omega}=\sum_{t=1}^{T}l_t\,\boldsymbol{\varphi}(\boldsymbol{x}_t)$$

$$\frac{\partial L}{\partial b}=0\rightarrow \sum_{t=1}^{T}l_t=0$$

$$\frac{\partial L}{\partial e_t}=0\rightarrow l_t=\gamma\, e_t \qquad (۴)$$

$$\frac{\partial L}{\partial l_t}=0\rightarrow \langle \boldsymbol{\omega},\boldsymbol{\varphi}(\boldsymbol{x}_t)\rangle + b + e_t - y^O{}_t=0$$

$$\forall t=1,2,...,T$$

اطلاعات کمی در مورد انتخاب تابع غیر خطی مناسب $\boldsymbol{\varphi}(\boldsymbol{x})$ در دسترس می‌باشد. در نتیجه می‌توان یک تابع کرنل در نظر گرفت که بیان‌کننده فضای ویژگی است که از فضای اصلی، نگاشت غیرخطی شده است.

$$(x_t,x_k)=\langle \boldsymbol{\varphi}(\boldsymbol{x}_t)',\boldsymbol{\varphi}(\boldsymbol{x}_k)\rangle \quad t=1,2,...,T \quad k=1,2,...,T \qquad (۵)$$

که در آن، $K(x_t, x_k)$ = تابع کرنل؛ k = شمارنده تعداد متغیرهای ورودی می‌باشد.

در نتیجه مقدار تخمین زده شده (محاسبه‌ای) برای خروجی y^O توسط روش LSSVR از رابطه (۶) به‌دست می‌آید.

$$y^E=\sum_{t=1}^{T}l_t\,K(x_t,x)+b \qquad (۶)$$

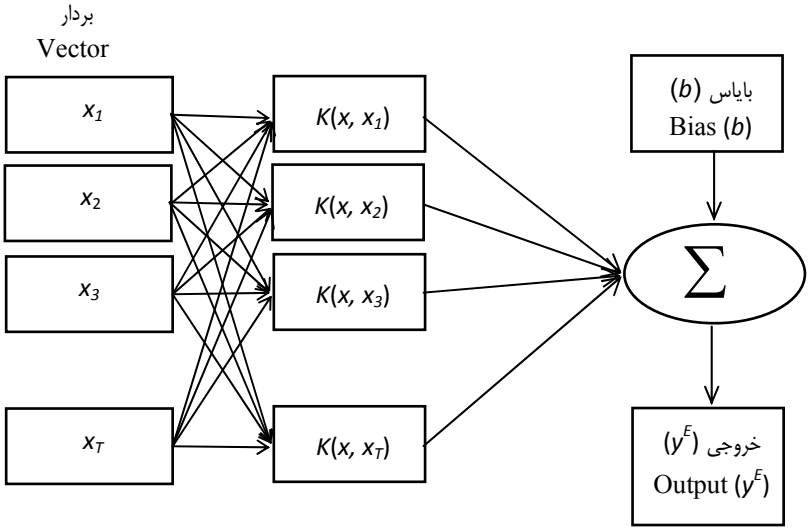
که در آن، y^E = مقدار تخمین زده شده (محاسباتی) توسط روش LSSVR برای مقدار y^O است که تفاوت بین مقادیر y^E و y^O توسط معیارهای آماری در ادامه تحلیل می‌گردد. تابع کرنل می‌تواند از بین توابع خطی[8]، چند جمله‌ای[9]، سیگموید[10]، RBF و MLP انتخاب شود. در این تحقیق از تابع کرنل RBF استفاده شده است. رابطه تابع کرنل RBF به شرح رابطه (۷) است.

1- Mapping
2- Feature Space
3- Hibert
4- Bias
5- Loss Function
6- Euclidean norm
7- Stationary Points

8- Linear
9- Polynomial
10- Sigmoid

$$K(x, x_t) = e^{\frac{-\|x - x_t\|^2}{\sigma^2}} \quad (۷)$$

که در آن، x_t = داده t ام بردار ورودی و σ = ضریب تابع کرنلRBF که پارامتری ثابت و قابل تنظیم توسط کاربر می‌باشد، هستند. معماری روش SVR به شرح شکل ۱ است.



شکل ۱- ساختار روش LSSVR
Figure 1- LSSVR method constructure

روش LSSVR راهکاری برای انتخاب ضرایب γ و σ پیشنهاد نکرده است. به دلیل این‌که دقت روش LSSVR، به انتخاب مقادیر این ضرایب وابسته است، در این تحقیق از GA برای به‌دست آوردن مقادیر بهینه این ضرایب استفاده می‌شود.

انتخاب بهینه ضرایب روش LSSVR با GA

الگوریتم‌های بهینه‌بندی فراکاوشی قابلیت یافتن جوابی در همسایگی جواب بهینه کلی مسئله را دارند و در حل بسیاری از مسائل بهینه‌بندی مرتبط با منابع آب استفاده شده‌اند (۱، ۲، ۳ و ۲۰). GA یک روش جست‌جوی ابتکاری الهام گرفته شده از تکامل جانداران در طبیعت است که برای حل مسائل بهینه‌بندی استفاده می‌شود (۵). این الگوریتم با الهام از تکامل[1] و وراثت[2] موجودات زنده توسعه داده شده است. در این الگوریتم، مسئله به صورت جامعه‌ای از افراد تعریف می‌شود. مطابق قانون بقا، بهترین افراد به دلیل توانایی سازگاری بیش‌تر با محیط برای تولید مثل انتخاب می‌شوند. افراد انتخاب شده طی فرآیندهای تزویج[3] و جهش[4] نسل جدیدی تولید می‌کنند. به این صورت افراد از نسلی به نسل دیگر تکامل می‌یابند. فرآیند تکامل منجر به ایجاد افرادی می‌شود که بیش‌ترین تطابق با محیط را دارند یا به عبارت دیگر راه حل بهینه مسئله هستند. در نتیجه در این تحقیق الگوریتم GA-LSSVR توسعه داده شد که توانایی تنظیم خودکار و بهینه ضرایب روش LSSVR را با GA دارد. روندنمای الگوریتم GA-LSSVR در شکل ۲ ارائه می‌شود.

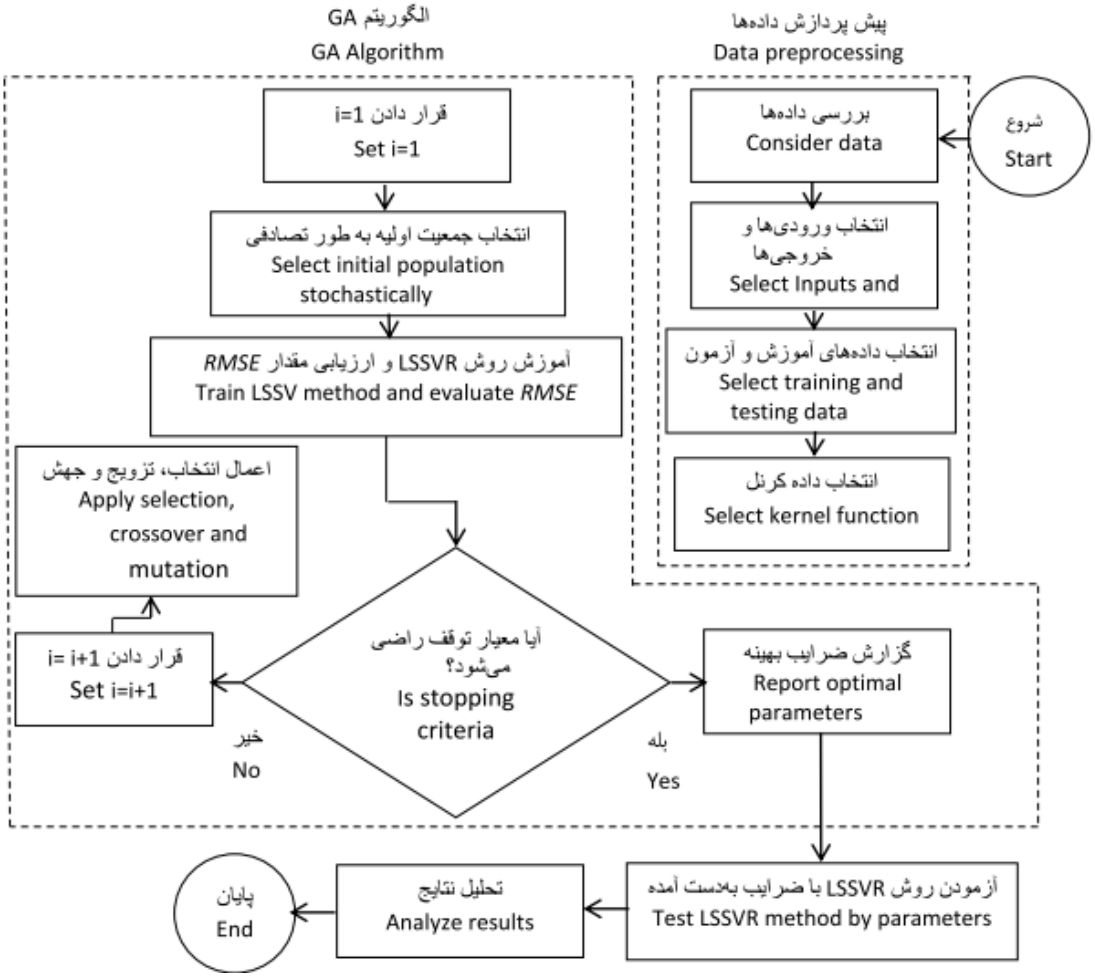
تابع هدف مورد استفاده در رابطه (۸) ارائه شده است. همچنین برای مقایسه، معیارهای ضریب تشخیص[5] (R^2) و نش-سایتکلیف[6] (NS) نیز استفاده می‌شوند. رابطه این معیارها در روابط (۹) و (۱۰) ذکر شد.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t^O - y_t^E)^2} \quad (۸)$$

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(y_t^O - \bar{y}_t^O)^2(y_t^E - \bar{y}_t^E)^2}{\sum_{t=1}^{T}(y_t^O - \bar{y}_t^O)^2 \times \sum_{t=1}^{T}(y_t^E - \bar{y}_t^E)^2} \quad (۹)$$

1- Evolution
2- Heredity
3- Crossover
4- Mutation

5- Coefficient of Determination
6- Nash-Sutcliff



شکل ۲- روندنمای الگوریتم GA-LSSVR

Figure 2- GA-LSSVR Flowchart

$$NS = 1 - \frac{\sum_{t=1}^{T}\left(y_t^O - y_t^E\right)^2}{\sum_{t=1}^{T}\left(y_t^O - \bar{y}_t^O\right)^2} \qquad (۱۰)$$

که در آن، y_t^O و y_t^E = به ترتیب داده‌های مشاهداتی و محاسباتی و $\bar{y}_t^O$ و $\bar{y}_t^E$ به ترتیب = متوسط داده‌های مشاهداتی و محاسباتی می‌باشند. معیارهای ارائه شده در روابط (۸) تا (۱۰) توسط وانگ و همکاران (۲۵)؛ رجایی و همکاران (۱۷)؛ سلطانی و همکاران (۱۹) و عروجی و همکاران (۱۴) نیز به کار گرفته شده‌اند.

پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها شامل سه مرحله‌ی ۱- بررسی داده‌ها به منظور کامل بودن سری زمانی و منطقی بودن مقدار متغیرها در سری آماری، ۲- شناسایی متغیرهای ورودی موثر بر متغیر خروجی مورد نظر با روش‌های ضریب همبستگی و PCA و ۳- انتخاب داده‌های دسته آموزش و آزمون بر اساس طول دوره آماری است.

روش ضریب همبستگی

ضریب همبستگی، یکی از معیارهای مورد استفاده در تعیین همبستگی دو متغیر است. رابطه ضریب همبستگی در زیر آورده شده است.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]}{\sigma_X \sigma_Y} \qquad (۱۱)$$

که در آن، cov = کوواریانس بین متغیر کمی X با Y؛ σ_X و σ_Y = به ترتیب انحراف معیار متغیر X و Y؛ μ_X و μ_Y = به ترتیب میانگین متغیر X و Y و E = امید ریاضی را نشان می‌دهد.

در این تحقیق برای انتخاب ورودی‌های هر الگو از ضریب همبستگی پیرسون بین داده‌ها استفاده شد و متغیرهایی به عنوان ورودی به الگو معرفی شدند که همبستگی معناداری در سطح ۹۹ درصد با متغیر مورد نظر داشتند.

روش PCA

در این تحقیق از روش PCA نیز به منظور مقایسه نتایج استفاده شد. یکی از کاربردهای مهم روش PCA، در رگرسیون است. با روش PCA می‌توان تعداد زیادی متغیر توضیحی (مستقل[1]) همبسته را با تعداد محدودی متغیر توضیحی جدید که مولفه‌های اصلی نامیده می‌شوند و ناهمبسته‌اند، جایگزین نمود (۴).

با این روش، ترکیباتی از p متغیر اولیه $x_1, x_2, \ldots, x_p$ برای ایجاد حداکثر p مولفه مستقل به صورت PC_1، PC_2، ...، PC_p ایجاد می‌شود. هر مولفه اصلی می‌تواند با دنباله ارائه شده در رابطه ۱۲ نشان داده شود (۸).

$$\begin{aligned} PC_1 &= w_{11}x_1 + w_{12}x_2 + \ldots + w_{1p}x_p \\ PC_2 &= w_{21}x_1 + w_{22}x_2 + \ldots + w_{2p}x_p \\ &\vdots \\ PC_p &= w_{p1}x_1 + w_{p2}x_2 + \ldots + w_{pp}x_p \end{aligned} \qquad (۱۲)$$

که در این رابطه، PC_i = مولفه i ام؛ w_{ij} = ضریب مربوط به مولفه i ام و متغیر اولیه jام؛ و x_i = متغیر اولیه i ام است. ضریب w_{ij} طوری تخمین زده می‌شود که اولین مولفه (PC_1) بیشینه واریانس داده‌ها را در نظر گرفته و دومین مولفه (PC_2)، بیشینه واریانس در نظر گرفته نشده توسط اولین مولفه را پیش‌بینی کرده و این روند ادامه می‌یابد تا آخرین مولفه (PC_P) نیز تمامی واریانس مورد نظر را در بر بگیرد.

$$w_{i1}^2 + w_{i2}^2 + \ldots + w_{ip}^2 = 1 \qquad \forall i = 1,\ldots,p \qquad (۱۳)$$

$$w_{i1}w_{j1} + w_{i2}w_{j2} + \ldots + w_{ip}w_{jp} = 0 \qquad \forall i \neq j \qquad (۱۴)$$

اطلاعات بیش‌تر در مورد روش PCA و نحوه محاسبه ضرایب w_{ij} را می‌توان در تاباچنیک و فیدل (۲۲)؛ اویانگ (۱۵) و نوری و همکاران (۱۳) مطالعه کرد.

1- Independent

به منظور حذف تاثیر مقیاس مقادیر متغیرهای ورودی مختلف نسبت به یکدیگر، مقادیر متغیرهای ورودی برای استفاده از روش PCA بهتر است استاندارد شوند. در این تحقیق از رابطه‌ی (۱۵) به منظور استانداردبندی داده‌ها استفاده شده است.

$$z = \frac{x_i - \mu}{\sigma} \qquad (۱۵)$$

که در آن x_i = متغیر ورودی i ام؛ μ = میانگین متغیرهای ورودی و σ = انحراف معیار متغیرهای ورودی هستند.

منطقه مورد مطالعه و داده‌ها

این تحقیق به تحلیل داده‌های کیفیت آب ایستگاه آستانه در مسیر رودخانه سفیدرود واقع در شمال ایران برای الگوبندی متغیرهای کیفی آب می‌پردازد. طول و مساحت حوضه رودخانه سفیدرود به ترتیب تقریباً ۶۷۰ کیلومتر و ۱۳٬۴۵۰ کیلومتر مربع هستند. این رودخانه در شهر رشت به دریای خزر می‌ریزد. شکل ۶ موقعیت منطقه مطالعاتی ایستگاه آستانه و رودخانه سفیدرود را نشان می‌دهد. متغیرهای مورد نظر در این تحقیق شامل Na^+، K^+، Mg^{2+}، So_4^{2-}، Cl^-، pH، EC و TDS هستند. سری زمانی پارامترهای کیفی مذکور به صورت ماهانه طی سال‌های ۱۳۶۴ تا ۱۳۷۷ و سال‌های ۱۳۷۹ تا ۱۳۸۴ موجود است.

نتایج

انتخاب و بررسی داده‌های آموزش و آزمون

در این تحقیق داده‌های دسته آموزش و آزمون به نسبت ۷۰ به ۳۰ درصد از کل داده‌ها به صورت متوالی انتخاب شدند. به‌این صورت، داده‌های دسته آموزش دوره آماری ۱۳۶۴ تا ۱۳۷۷ (۱۵۴ ماه) را در برمی‌گیرند و دوره آماری ۱۳۷۹ تا ۱۳۸۴ (۶۹ ماه) برای داده‌های دسته آزمون انتخاب شدند. شایان ذکر است که سری زمانی انتخاب شده برای داده‌های دسته آموزش و آزمون یک سری زمانی کامل بدون مقادیر گم شده است. در این بخش ابتدا خصوصیات آماری داده‌های دسته آموزش و آزمون در جدول ۱ بررسی می‌شود.

همان‌طور که در جدول ۱ قابل مشاهده است. پارامتر کیفی K و Na دارای بیش‌ترین ضریب تغییرات به ترتیب در بخش داده‌های آموزش و آزمون و پارامتر کیفی pH دارای کمترین ضریب تغییرات در بخش داده‌های آموزش و آزمون است.

انتخاب ورودی‌ها و تعیین ساختار الگو

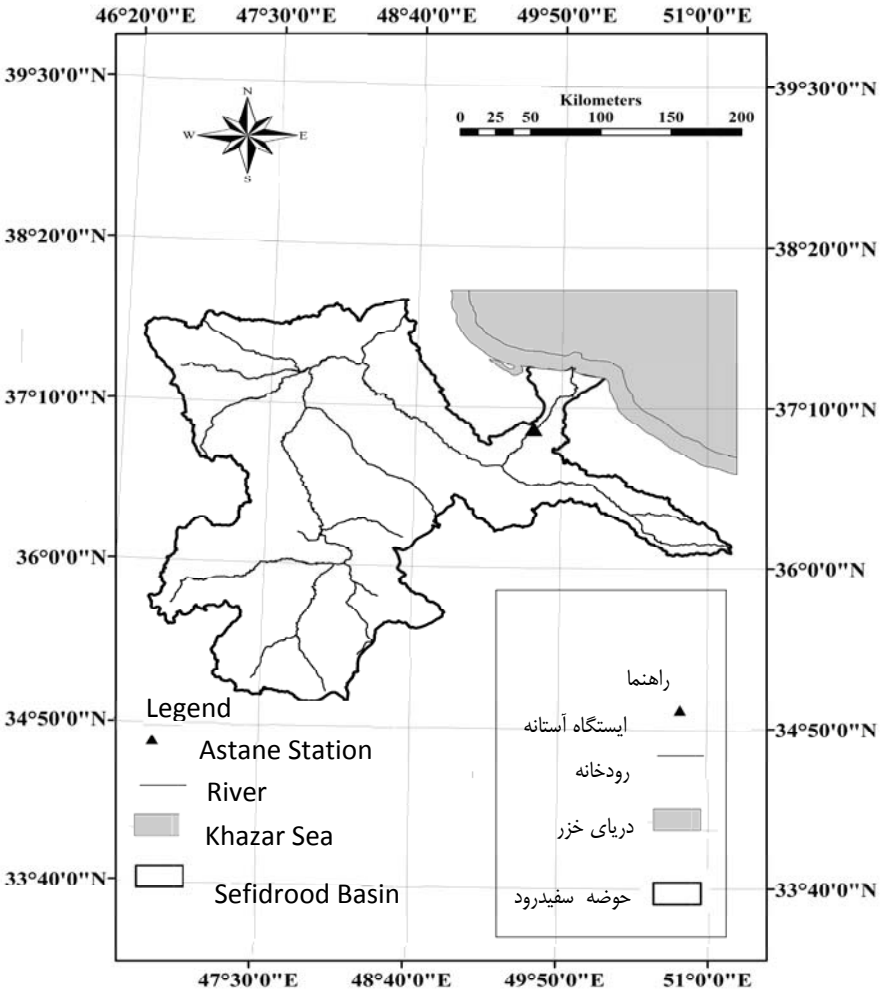
در این قسمت ضریب همبستگی پیرسون برای هر دو متغیر کیفی و برای هر یک از متغیرهای کیفی و دبی ورودی به مخزن در جدول ۲ محاسبه شده است. این ضرایب در طول کل دوره آماری بررسی

شده است.

که در آن Q = دبی ورودی ماهانه به مخزن بر حسب مترمکعب بر ثانیه می‌باشد. همان طور که در جدول ۲ نشان داده می شود، pH با هیچ یک از متغیرهای کیفی و دبی ورودی رابطه معناداری ندارد. در نتیجه برای pH الگو ورودی دیگری در نظر گرفته شد که در جدول ۳ آورده شده است. بر اساس جدول ۲ داده‌هایی که در سطح ۹۹ درصد با متغیر مورد نظر همبستگی و رابطه معناداری دارند، به عنوان ورودی برای تخمین پارامتر کیفی مورد نظر انتخاب شده و الگوهای ورودی در جدول ۳ برای تخمین هر پارمتر به طور جداگانه تشکیل داده شدند. همانطور که در جدول ۲ مشاهده می‌شود هر پارامتر کیفی با پارامترهای کیفی مختلفی رابطه معنا دارد، لذا نمی‌توان برای الگوبندی هر پارامتر کیفی از یک الگو ورودی یگانه استفاده کرد.

که در آن، Q_{t-1} و Q_{t-2} = به ترتیب برابر دبی جریان ورودی با یک ماه و دو ماه تاخیر و pH_{t-1} و pH_{t-2} = به ترتیب برابر با pH با یک ماه و دو ماه تاخیر می‌باشند. همان‌طور که گفته شد به دلیل این‌که هیچ یک از پارامترهای کیفی رابطه‌ی معناداری با پارامتر کیفی pH ندارند. لذا از پارامترهای Q_{t-1}، Q_{t-2}، pH_{t-1} و pH_{t-2} که به نسبت رابطه‌ی معنادارتری با پارامتر pH دارند برای تعین ساختار الگو ورودی این پارامتر استفاده شد.

در ادامه به منظور حذف تاثیر مقیاس مقادیر متغیرهای ورودی مختلف نسبت به یکدیگر، مقادیر متغیرهای ورودی برای استفاده از روش PCA با رابطه (۱۴) استاندارد شدند. سپس متغیرهای اصلی ماتریس ورودی به‌دست آمدند. شکل ۴ درصد تجمعی واریانس هر مولفه‌ی اصلی را نشان می‌دهد.



شکل ۳- موقعیت حوضه آبریز رودخانه سفیدرود و ایستگاه آستانه

Figre 3- The Sefidrood basin and Astane Station situation

جدول ۱- خصوصیات آماری داده‌های دسته آموزش و آزمون متغیرهای کیفیت آب رودخانه سفیدرود در ایستگاه آستانه

Table 1- The statistical properties of the train and test set of quality parameters in the Sefidrood statation

متغیر Parameter	مجموعه داده Data set	کمینه Minimum	میانگین Average	بیشینه Maximum	انحراف معیار Standard deviation	ضریب تغییرات Coefficient of variation
Na^+ (meq.l)	آموزش Training	0.05	5.78	15.65	2.64	45.72
	آزمون Testing	0.19	3.74	10.13	2.19	58.57
K^+ (meq.l)	آموزش Training	0.01	0.09	0.22	0.04	49.14
	آزمون Testing	0.01	0.08	0.15	0.03	39.76
Mg^{2+} (meq.l)	آموزش Training	0.20	2.31	5.50	0.93	40.37
	آزمون Testing	0.38	2.08	5.52	1.04	49.77
So_4^{2-} (meq.l)	آموزش Training	0.34	2.83	7.38	1.27	44.88
	آزمون Testing	0.21	1.93	4.12	1.00	51.89
Cl^- (meq.l)	آموزش Training	0.80	5.72	15.60	2.65	46.37
	آزمون Testing	0.20	4.27	10.90	2.45	57.31
pH	آموزش Training	6.70	7.74	8.80	0.37	4.81
	آزمون Testing	6.42	7.54	8.42	0.46	6.07
EC (µs.cm)	آموزش Training	244.43	1221.70	2336.00	381.49	31.23
	آزمون Testing	252.00	1024.41	2018.00	393.96	38.46
TDS (mg.l)	آموزش Training	263.00	772.59	1472.00	232.36	30.08
	آزمون Testing	159.00	641.75	1271.00	243.81	37.99

جدول۲- ضریب همبستگی بین متغیرهای کیفیت آب رودخانه سفیدرود

Table 2- The correlation coefficient between quality parameters in the Sefidrood river

TDS	EC	pH	Cl^-	So_4^{2-}	Mg^{2+}	K^+	Na^+	Q	متغیر Parameter
0.13	0.12	0.07	-0.11	0.03	-0.17**	-0.01	-0.06	1.00	Q
0.87**	0.87**	0.01	0.87**	0.68**	0.54**	0.32**	1.00	0.06	Na^+
0.37**	0.36**	0.09	0.39**	0.43**	0.10	1.00	0.32**	-0.01	K^+
0.68**	0.69**	0.04	0.66**	0.48**	1.00	0.10	0.54**	-0.17**	Mg^{2+}
0.75**	0.77**	0.07	0.63**	1.00	0.48**	0.43**	0.68**	0.03	So_4^{2-}
0.92**	0.94**	0.03	1.00	0.63**	0.66**	0.31**	0.87**	0.11	Cl^-
0.02	0.00	1.00	0.03	0.70	0.04	0.09	-0.01	-0.08	pH
0.97**	1.00	0.00	0.94**	0.77**	0.69**	0.36**	0.87**	-0.12	EC
1.00	0.98**	0.02	0.92**	0.75**	0.68 **	0.36**	0.87**	-0.13	TDS

** محدوده اطمینان ۹۹ ٪

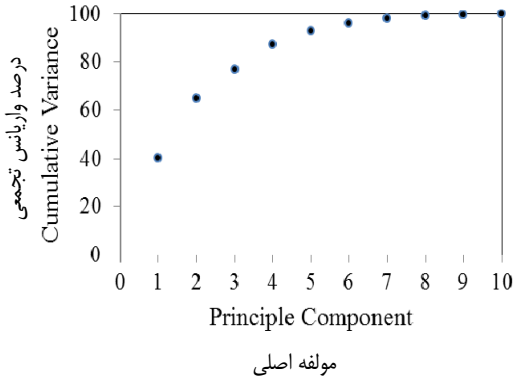
Confidence interval %99

در این تحقیق مطابق شکل ۴، ۱۰ متغیر اصلی محاسبه شده ۹۹/۹۹ درصد واریانس ماتریس ورودی را توجیه می‌کنند و ۲۵ متغیر بعدی تنها توجیه کننده ۰۱.۰ درصد از واریانس ماتریس ورودی هستند. لذا به دلیل افزایش دقت، ۱۰ مولفه اصلی اولیه به عنوان ورودی به الگوریتم GA-LSSVR داده شد.

جدول۳- الگو ورودی به GA-LSSVR بر اساس ضرایب همبستگی

Table 3- input model to GA-LSSVR based on correlation coefficeint

متغیر Parameter	الگو ورودی Input model											
	Q	Q_{t-1}	Q_{t-2}	Na^+	K^+	Mg^{2+}	So_4^{2-}	Cl^-	pH_{t-1}	pH_{t-2}	EC	TDS
Na^+				√	√	√	√	√			√	√
K^+				√	√	√	√				√	√
Mg^{2+}	√			√	√	√	√				√	√
So_4^{2-}				√				√			√	√
Cl^-					√	√	√	√			√	√
pH	√	√	√	√					√	√		
EC				√	√	√	√	√				√
TDS					√	√	√	√			√	



شکل ٤- درصد تجمعی واریانس به ازای مولفه‌های اصلی

Figure 7- Cumulative percentage of variance for main parameter

جدول٤- مقادیر بهینه ضرایب γ و σ به‌دست آمده از الگوریتم GA

Table 4- The obtained γ and σ of optimal value of coefficient from GA

متغیر	روش ضریب همبستگی		روش PCA	
Parameter	σ	γ	σ	γ
Na^+(meq.L)	43.80	30.15	58.31	48.75
K^+(meq.L)	22.60	58.07	3.03	13.43
Mg^{2+}(meq.L)	64.74	85.74	18.88	69.87
So_4^{2-}(meq.L)	96.87	38.00	15.61	57.61
Cl^-(meq.L)	68.28	8.70	15.77	58.04
pH	56.49	4.27	38.92	0.58
EC(µs.cm)	96.08	52.10	21.09	75.98
TDS(meq.L)	18.10	46.95	17.17	66.75

نتایج الگوریتم GA-LSSVR با دو روش پردازش داده ضرایب همبستگی و PCA

جدول ۴ ضرایب بهینه به‌دست آمده از الگوریتم GA-LSSVR را نشان می‌دهد. در این تحقیق GA با ۲۰ جمعیت، یک نخبه‌گرایی، نرخ تزویج[1] ۸۰.، نرخ جهش[2] ۰۲.۰، تعداد ۱۰۰ تکرار به‌دست آوردن ضرایب بهینه روش LSSVR استفاده شد. کران پایین و بالا بین ۰ تا

1- Crossover
2- Mutation

۱۰۰ برای تعیین محدوده متغیر تصمیم در GA تعریف شد. ضرایب بهینه ارائه شده در جدول ۴ نشان می‌دهند که ضرایب بهینه به‌دست آمده در روش ضریب همبستگی در بازه ۱۸ تا ۹۶، در روش PCA در بازه ۳ تا ۵۸ برای ضریب σ و در روش ضریب همبستگی در بازه ۴ تا ۸۵، در روش PCA در بازه ۵.۰ تا ۷۵ برای ضریب γ متغیر است که بازه وسیعی را در بر می‌گیرد. این بازه وسیع برای ضرایب روش LSSVR بیان‌گر این موضوع است که سری‌های زمانی متغیرهای کیفی آب مورد مطالعه دارای الگوهای مختلفی هستند و طیف وسیعی از سری‌های زمانی را دربر می‌گیرند.

با ضرایب بهینه ارائه شده در جدول ۴ روش LSSVR برای هر دو روش ضریب همبستگی و PCA به طور جداگانه اجرا شد. نتایج به‌دست آمده از الگوریتم GA-LSSVR برای داده‌های دسته آموزش و آزمون متغیرهای مختلف کیفیت آب در جدول ۵ ارائه می‌شوند.

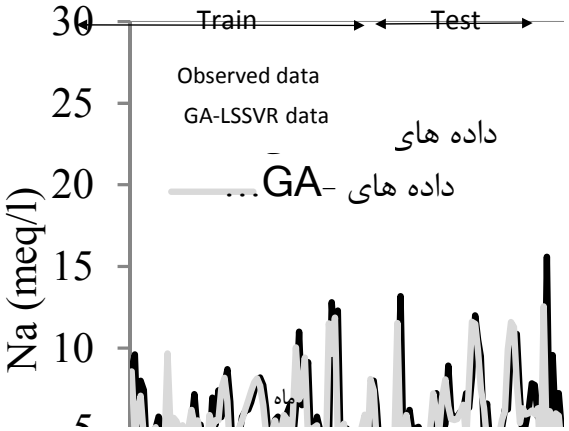
جدول ۵- آماره‌های به‌دست آمده از الگوریتم GA-LSSVR با دو روش ضریب همبستگی و PCA برای دسته‌های آموزش و آزمون

Table 5- The obtained statistics of GA-LSSVR using two methods of correlation coefficient and PCA for the train and test data

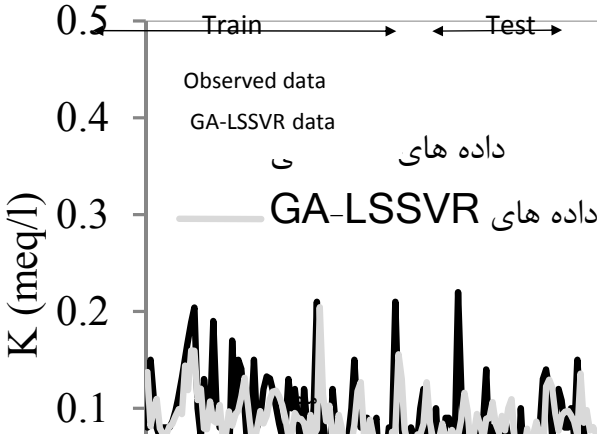
متغیر Parameter	روش Method	آموزش Train			آزمون Test		
		RMSE	R^2	*NS*	*RMSE*	R^2	*NS*
Na^+(meq.L)	ضریب همبستگی	1.10	0.91	0.82	0.88	0.94	0.85
	PCA	0.92	0.93	0.87	1.20	0.94	0.69
K^+(meq.L)	ضریب همبستگی	0.03	0.66	0.43	0.02	0.64	0.38
	PCA	0.03	0.63	0.36	0.02	0.64	0.38
Mg^{2+}(meq.L)	ضریب همبستگی	0.61	0.75	0.56	0.65	0.79	0.60
	PCA	0.64	0.72	0.52	0.66	0.79	0.58
So_4^{2-}(meq.L)	ضریب همبستگی	0.70	0.83	0.70	0.58	0.92	0.65
	PCA	0.76	0.79	0.63	0.67	0.84	0.54
Cl^-(meq.L)	ضریب همبستگی	0.91	0.94	0.88	0.61	0.97	0.94
	PCA	0.86	0.94	0.89	0.71	0.96	0.91
pH	ضریب همبستگی	0.32	0.55	0.30	0.38	0.69	0.30
	PCA	0.02	0.99	0.99	0.31	0.75	0.52
EC (µs.cm)	ضریب همبستگی	67.17	0.98	0.97	73.34	0.98	0.97
	PCA	69.08	0.98	0.98	61.79	0.98	0.97
TDS (meq.L)	ضریب همبستگی	40.50	0.98	0.97	65.45	0.98	0.93
	PCA	39.35	0.98	0.97	55.84	0.98	0.94

با توجه به جدول ۵، الگوریتم GA-LSSVR با به‌کارگیری روش ضریب همبستگی و PCA نتایج مشابهی را ارائه می‌دهد. در مورد متغیرهای کیفی pH، EC و TDS، هرچند که روش PCA از دقت بیش‌تری برخوردار است، ولی اختلاف بین *RMSE* روش ضریب همبستگی و روش PCA بسیار ناچیز است. به طوری که روش PCA نسبت به روش ضریب همبستگی منجر به بهبود مقدار *NS* به میزان ۲۲ و ۱/۰ درصد به ترتیب برای متغیرهای کیفی pH و TDS شده است و مقدار آن برای متغیر EC تغییر نکرده است. از آنجایی که مجموع معیار *NS* برای داده‌های آزمون در روش ضریب همبستگی ۶۲/۵ و در روش PCA برابر ۵۳/۵ است. لذا، نتایج بدست آمده از روش ضریب همبستگی و PCA به هم نزدیک است؛ هرچند که از نظر سادگی روش ضریب همبستگی ساده‌تر است ولی از نظر دقت بین این دو نمی‌توان اختلاف معنی‌داری قائل شد. در مجموع با توجه به مثبت بودن مقادیر *NS* می‌توان نتیجه گرفت که الگوریتم GA-LSSVR توانایی بالایی در الگوبندی متغیرهای کیفی با به‌کارگیری هر یک از روش‌های پیش‌پردازش ضریب همبستگی و PCA را دارد. نتایج الگوبندی متغیرهای کیفیت آب به‌دست آمده از الگوریتم GA-LSSVR در شکل ۸ ارائه می‌شوند.

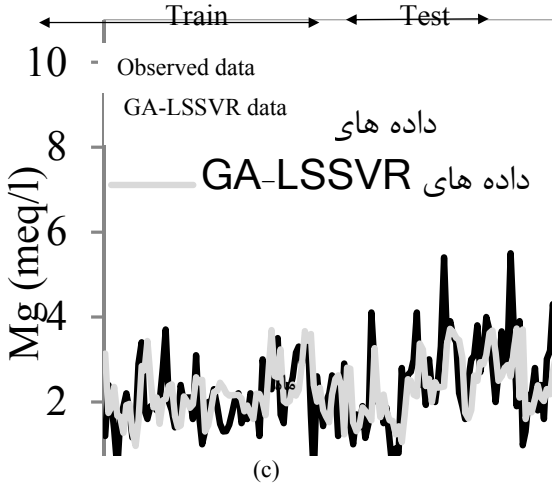
عروجی و همکاران (۱۴) با الگوهای ورودی پیش‌فرض (تعیین الگو ورودی به الگو داده‌مبنا به صورت فرضی و با توجه به تجربه)، متغیرهای کیفی Na^+، K^+، Mg^{2+}، So_4^{2-}، Cl^-، pH، EC و TDS رودخانه سفید رود در ایستگاه آستانه را با روش GP الگوبندی کردند. مقدار RMSE بهترین الگوهای انتخاب شده در مرحله آزمون برای متغیرهای کیفی فوق‌الذکر به ترتیب برابر ۱.۲، ۰۲.۰، ۸۵.۰، ۹۳.۰، ۱۸.۲، ۳۳.۰، ۱۵.۴۰۴ و ۱۵.۲۴۶ به‌دست آمدند. با مقایسه نتایج الگوبندی متغیرهای کیفی عروجی و همکاران (۱۸) در طول دوره آماری مشابه با تحقیق حاضر و جدول ۵ مشاهده می‌شود که انتخاب ورودی‌های الگو با روش ضریب همبستگی منجر به بهبود نتایج تا ۵.۵ برابر می‌شود.


Train
Test
Observed data
GA-LSSVR data
داده های
Na (meq/l)
30
25
20
15
10

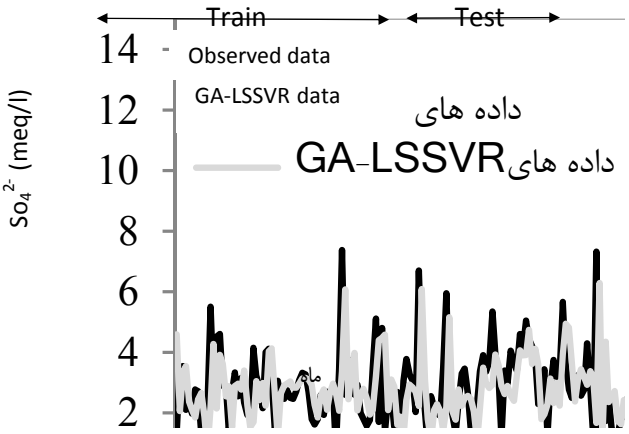
(a)


Train
Test
Observed data
GA-LSSVR data
داده های
K (meq/l)
0.5
0.4
0.3
0.2
0.1

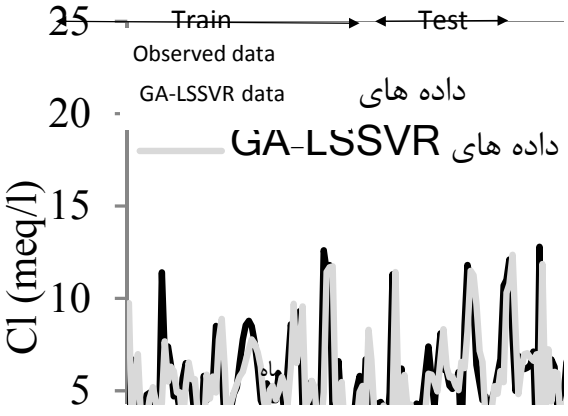
(b)


Train
Test
Observed data
GA-LSSVR data
داده های
Mg (meq/l)
10
8
6
4
2

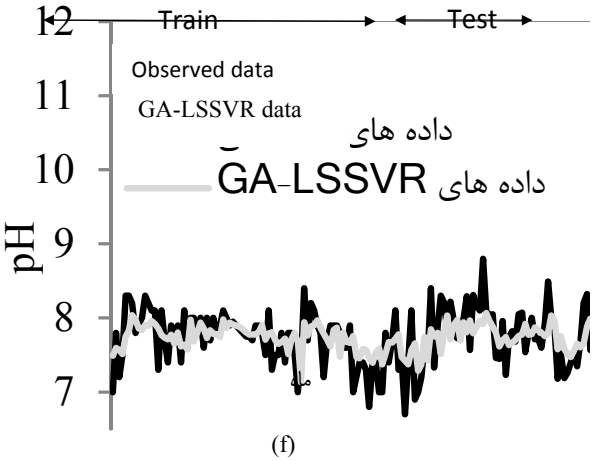
(c)


Train
Test
Observed data
GA-LSSVR data
داده های
So_4^{2-} (meq/l)
14
12
10
8
6
4
2

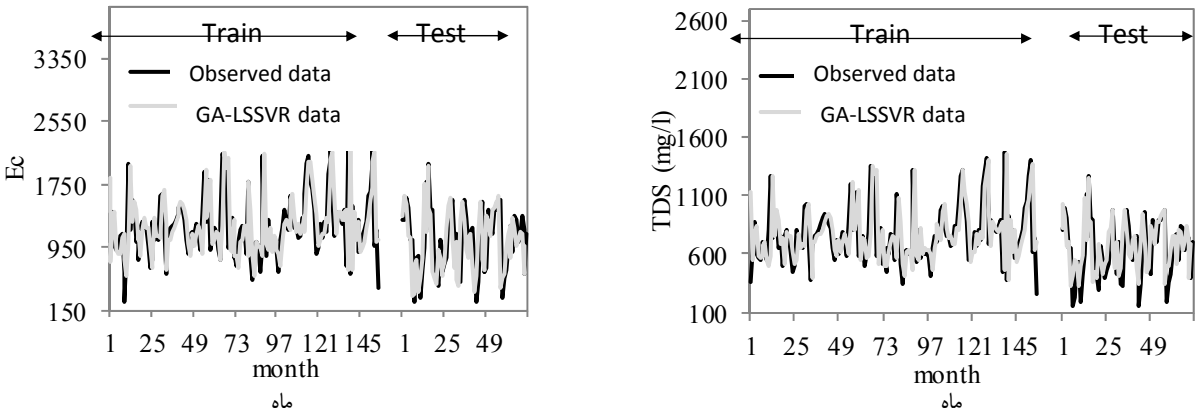
(d)


Train
Test
Observed data
GA-LSSVR data
داده های
Cl (meq/l)
25
20
15
10
5

(e)


Train
Test
Observed data
GA-LSSVR data
داده های
pH
12
11
10
9
8
7

(f)



شکل ۸- مقادیر مشاهداتی و محاسباتی با الگوریتم GA-LSSVR و روش پیش‌پردازش ضریب همبستگی متغیرهای کیفیت آب رودخانه سفید

Figure 8- The observation and calculated values using GA-LSSVR algotithm and corrolation coeifficent method of quality parameters in the Sefidrood river

(a) Na^{+}, (b) K^{+}, (c) Mg^{2+}, (d) So_4^{2-}, (e) Cl^{-}, (f) pH, (g) EC, and (h) TDS

نتیجه‌گیری کلی

در این تحقیق از الگوریتم GA که به عنوان یکی از پرکاربردترین الگوریتم‌های بهینه‌بندی در علوم مختلف مطرح است، جهت بهینه‌بندی ضرایب روش LSSVR استفاده شد و الگوریتم GA-LSSVR جهت الگوبندی متغیرهای کیفی آب توسعه داده شد. به منظور مقایسه روش‌های پیش‌پردازش ضریب همبستگی و PCA، داده‌های ورودی به الگوریتم GA-LSSVR از هر دو روش ضریب همبستگی و PCA بدست آمد. در ادامه الگوریتم GA-LSSVR برای متغیرهای ورودی که با روش ضریب همبستگی و PCA بدست آمده بودند، به طور جداگانه اجرا شد. در مورد متغیرهای کیفی pH، EC و TDS، هرچند که روش PCA از دقت بیش‌تری برخوردار است، ولی اختلاف بین *RMSE* روش ضریب همبستگی و روش PCA بسیار ناچیز است. به طوری که روش PCA نسبت به روش ضریب همبستگی منجر به بهبود مقدار *NS* به میزان ۲۲ و ۱.۰ درصد به ترتیب برای متغیرهای کیفی pH و TDS شده است و مقدار آن برای متغیر EC تغییر نکرده است. از آنجایی که مجموع معیار *NS* برای داده‌های آزمون در روش ضریب همبستگی ۶۲.۵ و در روش PCA برابر ۵۳.۵ است. لذا، نتایج بدست آمده از روش ضریب همبستگی و PCA اختلاف چندان و معناداری با هم ندارند. در هر دو روش ضریب همبستگی و PCA، مقادیر R^2 و *NS* نشان‌دهنده همبستگی بالا بین مقادیر محاسباتی و مشاهداتی می‌باشند، لذا استفاده از هر دو روش در الگوبندی داده‌های کیفی با به‌کارگیری روش LSSVR با به کارگیری ضرایب بهینه پیشنهاد می‌شود. بهبودی نتایج تحقیق حاضر تا ۵.۵ برابر نسبت به نتایج تحقیق عروجی و همکاران (۱۴) با دوره آماری مشابه نشان می‌دهد که الگوریتم تلفیقی پیشنهادی و انتخاب ورودی‌های الگو با روش ضریب همبستگی قابلیت تقریب بسیار بالایی در الگوبندی متغیرهای کیفی را دارد.

منابع

1- Bozorg Haddad O., Afshar A., and Mariño M.A. 2011. Multireservoir optimisation in discrete and continuous domains, Proceedings of the Institution of Civil Engineers: Water Management, 164(2), 57-72.

2- Bozorg Haddad O., Fallah-Mehdipour E., Mirzaei-Nodoushan F., and Mariño M.A. 2014a. Discussion of A GA-based support vector machine model for the prediction of monthly reservoir storage, Journal of Hydrologic Engineering, DOI: 10.1061.(ASCE)HE.1943-5584.0001086.

3- Bozorg Haddad O., Moravej M., and Loáiciga H. 2014b. Application of the water cycle algorithm to the optimal operation of reservoir systems, Journal of Irrigation and Drainage Engineering, DOI: 10.1061.(ASCE)IR.1943-4774.0000832 ,04014064.

4- Camdevyren H., Demyr N., Kanik A., and Keskyn, S. 2005. Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs, Ecological Modelling, 181 (4), 581-589.

5- Chiu D.Y., and Chen P.J. 2009. Dynamically exploring internal mechanism of stock market by fuzzy-based support

vector machines with high dimension input space and genetic algorithm, Expert Systems with Applications, 36(2), 1240-1248.
6- Fallah-Mehdipour E., Bozorg Haddad O., and Mariño M.A. 2013. Prediction and simulation of monthly groundwater levels by genetic programming, Journal of Hydro-Environment Research, 7(4), 253-260.
7- Ghavidel S. Z.Z., and Montaseri M. 2014. Application of different data-driven methods for the prediction of total dissolved solids in the Zarinehroud basin, Stochastic Environmental Research and Risk Assessment, 28(8), 2101-2118.
8- Johnson R.A., and Wichern D.W. 1982. Applied multivariate statistical analysis, Prentice Hall, No 3, Englewood Cliffs, SA.
9- Koza J. R. 1990. Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems, Department of Computer Science, Stanford University, 131pp.
10- Liu S., Tai H., Ding Q., Li D., Xu L., and Wei, Y. 2013, A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction, Mathematical and Computer Modelling, 58(3), 458-465.
11- Noori R., Ashrafi Kh., and Ajdarpour A. 2008. Comparison of ANN and PCA based multivariate linear regression applied to predict the daily average concentration of Co: A case study of Tehran, Journal of Physics Earth Space, 34(1), 135-152.
12- Noori R., Karbassi A., and Salman Sabahi, M. 2010. Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste prediction, Journal of Environmental Management, 91(3), 767-771.
13- Noori R., Karbassi A. R., Moghaddamnia A., Han D., Zokaei-Ashtiani M.H., Farokhnia A., and Gousheh M. G. 2011. 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(3), 177-189.
14- Orouji H., Bozorg Haddad O., Fallah-Mehdipour E., and Mariño M.A. 2013. Modeling of water quality parameters using data-driven models, Journal of Environmental Engineering, 139(7), 947-957.
15- Ouyang, Y. 2005. Evaluation of river water quality monitoring stations by principal component analysis, Water Research, 39(12), 2621-2635.
16- Raghavendra N.S., and Deka P.C. 2014. Support vector machine applications in the field of hydrology: A review, Applied Soft Computing, 19, 372-386.
17- Rajaee T., Mirbagheri S.A., Zounemat-Kermani M., and Nourani, V. 2009. Daily suspended sediment concentration simulation using ANN and neuro-fuzzy models, Science of the Total Environment, 407(17), 4916-4927.
18- Singh K.P., Basant N., and Gupta S. 2011. Support vector machines in water quality management, Analytica Chimica Acta, 703(2), 152-162.
19- Soltani F., Kerachian R., and Shirangi E. 2010. Developing operating rules for reservoirs considering the water quality issues: Application of ANFIS-based surrogate models, Expert Systems with Applications, 37(9), 6639-6645.
20- Su J., Wang X., Liang Y., and Chen B. 2013. A GA-based support vector machine model for the prediction of monthly reservoir storage, Journal of Hydrologic Engineering, 19(7), 1430-1437.
21- Suykens J.A.K., Van Gestel T., De Brabanter J., De Moor B., and Vandewalle J. 2002. Least squares support vector machines, World Scientific Publishing, No. 4, Singapore.
22- Tabachnick B.G., and Fidell, L.S. 2001. Using multivariate statistics, Pearson, No. 2, 963 pp.
23- Tan G., Yan J., Gao C., and Yang, S. 2012. Prediction of water quality time series data based on least squares support vector machine, Procedia Engineering, 31, 1194-1199.
24- Vapnik V.N. 1995. The nature of statistical learning theory, Springer, New York, USA.
25- Wang W.C., Chau K.W., Cheng C.T., and Qiu L. 2009. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series, Journal of Hydrology, 374(3), 294-306.
26- Yoon H., Jun S.C., Hyun Y., Bae G.O., and Lee K.K. 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, Journal of Hydrology, 396(1-2), 128-138.
27- Yunrong X., and Liangzhong J. 2009. Water quality prediction using LS-SVM with particle swarm optimization, Second International Workshop on Knowledge Discovery and Data Mining, IEEE 2009, Moscow, Russia, January 23-25.

Journal of Water and Soil
Vol. 30, No. 3, Jul.-Agu. 2016, p. 743-757

نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۰، شماره ۳، مرداد – شهریور ۱۳۹۵، ص. ۷۵۷-۷۴۳

Modeling Water Quality Parameters Using Data-driven Methods

Sh. Soleimani[1]- O. Bozorg Haddad[2]- M. Moravej[3*]

Received: 26-01-2015
Accepted: 11-10-2015

Introduction: Surface water bodies are the most easily available water resources. Increase use and waste water withdrawal of surface water causes drastic changes in surface water quality. Water quality, importance as the most vulnerable and important water supply resources is absolutely clear. Unfortunately, in the recent years because of city population increase, economical improvement, and industrial product increase, entry of pollutants to water bodies has been increased. According to that water quality parameters express physical, chemical, and biological water features. So the importance of water quality monitoring is necessary more than before. Each of various uses of water, such as agriculture, drinking, industry, and aquaculture needs the water with a special quality. In the other hand, the exact estimation of concentration of water quality parameter is significant.

Material and Methods: In this research, first two input variable models as selection methods (namely, correlation coefficient and principal component analysis) were applied to select the model inputs. Data processing is consisting of three steps, (1) data considering, (2) identification of input data which have efficient on output data, and (3) selecting the training and testing data. Genetic Algorithm-Least Square Support Vector Regression (GA-LSSVR) algorithm were developed to model the water quality parameters. In the LSSVR method is assumed that the relationship between input and output variables is nonlinear, but by using a nonlinear mapping relation can create a space which is named feature space in which relationship between input and output variables is defined linear. The developed algorithm is able to gain maximize the accuracy of the LSSVR method with auto LSSVR parameters. Genetic algorithm (GA) is one of evolutionary algorithm which automatically can find the optimum coefficient of Least Square Support Vector Regression (LSSVR). The GA-LSSVR algorithm was employed to model water quality parameters such as Na^{+}, K^{+}, Mg^{2+}, So_4^{2-}, Cl^{-}, pH, Electric conductivity (EC) and total dissolved solids (TDS) in the Sefidrood River. For comparison the selected input variable methods coefficient of determination (R^2), root mean square error (*RMSE*), and Nash-Sutcliff (*NS*) are applied.

Results and Discussion: According to Table 5, the results of the GA-LSSVR algorithm by using correlation coefficient and PCA methods approximately show similar results. About *pH*, *EC*, and *TDS* quality parameters, the results of PCA method have, the more accuracy, but the difference of *RMSE* between the PCA method and correlation coefficient method is not significant. The PCA method cause improvement in *NS* values to 22 and 0.1 percentages in pH and TDS water quality parameters to the correlation coefficient method, respectively,and *NS* criteria value for EC water quality parameter did not change in both methods. As a result, according to positive values of *NS* criteria in both PCA and correlation methods, it is clear that GA-LSSVR has a high ability for modeling of water quality parameters. Because of summation of *NS* criteria for PCA method is 5.53 and for correlation coefficient is 5.62, we can say that the correlation coefficient method has more applicable as a data processing method, but both methods have a high ability. Orouji et all. (18) used assumed models to model Na^{+}, K^{+}, Mg^{2+}, So_4^{2-}, Cl^{-}, pH, EC, and TDS by Genetic programming (GP) method. The *RMSE* criteria of the better models for testing data are 2.1, 0.02, 0.85, 0.93, 2.18, 0.33, 404.15, and 246.15, respectively. For comparison the orouji et al. (18) and table (5), the Results show using the correlation coefficient method as a data processing method can improve the results to 5.5 times. The results indicate the superiority of developingalgorithm increases the modeling accuracy. It is worth mentioning that according to *NS* criteria both selected inputs variable methods (correlation coefficient and PCA) are capable to model the water quality parameters. Also the result shows that using correlation coefficient method lead to more accurate results than PCA.

Conclusion: In this study, GA algorithm as one of the most applicable optimization algorithms in the different sciences was used to optimize the LSSVR coefficients and Then GA-LSSVR was developed to model the water quality parameters. To comparison data processing methods (correlation coefficient and PCA methods), the input variables of both methods were determined and GA-LSSVR was performed for each of the input variables. To compare the results of the PCA and correlation coefficient methods, some statistics were

1, 2 and 3- M.Sc. Student, Associate Professor and Ph.D. Candidate, Department of Irrigation and Reclamation, Faculty of Agricultural Engineering and Technology, College of Agriculture and Natural Resources, University of Tehran
(*-Coressponding Author Email: ShimaSoleimani@ut.ac.ir)

used. It is worth mentioning that according to *NS* criteria both input selection methods are capable to model water quality parameters. Also the results show that using correlation coefficient method lead to more accurate results than PCA.

Keywords: GA-LSSVR Algorithm, Pearson correlation coefficient, PCA method