توسعه روش رگرسیون K- نزدیک‌ترین همسایگی در پیش‌بینی جریان رودخانه

محمد عزمی[1] شهاب عراقی‌نژاد[2]

(دریافت ۸۸/۱۲/۱۰ پذیرش ۹۰/۵/۱۷)

چکیده

روشهای مختلف آماری، غیرآماری و جعبه سیاه در فرایندهای پیش‌بینی جریان رودخانه استفاده می‌شوند. از میان روشهای آماری، روش رگرسیون ناپارامتری K- نزدیک‌ترین همسایگی به‌واسطه پایه ریاضی و سادگی ذاتی، یکی از روشهای مناسب در فرایندهای پیش‌بینی است. در این تحقیق ضمن معرفی کامل روش K-NN به تشریح راهکارهای توسعه و بهبود این روش پرداخته می‌شود که از آن جمله می‌توان به معرفی روشهای تخمین بهترین همسایگی، توابع انتقال اطلاعات (پیش‌پردازش)، توابع فاصله‌سنجی و روش پیشنهادی برای برونیابی اشاره کرد. روش پیش‌بینی K-NN به‌همراه راهکارهای توسعه آن بر روی مطالعه موردی پیش‌بینی آورد حوضه بالادست سد زاینده‌رود اجرا شد. مقایسه نتایج نهایی روش K-NN کلاسیک با روش اصلاح شده K-NN (تعداد همسایگی ۵، تابع انتقال دامنه مقیاس، تابع فاصله سنجی ماهانالوبیس و اعمال روش برونیابی پیشنهادی) نشان می‌دهد که مدل بهبود یافته در پارامترهای نکویی برازش، ریشه میانگین مربعات خطا، درصد حجم خطا و میزان همبستگی به‌ترتیب ٤٥ ، ٥٩ و ١٧درصد بهبود عملکرد داشته است. این نتایج، ضرورت اعمال راهکارهای ذکر شده را برای استخراج پیش‌بینی‌های دقیق‌تر نشان می‌دهد.

واژه‌های کلیدی: روش نزدیک‌ترین همسایگی، توابع فاصله‌سنجی، فاصله ماهانالوبیسی، برونیابی، رودخانه زاینده‌رود

Development of K-Nearest Neighbour Regression Method in Forecasting River Stream Flow

Mohammad Azmi[1] *Shahab Araghinejad*[2]

(Received Feb. 28, 2010 Accepted Aug. 7, 2011)

Abstract

Different statistical, non-statistical and black-box methods have been used in forecasting processes. Among statistical methods, K-nearest neighbour non-parametric regression method (K-NN) due to its natural simplicity and mathematical base is one of the recommended methods for forecasting processes. In this study, K-NN method is explained completely. Besides, development and improvement approaches such as best neighbour estimation, data transformation functions, distance functions and proposed extrapolation method are described. K-NN method in company with its development approaches is used in streamflow forecasting of Zayandeh-Rud Dam upper basin. Comparing between final results of classic K-NN method and modified K-NN (number of neighbour 5, transformation function of Range Scaling, distance function of Mahanalobis and proposed extrapolation method) shows that modified K-NN in criteria of goodness of fit, root mean square error, percentage of volume of error and correlation has had performance improvement 45% , 59% and 17% respectively. These results approve necessity of applying mentioned approaches to derive more accurate forecasts.

Keywords: Nearest Neighbour Method, Distance Functions, Mahanalobis Distance, Extrapolation, Zayandeh-Rud River.

۱- دانشجوی دکترای مهندسی منابع آب، دانشکده فناوری و مهندسی کشاورزی، دانشگاه تهران (نویسنده مسئول) (۰۲۱) ۲۳۹۶۱۲۲۸ mazami@ut.ac.ir

۲- استادیار مهندسی منابع آب، دانشکده فناوری و مهندسی کشاورزی، دانشگاه تهران

1. Ph. D. Student of Water Resources Eng., College of Tech. and Agricultural Eng., Tehran University, Tehran (Corresponding Author) (+98 21) 23961228 mazami@ut.ac.ir
2. Assist. Prof. of Water Resources Eng., College of Tech. and Agricultural Eng., Tehran University, Tehran

۱- مقدمه

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 تخمین پارامترها در مرحله واسنجی مدل، تخمین زده می‌شوند. در مدل‌های ناپارامتری، مرحله تخمین پارامترها وجود ندارد. یکی از شناخته شده‌ترین مدل‌های ناپارامتری روش K- نزدیک‌ترین همسایه است. این روش به‌طور گسترده‌ای در علوم مختلف از جمله شناخت الگوهای داده‌ها و دسته‌بندی اطلاعات مورد استفاده قرار گرفته است [۱ و ۲]. دلیل گستردگی استفاده از این روش را شاید بتوان در سادگی ذاتی این روش و کاربردها و قابلیتهای آن دانست. پیش‌بینی انجام شده با این روش به‌صورت مجموعه‌ای از جوابهای محتمل با ارائه مقدار احتمال هر یک از آنها انجام می‌شود. محققان بسیاری در دهه‌های اخیر مزایای استفاده از این روش ناپارامتری را در مدل‌سازی سری‌های زمانی مارکوفی، در تئوری و کاربرد نشان دادند[۳ و ۴].

در دهه اخیر روشهایی برای بهبود روش K-NN در فرایندهای پیش‌بینی ارائه گردیده است. شرما و همکاران[1] با استفاده از تابع چگالی غیر پارامتری کرنل و استفاده از شاخصهای انزو[2]، توزیع احتمالاتی بارش حوضه سد واراگامبای استرالیا را ارائه نمودند [۵]. عراقی‌نژاد و برن[3] نشان دادند که می‌توان با استفاده از تلفیق روشهای زمین‌آماری و روش K-NN به پیش‌بینی دقیق‌تر پارامترهای هواشناسی و هیدرولوژیکی پرداخت و از نتایج پیش‌بینی‌ها برای مدیریت بلندمدت منابع آب اقدام نمود [۶]. اسدیانی یکتا از الگوریتم K-NN و مدل استنتاج تطبیقی عصبی- فازی برای برآورد میزان بار معلق ورودی به مخزن سد اکباتان استفاده کرد [۷]. هدف از این تحقیق یافتن رابطه‌ای بین میزان رسوبات معلق با دبی رودخانه با کمترین میزان خطا نسبت به‌روش سنتی سنجه رسوب بوده است. نوری و همکاران نشان دادند که اعمال روشهای پیش‌پردازش آنالیز مؤلفه‌های اصلی و موجک می‌تواند تأثیر بسزایی در بهبود عملکرد پیش‌بینی جریان ماهانه با استفاده از شبکه‌های عصبی داشته باشد [۸].

عزمی و همکاران نشان داده‌اند که می‌توان با استفاده از الگوریتم نزدیک‌ترین همسایگی، نتایج خروجی حاصل از مدل‌های منفرد پیش‌بینی را به‌گونه‌ای ترکیب کرد که نتایج نهایی حاصل از این ترکیب اطلاعات[4] دقیق‌تر و مطلوب‌تر از نتایج مدل‌های منفرد و حتی روشهای ترکیب اطلاعات کلاسیک باشد [۹].

در روش آماری ناپارامتری K-NN دو عامل مهم دخیل است. اولین عامل، انتخاب تابع فاصله‌سنجی و وزن‌های مربوطه و دومی انتخاب بهترین تعداد همسایگی است. در انجام پیش‌بینی‌ها توسط K-NN، استفاده از تابع فاصله سنجی اقلیدسی وزن‌دارشده[5] بسیار معمول است. روش K-NN با اختصاص وزن‌های بیشتر به همسایگی‌های منتخبی که از نظر فاصله زمانی نسبت به زمان حاضر نزدیک‌تر می‌باشند توانسته است نتایج پیش‌بینی را نسبت به اتفاقات همسایگی‌های نزدیک‌تر به زمان فعلی وابسته‌تر گرداند[۱۰ و ۱۱]. تحقیقات گسترده‌ای برای انتخاب بهترین همسایگی و وزن‌ها انجام شده است [۱۲-۱۴]. همچنین تصحیح و یا حذف اطلاعات پرت و یا مشکوک می‌تواند توانایی و قدرت کارایی این روش را افزایش دهد. لذا انتخاب تعداد متغیرهای پیش‌بینی کننده و طول آماری مناسب می‌تواند تأثیر شایانی را برروی دقت مدل داشته باشد [۱۵].

هدف از این تحقیق، اعمال راهکارهایی به‌منظور بهبود و رفع کمبودهای روش K- نزدیک‌ترین همسایگی برای استفاده در فرایندهای پیش‌بینی بود. از آنجایی که بارزترین نقطه ضعف موجود در این روش پیش‌بینی، عدم توانایی در برونیابی مقادیر پیش‌بینی است لذا معرفی یک روش پیشنهادی به‌منظور برونیابی می‌تواند مهم‌ترین نوآوری موجود در این تحقیق باشد.

۲- مواد و روشها

۲-۱- مفاهیم اولیه و الگوریتم کلی استفاده از K-NN در فرایندهای پیش‌بینی

از مزایای استفاده از الگوریتم K-NN در فرایندهای پیش‌بینی، می‌توان به موارد زیر اشاره کرد [۱، ۳ و ۱۳]

۱- اجرای ساده

۲- عدم نیاز به مرحله تخمین پارامترها

۳- قابلیت مدل‌سازی غیرخطی

۴- مؤثر بودن و عملکرد با بازدهی بالا در برخورد با تعداد دسته‌های زیاد از داده‌ها.

از محدودیتهای استفاده از الگوریتم K-NN در فرایندهای پیش‌بینی می‌توان به موارد زیر اشاره کرد [۱۳ و ۱۶]:

از آن جا که این مدل سعی در شناسایی الگوهای مشابه در سری تاریخی و استفاده از آنها در پیش‌بینی دارد، وجود اطلاعات کافی برای واسنجی آن لازم است. اطلاعات کوتاه مدت ممکن است

[1] Sharma et al.
[2] ENSO
[3] Burn
[4] Data Fusion
[5] Weighted Euclidean

منجر به خطاهای زیادی در مدل‌سازی با استفاده از این الگوریتم شود.

همان‌گونه که از روابط مربوط به ساختار روش K-NN برای تخمین اطلاعات توسط این الگوریتم بر می‌آید این الگوریتم قادر نیست مقادیر بزرگ‌تر از بیشترین مقدار مشاهده شده تاریخی و کوچک‌تر از کمترین مقدار مشاهده شده تاریخی تولید کند. به‌عبارت دیگر این الگوریتم تنها توانایی درونیابی بین اطلاعات را دارد و قادر به انجام برونیابی نیست. بنابراین استفاده از این الگوریتم در پیش‌بینی مقادیر تا حدی ممکن است منجر به تولید خطاهای چشمگیر شود.

در روش K-NN تابع توزیع مقادیر پیش‌بینی با استفاده از توزیع ناپارامتری تابع کرنل به‌دست می‌آید. مفهوم مورد استفاده 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 به‌عنوان گزینه‌های محتمل در شرایط فعلی درنظر 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. دو عامل مهم در به‌کارگیری روش K-NN، تعداد همسایه‌ها (k) و وزن پیش‌بینی کننده‌ها (w_j) هستند. بر اساس توضیحات ارائه شده، الگوریتم انجام پیش‌بینی با استفاده از روش K-NN به‌صورت زیر است:

۱- بردار سطری m ستونه مقادیر متغیرهای پیش‌بینی کننده x_j در زمان t به‌صورت زیر است

$$Pr_{jt} = (x_{jt}) \quad j = 1...m \qquad (۱)$$

ماتریس m ستونه و n سطری از مقادیرمتغیرهای پیش‌بینی کننده x_j درسری زمانی تاریخی به‌صورت زیر است

$$Pr_{j,(t-i)} = (x_{j,(t-i)}) \quad j = 1...m, i = 1...n \qquad (۲)$$

با استفاده از تابع فاصله‌سنج Dist، فواصل بین‌بردار Prj,t با سطرهای ماتریس Prj,(t-i) استخراج می‌گردد

$$Dist(t-i) = f(w_j, x_{j,(t-i)}, x_{jt}) \qquad (۳)$$

که در این رابطه

اندیس j نشان‌دهنده متغیرهای پیش‌بینی کننده و اندیس i بیان کننده گام زمانی در سری تاریخی است. مقادیر w_j، وزن‌هایی است که برای پیش‌بینی کننده‌ها درنظر گرفته می‌شود.

در این تحقیق برای تعیین وزن‌ها از روش صحت‌سنجی متقاطع تعمیم داده شده[1] که توسط تاربوتون و همکاران[2] نیز به‌کار گرفته شده، استفاده شد. رابطه محاسبه GCV به‌صورت زیر است[۱۵]

$$GCV = \frac{\sum_{i=1}^{n} e_i^2 / n}{(1 - 1/\sum_{j=1}^{k} 1/j)^2} \qquad (۴)$$

که در این رابطه

n تعداد داده‌ها و e_i خطای پیش‌بینی‌ها و k تعداد بهترین همسایگی است. در این فرایند پس از انتخاب بهترین همسایگی، مقادیر مختلفی برای وزن‌ها درنظر گرفته می‌شود و در نهایت وزن‌هایی که کمترین GCV را نتیجه می‌دهند، انتخاب می‌گردند. محدوده وزن‌ها را می‌توان به نسبت میزان همبستگی بین هر یک از متغیرهای پیش‌بینی کننده با متغیر وابسته تعیین کرد.

۲- از یک تابع کرنل گسسته برای وزن‌دهی به همسایه‌ها استفاده می‌شود. یاکوویتز[3]، تابع کرنل زیر را پیشنهاد داده است [۱]

$$K(Dist(t-i)) = \frac{1/Dist(t-i)}{\sum_{i=1}^{k} 1/Dist(t-i)} \qquad (۵)$$

نزدیک‌ترین همسایه‌ها براساس وزن محاسبه شده از ۱ تا K رتبه‌بندی می‌شوند به‌نحوی که همسایه با بیشترین وزن، کمترین رتبه را دارد و بر عکس. چنانچه دو یا چند همسایه دارای یک وزن باشند، رتبه همسایه‌ای که در سری تاریخی فاصله زمانی کمتری با زمان پیش‌بینی دارد، کمتر درنظر گرفته می‌شود. اینکه بتوان تعیین کرد کدام تعداد همسایگی به‌عنوان بهترین تعداد همسایگی است یکی از دغدغه‌های موجود در این روش است. در نهایت مقدار پیش‌بینی از رابطه زیر محاسبه می‌شود

$$D(t) = \sum_{i=1}^{k} K(Dist(t-i)) \times D(t-i) \qquad (۶)$$

که در این رابطه

$D(t-i)$، مقدار متغیر وابسته در زمان t-i و D(t) مقدار متغیر وابسته در زمان t است. شکل ۱ شماتیک روش پیش‌بینی K-NN را نشان می‌دهد.

۲-۲- روشهای بهبود تخمین‌های مدل K-NN

به‌طور کلی برای توسعه روش K-NN به چهار مورد زیر می‌توان اشاره کرد:

۱- توسعه روشهایی برای تخمین بهترین همسایه‌ها

۲- توسعه توابع انتقال اطلاعات

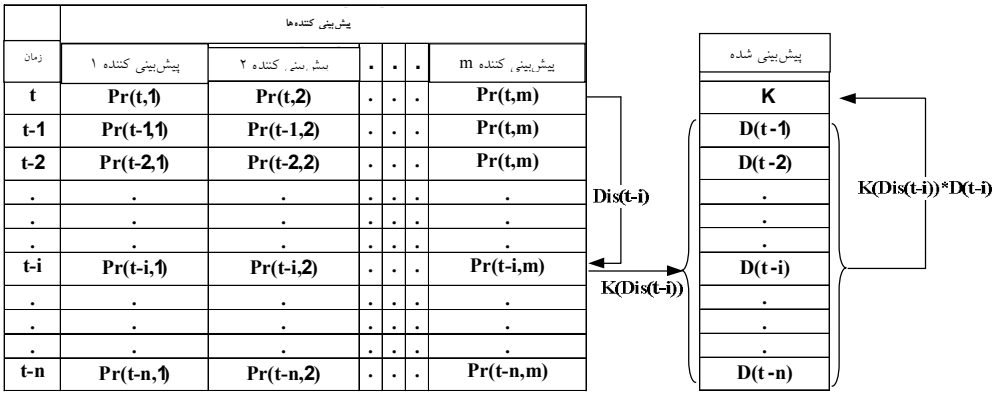
۳- توسعه توابع فاصله‌سنجی

۴- ارائه روشی برای برطرف کردن مشکل برونیابی روش K-NN

[1] Generalized Cross Validation (GCV)

[2] Varboton et al.

[3] Yakowitz



شکل ۱- شماتیک روش پیش‌بینی K-NN با استفاده از تابع کرنل

۲-۲-۱- توسعه روشهایی برای تخمین بهترین همسایگی

برای تخمین بهترین همسایگی‌ها در روش ناپارامتری K-NN ، روش‌های مختلفی ارائه شده‌اند که بسته به دقت و مورد استفاده و پیچیدگی و حجم مسئله قابل استفاده هستند. به‌طور کلی چهار روش معمول زیر برای تخمین بهترین همسایگی مورد استفاده قرار می‌گیرد:

۱- روابط تجربی

فرمول $K = \sqrt{n}$ به‌عنوان تقریب مناسبی از محدوده بهترین همسایگی مورد استفاده قرار می‌گیرد که در این رابطه n طول سری زمانی و K بهترین تعداد همسایگی مورد استفاده در روش K-NN است. میزان کارایی این رابطه با افزایش طول سری زمانی افزایش می‌یابد[۱۷].

۲- روش سعی و خطا

این روش با انتخاب همسایگی‌های مختلف در محدوده بهترین همسایگی و استخراج خطاهای پیش‌بینی، سعی در پیدا کردن بهترین همسایگی‌ها دارد.

۳- روش واریوگرافی[1]

این روش با تلفیق مفاهیم زمین‌آماری سعی دارد که بهترین همسایگی را تخمین بزند. به‌طور کلی مفاهیم زمین‌آماری در بعد مکانی طرح‌ریزی شده و مورد استفاده قرار می‌گیرند هرچند که می‌توان این مفاهیم را به بعد زمانی منتقل کرده و مورد استفاده قرار داد. عراقی‌نژاد و برن نشان دادند که چگونه می‌توان مفاهیم مکانی زمین آماری را با جایگزینی متغیرهای پیش‌بینی کننده به‌جای ابعاد x و y در بعد زمانی مورد استفاده قرار داد [۶]. در این تحقیق نیز با توجه به همین مسئله دو متغیر پیش‌بینی کننده که شامل میزان آورد تجمعی رودخانه در سه ماهه پاییز و مقدار میانگین شاخص اقلیمی نوسانات شمالی[2] در شش ماهه تابستان و پاییز هستند به‌جای ابعاد x و y در بعد مکانی جایگزین می‌گردند. واریوگرام تجربی مربوط به این فضای مجازی ترسیم شده و پس از برازش بهترین واریوگرام تئوریکی به این واریوگرام تجربی، پارامتر دامنه[3] واریوگرام استخراج می‌شود که این مقدار، محدوده بهترین همسایگی را به‌خوبی نمایان می‌سازد.

۴- روشهای بهینه‌سازی[4]

اساس کار در روشهای بهینه‌سازی بر مبنای یک الگوریتم بهینه‌سازی است که در جهت بهینه کردن تابع هدف حرکت می‌کنند. این تابع هدف می‌تواند برابر حداقل میزان خطا بین متغیرهای واقعی و متغیرهای پیش‌بینی شده باشد. الگوریتم‌های بهینه‌سازی در مرحله آموزش روش K-NN برای بهینه کردن بهترین تعداد همسایگی (K) و همچنین وزن‌های مربوط به هر یک از متغیرهای پیش‌بینی کننده در تابع فاصله‌سنجی (w_i)، مورد استفاده قرار می‌گیرد. در این تحقیق از روش اول و سوم برای پیدا کردن بهترین محدوده همسایگی‌ها و از روش دوم برای استخراج بهترین همسایگی از این محدوده استفاده شد.

۲-۲-۲- توابع انتقال اطلاعات (پیش‌پردازش اطلاعات متغیرهای پیش‌بینی کننده)

با توجه به ابعاد و انواع مختلف اطلاعات و محدوده‌های متفاوت آنها که می‌توانند به‌عنوان ورودی‌های مدل (مقادیر متغیرهای

[1] Variography

[2] Southern Oscillations Index (SOI)
[3] Range parameter
[4] Evolutionary optimization methods

پیش‌بینی کننده) درنظر گرفته شوند لزوم همگن کردن آنها برای جلوگیری از خطاهای مربوط به عدم همگن بودن آنها در مدل احساس می‌شود. با توجه به اطلاعات خام می‌توان چهار رابطه معمول برای پردازش اطلاعات را معرفی کرد [۱۶]:

۱- استاندارد سازی[1]

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

$$X_{ij} = \frac{x_{ij} - m_j}{\sigma_j} \qquad (۷)$$

که در این رابطه

m_j میانگین متغیر j در طول دوره آماری i، σ_j انحراف معیار متغیر j در طول دوره آماری i و x_{ij} متغیر j که در طول دوره آماری i دارای اطلاعات است.

همان‌طور که بیان شد i معرف تغییر مقدار متغیر j در طول دوره آماری مربوطه است که در تمامی رابطه‌های انتقال، تعریف یکسانی دارد.

۲- دامنه مقیاس[2]

در این روش با توجه به حداکثر و حداقل مقدار متغیر j در طول دوره آماری، مقادیر متغیرها کوچک می‌شوند. رابطه مربوطه این انتقال به‌شرح زیر است

$$X_{ij} = \frac{x_{ij} - L_j}{U_j - L_j} \qquad (۸)$$

که در آن

L_j حداقل مقدار متغیر j در طول دوره آماری i ، U_j حداکثر مقدار متغیر j در طول دوره آماری i، x_{ij} متغیر j که در طول دوره آماری i دارای اطلاعات است.

۳- حداکثر مقیاس[3]

در این روش با توجه به حداکثر مقدار متغیر j در طول دوره آماری، مقادیر متغیرها کوچک می‌شوند. رابطه این انتقال به‌شرح زیر است

$$X_{ij} = \frac{x_{ij}}{U_j} \qquad (۹)$$

که در این رابطه

U_j حداکثر مقدار متغیر j در طول دوره آماری i، x_{ij} متغیر j که در طول دوره آماری i دارای اطلاعات است.

۴- نمایه‌ها[4]

در این روش با توجه به مقادیر مشاهده شده در سطرها و ستون‌های ماتریس متغیرها در طول سری زمانی تاریخی، کلیه مقادیر متغیرهای مختلف به یک نسبت کوچک می‌شوند.

رابطه تابع انتقال این روش به‌شرح زیر است

$$X_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^2 \sum_j x_{ij}^2}} \qquad (۱۰)$$

که در آن

$\sum_i x_{ij}^2$ مجموع مربعات مقادیر هر متغیر در طول دوره زمانی مربوطه و $\sum_j x_{ij}^2$ مجموع مربعات مقادیر متغیرها به ازای هر گام زمانی است.

۵- آنالیز مؤلفه‌های اصلی[5]

روش آنالیز مؤلفه‌های اصلی، وابستگی‌ها و همبستگی‌های خطی میان متغیرها را نشان می‌دهد و سپس مقادیر همبسته متغیرها را با مقادیر غیر همبسته جایگزین می‌کند که این مقادیر جدید، ترکیبهای اصلی[6] نام دارد. مقادیر جدید متغیرها که از ترکیبات خطی مقادیر پیشین وزن‌دار شده به‌دست می‌آیند، امتیازات ترکیبات اصلی[7] نامیده می‌شوند [۱۸].

در این تحقیق، هدف اصلی از اجرای آنالیز مؤلفه‌های اصلی بر روی متغیرهای پیش‌بینی کننده، کاهش ابعاد و تعداد متغیر ورودی نبوده بلکه هدف، از بین بردن وابستگی میان متغیرها و استفاده از مقادیر مستقل برای متغیرهای پیش‌بینی کننده در مدل‌سازی بود.

۲-۲-۳- توابع فاصله‌سنجی

همانطور که بیان شد در روش K-NN نیاز به یک تابع فاصله‌سنجی به‌منظور یافتن بهترین همسایگی‌ها است که دو تابع فاصله‌سنجی اقلیدسی و ماهانالوبیس از معمول‌ترین این توابع هستند. تابع فاصله‌سنج اقلیدسی بر اساس فاصله متریک دو نقطه در فضای اقلیدسی تعریف می‌شود. چنانچه m متغیر پیش‌بینی کننده با طول سری زمانی تاریخی n درنظر گرفته شود این تابع فاصله‌سنجی را می‌توان به‌صورت زیر تعریف نمود

[1] Standardized
[2] Range scaling
[3] Maximum scaling
[4] Profiles
[5] Principal Component Analysis (PCA)
[6] Principal components
[7] Principal component scores

$$Dist(t-i)=\sqrt{\sum_{j=1}^{m} w_j (x_{j,(t-i)}-x_{j,t})^2} \qquad (۱۱)$$

که در این رابطه

$x_{j,(t-i)}$ متغیر پیش‌بینی کننده مشاهده شده j در سری تاریخی $i=1,...,n$، $x_{j,t}$ متغیر پیش‌بینی کننده مشاهده شده j در زمان t که متغیر وابسته در زمان t توسط این مقدار تخمین زده می‌شود و w_j وزن‌های مربوطه به هریک از متغیرهای پیش‌بینی کننده هستند.

تابع فاصله‌سنجی دیگری که توسط شریف و برن برای پیش‌بینی‌های هیدرولوژیکی مورد استفاده قرار گرفت تابع فاصله‌سنجی ماهانالوبیس بود که به‌واسطه دارا بودن فاکتوری که همبستگی میان متغیرهای پیش‌بینی کننده را درنظر می‌گیرد می‌تواند پاسخ‌گویی مناسبی را در پیش‌بینی‌ها داشته باشد [۱۷]. رابطه ۱۲ این تابع فاصله‌سنجی را نشان می‌دهد

$$Dist(t-i)=\sqrt{(G_{j,t}-G_{j,(t-i)})^T P^{-1}(G_{j,t}-G_{j,(t-i)})} \qquad (۱۲)$$

که در آن

G یک بردار سطری m ستونه از متغیرهای پیش‌بینی کننده و P^{-1}، ماتریس کواریانس میان سری‌های زمانی متغیرهای پیش‌بینی کننده است. همان‌طور که در فرمول ملاحظه می‌شود عامل ماتریس کواریانس در این رابطه مبین همبستگی میان متغیرهای پیش‌بینی‌کننده است و میزان تأثیر این همبستگی‌ها را نمایش می‌دهد. اساس و منطق توابع دیگری که در اینجا ارائه می‌گردند نیز در ریاضیات اثبات شده و در علوم مختلف و به‌خصوص مسائل اقتصادی برای کلاسه‌بندی و پیش‌بینی، مورد استفاده قرار گرفته‌اند [۱۶]. چنانچه m متغیر پیش‌بینی کننده برای فاصله‌سنجی در یک سری زمانی با طول n درنظر گرفته شود روابط توابع فاصله‌سنجی به‌صورت زیر تعریف می‌گردند

۱- تابع فاصله‌سنجی منهتن[1]

$$Dist(t-i)=\sum_{j=1}^{m} w_j \mid x_{j,t}-x_{j,(t-i)} \mid \qquad (۱۳)$$

۲- تابع فاصله سنجی کمبرا[2]

$$Dist(t-i)=\sum_{j=1}^{m} w_j \left(\frac{x_{j,t}-x_{j,(t-i)}}{x_{j,t}+x_{j,(t-i)}}\right)^2 \qquad (۱۴)$$

۳- تابع فاصله‌سنجی لنس- ویلیامز[3]

$$Dist(t-i)=\frac{\sum_{j=1}^{m} w_j \mid x_{j,t}-x_{j,(t-i)} \mid}{\sum_{j=1}^{m} w_j \mid x_{j,t}+x_{j,(t-i)} \mid} \qquad (۱۵)$$

۴- ضریب کوسین[4]

$$Dist(t-i)=\frac{1}{m}\left(\sum_{j=1}^{m} w_j \frac{x_{j,t}\times x_{j,(t-i)}}{x_{j,(t-i)}^2}\right) \qquad (۱۶)$$

که در این روابط

$x_{j,(t-i)}$ مقدار مشاهده شده متغیر پیش‌بینی کننده j در زمان t-i در سری تاریخی $i=1,...,n$، $x_{j,t}$ مقدار مشاهده شده متغیر پیش‌بینی کننده j در زمان t در سری تاریخی $i=1,...,n$ که متغیر وابسته در زمان t توسط این مقدار تخمین زده می‌شود و w_j : وزن‌های مربوطه به هریک از متغیرهای پیش‌بینی کننده هستند.

مقادیر وزن‌های w_j براساس نوع مسئله می‌تواند متفاوت باشد. به‌این معنی که چنانچه تمامی متغیرهای پیش‌بینی کننده دارای اهمیت یکسانی در فرایند پیش‌بینی باشند تمامی w_jها، مقدار یک را به‌خود اختصاص خواهند داد و چنانچه اهمیت متغیرها در فرایند پیش‌بینی یکسان نباشد مقدار وزن‌ها متناسب با اهمیت متغیرها تغییر خواهند کرد. مجموع وزن‌ها در زمانی که متغیرهای پیش‌بینی کننده دارای اهمیت یکسانی نیستند برابر یک است. روشهای بسیاری برای تعیین مقدار وزن‌های توابع فاصله‌سنجی ارائه شده است که از روشهای تجربی تا الگوریتم‌های بهینه‌سازی فراکاووشی گسترده شده‌اند. در این تحقیق میزان اهمیت کلیه متغیرهای پیش‌بینی کننده یکسان و برابر یک درنظر گرفته شد.

۲-۲-۴- روش پیشنهادی برای برونیابی در روش K-NN

همان‌طور که پیش از این نیز بیان شد روش K-NN توانایی پیش‌بینی مقادیر خارج از محدوده مقادیر سری تاریخی را ندارد. به بیان دیگر روش K-NN قادر به برونیابی مقادیر پیش‌بینی نیست. برای حل این مسئله باید مراحل زیر دنبال شود:

۱- استخراج مقادیر خطای پیش‌بینی‌های مرحله آموزش

$$E=[Z_i]-[\hat{Z}_i] \qquad (۱۷)$$

که در این رابطه

Z_i بردار مقادیر مشاهده‌ای متغیر وابسته در سری تاریخی آموزش، $\hat{Z}_i$ بردار مقادیر پیش‌بینی شده متغیر وابسته توسط روش K-NN در مرحله آموزش، E بردار مقادیر خطا در مرحله آموزش است.

[1] Manhattan distance function
[2] Cambera distance function
[3] Lance-villiams distance function
[4] Cosine coefficient

۲- اجرای مرحله آزمایش K-NN و ترکیب مقادیر خطای مرحله آموزش با مقادیر پیش‌بینی شده در این مرحله

$$Z_p = KNN_{Z_i}(x_i) + KNN_E(x_i) \qquad (۱۸)$$

که در این رابطه

$KNN_{Z_i}(x_i)$ نتیجه اجرای روش K-NN در مرحله آزمایش برروی مقادیر متغیر وابسته، $KNN_{E_i}(x_i)$ نتیجه اجرای روش K-NN در مرحله آزمایش برروی مقادیر خطای مرحله آموزش و Z_P مقدار پیش‌بینی شده متغیر وابسته در مرحله آزمایش است.

به‌طور خلاصه در این بخش، راهکارهای بهبود نتایج در روش K-NN ارائه گردید و همان‌طور که بیان شد این راهکارها در چهاربخش تخمین بهترین همسایگی، انتقال اطلاعات، توابع فاصله‌سنجی و نهایتاً روش برونیابی بود. از روشهای بهبود تخمین بهترین همسایگی می‌توان به روابط تجربی، روش سعی و خطا، واریوگرافی و روش فراکاوشی اشاره کرد. توابع انتقال اطلاعات را نیز می‌توان در پنج رابطه استاندارد کردن، دامنه مقیاس، حداکثر مقیاس، نمایه‌ها و نهایتا آنالیز مؤلفه‌های اصلی خلاصه نمود. از توابع فاصله‌سنجی که به‌طور معمول در علوم اقتصاد و ریاضیات برای کلاسه‌بندی و پیش‌بینی مقادیر سری‌های زمانی استفاده می‌شوند می‌توان به چهار تابع منهتن، کمبرا، لنس-ولیامز و ضریب کوسین اشاره نمود. در قسمت آخر این بخش روشی برای رفع مشکل برونیابی روش K-NN ارائه گردید که در این روش با استخراج مقادیر خطای پیش‌بینی‌ها در مرحله آموزشی و سپس افزودن آنها در مراحل آزمایش، عملیات برونیابی صورت می‌گیرد. سه معیار نکویی برازش برای مقایسه نتایج روشهای مختلف در دو بخش آموزش و آزمایش مورد استفاده قرار گرفت. اولین معیار سنجش، مجذور میانگین مربعات خطا[1] است که به‌صورت زیر ارائه می‌شود

$$RMSE = \sqrt{\frac{\sum_i^n (obs_i - for_i)^2}{n}} \qquad (۱۹)$$

معیار سنجش دیگری که در این تحقیق برای ارزیابی روشها به‌کار گرفته شد معیار میانگین حجم خطا[2] است که از رابطه زیر به‌دست می‌آید

$$\%VE = \frac{\sum_i^n \left| \frac{obs_i - for_i}{obs_i} \right|}{n} \times ۱۰۰ \qquad (۲۰)$$

[1] Root Mean Square Error (RMSE)
[2] Volume error

که در این رابطه

obs_i مقادیر مشاهده‌ای، for_i مقادیر پیش‌بینی شده، n تعداد مقادیر است.

آخرین معیار سنجش که در پیش‌بینی‌های هیدرولوژیکی دارای ارزش بالایی است معیار ضریب همبستگی[3] بین مقادیر مشاهده‌ای و پیش‌بینی شده است. بالا بودن این معیار برای یک مدل بیان می‌کند که مدل، ین توانایی را دارد که بتواند تغییرات ناگهانی رخدادها را به‌خوبی پیش‌بینی کند. این معیار سنجش را می‌توان از رابطه زیر تخمین زد

$$Corr\% = \frac{Cov(obs, for)}{\sigma_{obs} \times \sigma_{for}} \qquad (۲۱)$$

که در این رابطه

obs سری زمانی مشاهده‌ای، for سری زمانی پیش‌بینی شده، Cov(obs,for) کواریانس بین مقادیر مشاهده‌ای و مقادیر پیش‌بینی شده، σ_{obs} انحراف معیار مقادیر مشاهده‌ای، σ_{for}: انحراف معیار مقادیر پیش‌بینی شده است.

در بخش بعدی تمامی راه‌کارهای بهبود روش K-NN به‌مراه روش کلاسیک این روش در پیش‌بینی آورد حوضه بالادست سد زاینده‌رود مورد استفاده قرار گرفته و نتایج آن مورد ارزیابی قرار می‌گیرد.

۲-۳- مطالعه موردی

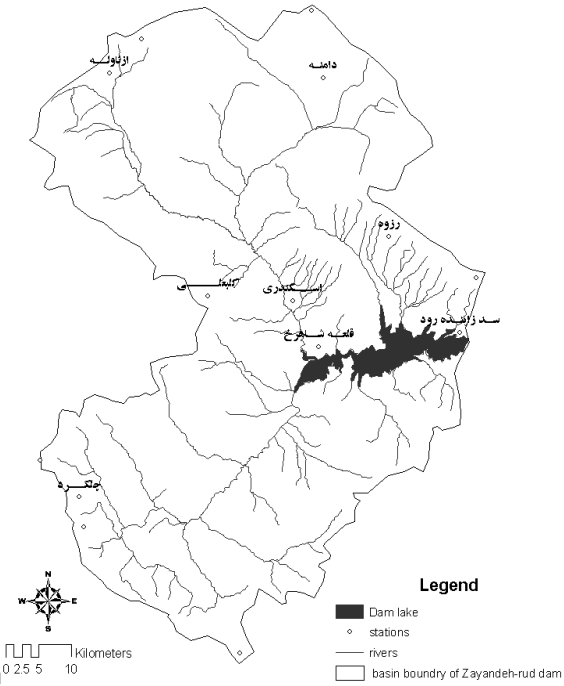
منطقه مورد مطالعه، محدوده حوضه بالادست سد زاینده‌رود در استان اصفهان بود. این منطقه دارای تعدادی ایستگاههای هواشناسی سینوپتیک و اقلیمی است (شکل ۲). متوسط میزان بارش سالانه از سال ۱۳۴۹ تا ۱۳۷۹ برابر ۲۱۱ میلی‌متر در هر سال بوده است. این حوضه دارای هشت ایستگاه سینوپتیک به نامهای ایستگاه ازناوله، دامنه، کلبعلی، رزوه، اسکندری، چلگرد، قلعه شاهرخ و ایستگاه سد زاینده‌رود است. متوسط میزان جریان ورودی به سد زاینده‌رود در همین ۳۰ ساله آماری برابر ۱۴۳۹ میلیون مترمکعب در هر سال است. عزمی و همکاران نشان داده‌اند که در مطالعه موردی این تحقیق، دو متغیر پیش‌بینی‌کننده میزان‌آورد تجمعی رودخانه در سه ماهه پاییز و متغیر پیش‌بینی‌کننده میانگین شاخص اقلیمی نوسانات شمالی در شش ماهه تابستان و پاییز می‌توانند بر روی پیش‌بینی مجموع آورد حوضه بالادست سد زاینده‌رود در نه ماهه بعد از پاییز (نقطه شروع پیش‌بینی) تأثیرگذار باشد [۹].

[3] Correlation

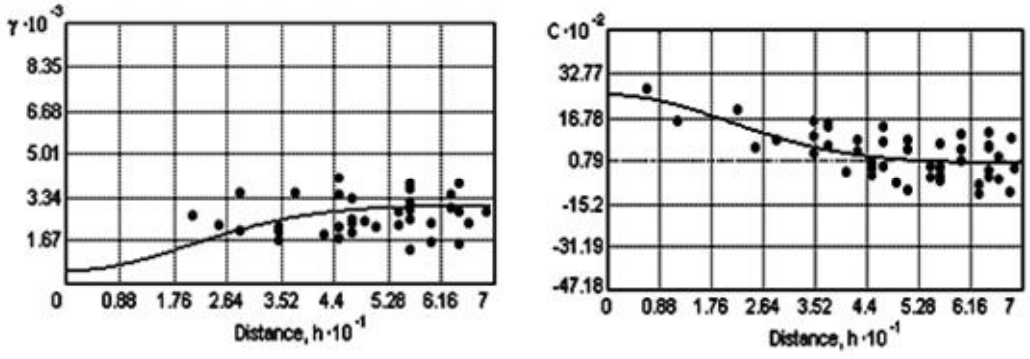
۳- نتایج و بحث

به‌دلیل سالانه بودن مقیاس پیش‌بینی‌ها که باعث محدود شدن تعداد اعضا در مجموعه سری تاریخی مورد استفاده در جستجوی الگوهای مشاهداتی مشابه وضع کنونی می‌شود، کلیه حالتهای مدل‌های اجرا شده در این تحقیق با روش صحت‌سنجی متقاطع[1] صورت گرفت. به‌طور خلاصه در این تحقیق ابتدا توسط سه روش روابط تجربی، روش سعی و خطا و روش واریوگرافی، بهترین تعداد همسایگی ارزیابی شد که با توجه به تعداد داده‌های سری زمانی (۳۰ سال) بهترین تعداد همسایگی ۵ یا ۶ ارزیابی شد. با توجه به‌روش واریوگرافی که نتایج آن در شکل ۳ آمده است، بهترین تعداد همسایگی عدد ۵ به‌دست آمد. همچنین برای اطمینان از استخراج بهترین همسایگی از روش سعی و خطا، همسایگی‌های ۷، ۶، ۵، ۴ مورد بررسی قرار گرفتند که نتیجه این قسمت نیز تعداد بهترین همسایگی را برای این مطالعه موردی، عدد ۵ معرفی می‌کند. نتایج این قسمت، دقت مناسب روش واریوگرافی و تا حدی روش روابط تجربی برای تخمین بهترین همسایگی را نشان می‌دهد. در روش سعی و خطا از تابع فاصله‌سنج اقلیدسی با اطلاعات بدون پیش‌پردازش (بدون تأثیر توابع انتقال اطلاعات) برای استخراج بهترین همسایگی استفاده شد.

[1] Cross-validation



شکل۲- منطقه مورد مطالعه، حوضه بالادست سد زاینده‌رود و ایستگاههای هواشناسی واقع شده در حوضه



$f \forall$ x

شکل۳- واریوگرام متغیرهای پیش‌بینی کننده (الف). نمودار کواریانس بین متغیرهای پیش‌بینی کننده (ب)

پس از تعیین بهترین تعداد همسایگی، شش تابع فاصله‌سنجی معرفی شده، با تعداد همسایگی۵، مورد استفاده قرار گرفتند. لازم به‌ذکر است که در این مرحله از اطلاعات ورودی خام (بدون پیش‌پردازش) برای اجرای مدل‌های K-NN استفاده گردید (جدول ۱). با توجه به جدول ۱ دیده می‌شود که تابع فاصله‌سنجی ماهانالوبیس توانسته است نسبت به سایرین، نتایج مطلوب‌تری را ارائه نماید به این معنی که در دو شاخص RMSE و VE٪ که مبین خطای مدل‌ها هستند، کمترین مقادیر (۴۴ و ۱۷درصد) و در شاخص CORR مقدار ۸۴ درصد را کسب نمود. علت این برتری را شاید بتوان در لحاظ کردن میزان همبستگی بین متغیرهای پیش‌بینی کننده درحین سنجش فاصله بین شرایط کنونی با شرایط سری تاریخی دانست. در پیش‌بینی‌های هیدرولوژیکی و هواشناسی، میزان همبستگی بین متغیرهای پیش‌بینی کننده مؤثر است لذا می‌توان این تابع فاصله‌سنجی را همواره یکی از گزینه‌های مورد نظر برای استفاده در بخش تابع فاصله‌سنجی روش K-NN دانست.

جدول۱- نتایج روش K-NN به ازای بهترین تعداد همسایگی ۵ و تغییر نوع تابع فاصله‌سنجی (بدون پیش‌پردازش اطلاعات ورودی)

بهترین تعداد همسایگی	نوع تابع فاصله-سنجی	RMSE	%VE	CORR
۵	اقلیدسی	۴۸	۲۱	۸۰
	ماهانالوبیس	۴۴	۱۷	۸۴
	منهتن	۴۷	۲۲	۸۳
	کمبرا	۴۹	۲۵	۸۰
	لنس-ویلیمز	۴۶	۲۴	۷۴
	کوسین	۵۴	۲۹	۷۳

پس از این تابع، دو تابع اقلیدسی و منهتن توانسته‌اند بهترین نتایج را کسب نمایند که با مقایسه دقیق‌تر بین این دو تابع، تابع منهتن نسبت به اقلیدسی برتری نه چندان محسوسی دارد. مقادیر شاخصهای آماری تابع فاصله‌سنجی منهتن برای RMSE برابر ۴۷، VE٪ برابر ۲۲درصد و برای CORR برابر ۸۳ درصد است. در میان شش تابع فاصله‌سنجی معرفی شده در این تحقیق، تابع کوسین با دارا بودن مقادیر شاخصهای آماری RMSE برابر ۵۴، VE٪ برابر ۲۹درصد و CORR برابر ۷۳ درصد ضعیف‌ترین نتایج را ارائه نمود.

از نتایج حاصله مشخص می‌شود که روش K-NN با تعداد همسایگی ۵ و تابع فاصله‌سنجی ماهانالوبیس می‌تواند بهترین ترکیب کارآمد برای برآورد مقادیر پیش‌بینی توسط روش K-NN باشد. برای بررسی میزان تأثیر پردازش اولیه اطلاعات توسط توابع انتقال، مقادیر متغیرهای پیش‌بینی کننده و وابسته توسط توابع انتقال اطلاعات مورد پردازش قرار گرفت و سپس توسط روش K-NN با تعداد همسایگی ۵ و تابع فاصله‌سنجی ماهانالوبیس، پیش‌بینی‌ها صورت گرفت.

نتایج تأثیر توابع انتقال اطلاعات در جدول ۲ ارائه شده است. در سطر دوم جدول ۲، نتایج روش K-NN بدون تأثیر توابع انتقال ذکر شد که این نتایج در حقیقت همان نتایج روش K-NN با تابع فاصله‌سنجی ماهانالوبیس است که در جدول ۱ در سطر سوم ارائه شده بود. تابع انتقال اطلاعات دامنه مقیاس توانسته است با شاخصهای آماری RMSE برابر ۳۹، VE٪ برابر ۱۳درصد و CORR برابر ۸۹ درصد بهترین تأثیر را بر روی روند پیش‌بینی‌ها داشته باشد. می‌توان دلیل برتری این تابع انتقال را در این مطالعه موردی، همگن کردن اطلاعات متغیرهای پیش‌بینی کننده دانست به‌این معنی که این تابع انتقال نه تنها تمامی مقادیر را به یک نسبت متناسب کوچک می‌کند بلکه مقادیر را مثبت نموده و تمامی متغیرهای پیش‌بینی کننده را در حد فاصل صفر تا یک قرار خواهد داد. این روش می‌تواند برای پردازش متغیرهای هیدرولوژیکی که تنها مقادیر صفر یا مثبت را به‌خود اختصاص می‌دهند در کنار متغیرهای هواشناسی مثل دما و اندیس‌های اقلیمی که مقادیر منفی را نیز می‌توانند قبول نمایند، روشی مؤثر و کارآمد برای همگن کردن باشد.

در جدول ۲ ضعیف‌ترین نتایج، مربوط به تابع آنالیز مؤلفه‌های اصلی است. نتایج این تابع بیانگر آن است که استفاده از این روش نه تنها جوابها را به‌سوی دقت بیشتر سوق نمی‌دهد بلکه می‌تواند نتایج را از حالت بدون تأثیر توابع انتقال نیز بدتر نماید. علت این است که در فرایندهای پیش‌بینی، دارا بودن همبستگی بین متغیرهای پیش‌بینی کننده با یکدیگر و نیز بین متغیرهای پیش‌بینی کننده و متغیر پیش‌بینی شده امری اساسی بوده و این روش به‌دلیل حذف این ویژگی از مجموعه اطلاعات، مدل را دچار خطاهای زیادی خواهد کرد. بنابراین ترکیب نهایی روش K-NN برای پیش‌بینی پیش از اعمال روش پیشنهادی برونیابی، استفاده از تابع انتقال اطلاعات دامنه مقیاس، تعداد همسایگی پنج و استفاده از تابع فاصله‌سنجی ماهانالوبیس خواهد بود.

جدول۲- نتایج روش K-NN به ازای تعداد همسایگی ۵ و تابع فاصله‌سنجی ماهانالوبیس با تاثیر و بدون تاثیر توابع انتقال

حالتهای مختلف	RMSE	%VE	CORR
بدون تاثیر توابع انتقال	۴۴	۱۷	۸۴
تابع انتقال استاندارد کردن	۴۳	۱۷	۸۵
تابع انتقال دامنه مقیاس	۳۹	۱۳	۸۹
تابع انتقال حداکثر مقیاس	۴۲	۱۵	۸۶
تابع انتقال نمایه‌ها	۴۶	۱۸	۸۱
تابع انتقال آنالیز مؤلفه‌های اصلی	۵۰	۲۵	۷۸

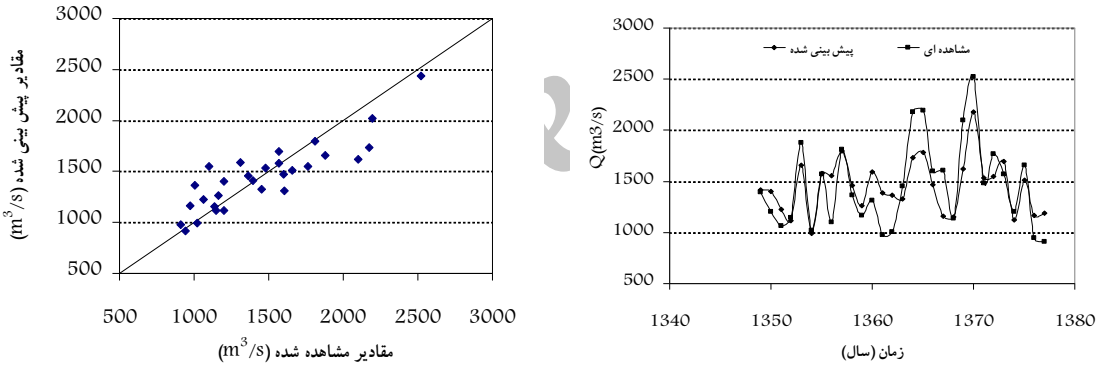
در بخش نهایی این تحقیق برای برطرف نمودن مشکل برونیابی روش K-NN بر روی ترکیب منتخب K-NN ، روش پیشنهادی برونیابی اعمال شد که نتایج آن در جدول ۳ ارائه شده است. نتایج، بیان کننده تأثیر مثبت این روش بر روی نتایج نهایی پیش‌بینی است. اگر سری تاریخی طولانی‌تر و دارای تنوع مقادیر حدی و غیرحدی باشد آنگاه تأثیر این روش برونیابی می‌تواند بیشتر نیز باشد زیرا مدل برای پیش‌بینی هر مقدار در آینده، وسعت انتخاب و تأثیرگذاری بیشتری خواهد داشت.

جدول۳- نتایج اجزا منتخب روش K-NN بدون تاثیر و با تاثیر روش برونیابی پیشنهادی

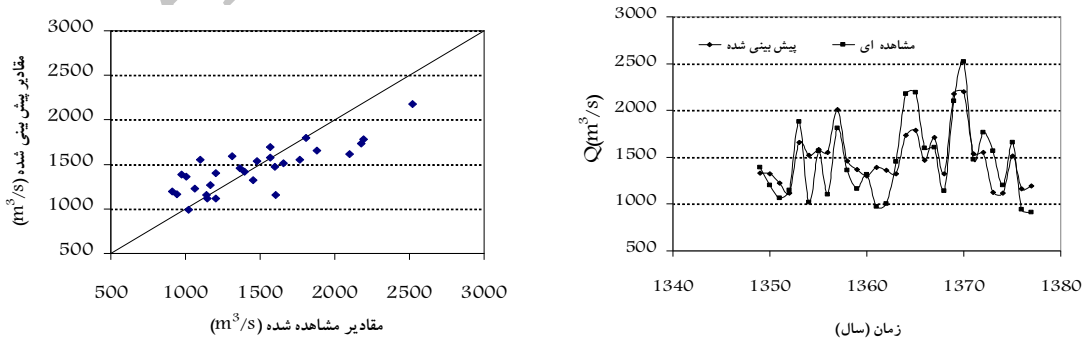
نوع رگرسیون	کل داده‌ها			بیشترین مقدار	کمترین مقدار
	RMSE	%VE	CORR	%VE	%VE
معمولی	۳۹	۱۳	۸۹	۴۰	۰/۵
با لحاظ کردن روش برونیابی	۲۶	۸/۵	۹۴	۲۵	۰/۵

نتایج تأثیر روش پیشنهادی برونیابی را می‌توان در شکلهای ۴ و ۵ ملاحظه کرد. همان‌طور که در این شکلها نیز مشخص است اعمال روش برونیابی به‌خصوص در سالهای ۱۳۷۰ (از نوع مقادیر حدی حداکثر) و سالهای ۱۳۷۷ تا ۱۳۷۹ (از نوع مقادیر حدی حداقل) توانسته است تأثیر مطلوبی بر روی نتایج پیش‌بینی‌ها داشته باشد. اثر این روش باعث شده است که مقادیر RMSE و VE٪ از ۳۹ و ۱۳درصد که برای ترکیب منتخب اجزای روش K-NN بدون تأثیر روش برونیابی است به مقادیر ۲۶ و ۸/۵ درصد کاهش یافته و نیز مقدار شاخص CORR از ۸۹ درصد به ۹۴درصد افزایش یابد که این حجم خطا در موارد پیش‌بینی مقادیر حدی می‌تواند بسیار با ارزش باشد.

مقایسه نتایج نهایی مدل K-NN کلاسیک با مدل نهایی K-NN با اعمال کلیه راهکارهای بهبود عملکرد نشان می‌دهد که مدل بهبود یافته در پارامترهای نکویی برازش RMSE ، VE٪ و CORR به ترتیب ۴۵، ۵۹ و ۱۷درصد بهبود عملکرد داشته است که این ارقام ضرورت اعمال راهکارهای ذکر شده را به‌منظور استخراج نتایج دقیق‌تر نشان می‌دهد.



شکل۴- نمودار نتایج روش K-NN با ترکیب منتخب بدون اعمال روش برونیابی پیشنهادی



شکل۵- نمودار نتایج روش K-NN با ترکیب منتخب با اعمال روش برونیابی پیشنهادی

۴- نتیجه‌گیری

روش K-NN به‌دلیل سادگی و عدم پیچیدگی بالا می‌تواند همواره یکی از بهترین گزینه‌ها برای انجام پیش‌بینی در علوم مختلف به‌خصوص هیدرولوژی و هواشناسی باشد. این روش برای شرایطی که دارای سری‌های تاریخی بلندمدت است دارای ارجحیت بالاتری خواهد بود زیرا تأثیر وقایع متنوع حدی و غیرحدی در پیش‌بینی‌های آتی می‌تواند بسیار مؤثر باشد. در این تحقیق تأثیر توابع انتقال اطلاعات برای پیش‌پردازش مقادیر متغیرهای پیش‌بینی کننده و نیز روشهای مختلف تخمین بهترین همسایگی مورد ارزیابی قرار گرفتند. همچنین توابع مختلف فاصله‌سنجی برای ارزیابی میزان اختلاف بین شرایط کنونی و شرایط گذشته بررسی شد. نهایتاً اعمال روش پیشنهادی برونیابی می‌تواند نتایج را، به‌ویژه در خصوص مقادیر حدی، به میزان قابل توجهی بهبود بخشد.

در مطالعه موردی این تحقیق، بهترین ترکیب منتخب برای روش K-NN ، استفاده از تابع انتقال اطلاعات دامنه مقیاس با تعداد همسایگی ۵ و تابع فاصله‌سنجی ماهانالوبیس و نهایتاً اعمال روش برونیابی بر روی مقادیر بود. مقایسه نتایج نهایی مدل K-NN کلاسیک با مدل نهایی K-NN با اعمال کلیه راهکارهای بهبود عملکرد نشان می‌دهد که مدل بهبود یافته در پارامترهای نکویی برازش RMSE ، VE ٪ و CORR به‌ترتیب ۴۵، ۵۹ و ۱۷درصد بهبود عملکرد داشته است که این ارقام ضرورت اعمال راهکارهای ذکر شده را برای استخراج نتایج دقیق‌تر نشان می‌دهد.

۵- مراجع

1- Yakowitz, S. J. (1985). "Nonparametric density estimation, prediction, and regression for markov sequences." *J. Am. Stat. Assoc.*, 80, 215-221.

2- Lall, U., and Sharma, A. (1996). "A nearest neighbor bootstrap for resampling hydrologic time series." *Water Resources Research*, 32(3), 679-694.

3- Karlsson, M., and Yakowitz, S. (1987). "Nearest-neighbor methods for nonparametric rainfall-runoff forecasting." *Water Resources Research*, 23(7), 1300-1308.

4- Toth, K., Brath A., and Montanari, A. (2000). "Comparison of short-term rainfall prediction models for real-time flood forecasting." *J. of Hydrology*, 239(4), 132-147.

5- Sharma, A., Luck, K. C., Cordery, I., and Lall, U. (2000). "Seasonal to interannual rainfall probabilistic forecast for improved water supply management: Part 2-Predictor Identification of quarterly rainfall using ocean-atmosphere information." *J. of Hydrology*, 239, 240-248.

6- Araghinejad, S. H., and Burn, D. (2005). "Probabilistic forecasting of hydrological events using geostatistical analysis." *Hydrological Sciences Journal- des Sciences Hydrologiques*, 50(5), 57-66.

7- Asadiani, Yekta, A., and Sultani, F. (2007). "Comparing sediment estimation of inflow load to Ekbatan dam between ANFIS and K-NN algorithm." *7th Conf. of Iran Hydraulic*, Shahid Abbaspour University, Tehran. (In Persian)

8- Noori, R., Farokhnia, A., Morid, S., and Riahi Madvar, H. R. (2008). "Effect of input variable properocessing in artificial neural network on monthly flow predication by PCA and wavelet transformation." *J. of Water and Wastewater*, 69, 13-23. (In Persian)

9- Azmi, M., Araghinejad, S., and Kholghi, M. (2010). "Multi model data fusion for hydrological forecasting using k- nearest neighbour method." *Iranian J. of Science and Technology, Transaction B, Engineering*, 34, 81-92.

10- Yates, D., Gangopadhyay, S., Rajagopalan, B., and Strzepek, K. (2003). "A technique for generating regional climate scenarios using a nearest-neighbor algorithm." *Water Resoures Research*, 39 (7), 1114-1121.

11- Sorjamaa, A., Reyhani, N., and Lendasse,A. (2005). "Input and structure selection for K-NN approximator." *8th Internatinal Conference on Artificial Neural Networks*, Lecture Notes in Computer Science Springer, IWANN, Berlin , 958-992.

12- Meade, N. (2002). "A comparison of the acuracy of short term foreign exchange forecasting methods." *International J. of Forecasting*, 18(1), 67-83.
13- Jayawardena, A. W., Li, W. K., and Xu, P. (2002). "Neighbor selection for local modelling and prediction of hydrological time series." *J. of Hydrology*, 258, 40-57.
14- Piechota, T.C., Chiew, F.H.S., Dracup, J.A., and McMahon, T.A. (2001). "Development of exceedence probability streamflow forecast." *J. of Hydrologic Engineering*, 6(1), 20-28.
15- Tarboton, D. G., Sharma, A., and Lall, U. (1993). "The use of non-parametric probability distribution in streamflow modeling." *In Proceeding of the 6 South African National Hydrological Symposium*, Ed. S. A. Lorentz et. Al., University of Natal, Pietermaritzburg, South Africa, 315-327.
16- Todeschini, R. (1989). "K-nearest neighbour method: Influence of data transformations and metrics." *Chemometrics and Intelligent Laboratory Systems*, 6, 213-220.
17- Sharif, M. H., and Burn, D. (2006). "Simulating climate change scenarios using an improved K-nearest neighbor model." *J. of Hydrology*, 325, 179-196.
18- Kshirsagar, A.M. (1972). *Multivariate analysis*, Marcel Decker, Inc., New York.