

توسعه روش رگرسیون K- نزدیک ترین همسایگی در پیش‌بینی جریان رودخانه

شهاب عراقی‌نژاد^۲

محمد عزمی^۱

(دریافت ۸۸/۱۲/۱۰ پذیرش ۹۰/۵/۱۷)

چکیده

روشهای مختلف آماری، غیر آماری و جعبه سیاه در فرایندهای پیش‌بینی جریان رودخانه استفاده می‌شوند. از میان روشهای آماری، روش رگرسیون ناپارامتری K- نزدیک‌ترین همسایگی به واسطه 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 به همراه 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 فاصله‌سنجی 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 به ترتیب ۴۵، ۵۹ و ۱۷ درصد 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 ماهانالوبیس، برونیابی، رودخانه زاینده‌رود

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

Mohammad Azmi¹

Shahab Araghinejad²

(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.

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

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

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-NN - 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^۱ با استفاده از تابع چگالی غیر پارامتری کرنل و استفاده از شاخصهای انزو^۲، توزیع احتمالاتی بارش حوضه سد واراگامبای استرالیا را ارائه نمودند [۵]. عراقی‌نژاد و برن^۳ 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 به روش سنتی سنج 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^۴ 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 وزن‌دار شده^۵ 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 بر روی دقت مدل داشته باشد [۱۵].

هدف از این تحقیق، اعمال راهکارهایی به‌منظور بهبود و رفع کمبودهای روش K-NN - 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 مدل سازی با استفاده از این الگوریتم شود.

همانگونه که از روابط مربوط به ساختار روش K-NN برای تخمین اطلاعات توسط این الگوریتم بر می آید این الگوریتم قادر نیست مقادیر بزرگ تر از بیشترین مقدار مشاهده شده تاریخی و کوچک تر از کمترین مقدار مشاهده شده تاریخی تولید کند. به عبارت دیگر این الگوریتم تنها توانایی درونیابی بین اطلاعات را دارد و قادر به انجام برونیابی نیست. بنابراین استفاده از این الگوریتم در پیش بینی مقادیر تا حدی ممکن است منجر به تولید خطاهای چشمگیر شود.

در روش K-NN تابع توزیع مقادیر پیش بینی با استفاده از توزیع ناپارامتری تابع کرنل به دست می آید. 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 می پردازد. وقایعی که در سری تاریخی در این الگوها پیش آمده اند می توانند به عنوان گزینه های محتمل در شرایط فعلی در نظر گرفته شوند. احتمال وقوع هر یک از این حالتها در شرایط حاضر، بستگی به شباهت بردار متغیرهای مستقل فعلی با بردار متغیرهای مستقل مشاهداتی در سری تاریخی دارد. دو عامل مهم در به کارگیری روش K-NN، تعداد همسایه ها (k) و وزن پیش بینی کننده ها (w_j) هستند. بر اساس توضیحات ارائه شده، الگوریتم انجام پیش بینی با استفاده از روش K-NN به صورت زیر است:

۱- بردار سطری m ستونه مقادیر متغیرهای پیش بینی کننده x_j در زمان t به صورت زیر است

$$Pr_{jt} = (x_{jt}) \quad j = 1 \dots m \quad (1)$$

ماتریس m ستونه و n سطری از مقادیر متغیرهای پیش بینی کننده x_j در سری زمانی تاریخی به صورت زیر است

$$Pr_{j,(t-i)} = (x_{j,(t-i)}) \quad j = 1 \dots m, i = 1 \dots n \quad (2)$$

با استفاده از تابع فاصله سنج Dist، فواصل بین بردار $Pr_{j,t}$ با سطریهای ماتریس $Pr_{j,(t-i)}$ استخراج می گردد

$$Dist(t-i) = f(w_j, x_{j,(t-i)}, x_{jt}) \quad (3)$$

که در این رابطه

اندیس j نشان دهنده متغیرهای پیش بینی کننده و اندیس i بیان کننده گام زمانی در سری تاریخی است. مقادیر w_j ، وزن هایی است که برای پیش بینی کننده ها در نظر گرفته می شود.

در این تحقیق برای تعیین وزن ها از روش صحت سنجی متقاطع تعمیم داده شده^۱ که توسط تاروتون و همکاران^۲ نیز به کار گرفته شده، استفاده شد. رابطه محاسبه GCV به صورت زیر است [۱۵]

$$GCV = \frac{\sum_{i=1}^n e_i^2 / n}{(1 - 1 / \sum_{j=1}^k 1 / j)^2} \quad (4)$$

که در این رابطه

n تعداد داده ها و e_i خطای پیش بینی ها و k تعداد بهترین همسایگی است. در این فرایند پس از انتخاب بهترین همسایگی، مقادیر مختلفی برای وزن ها در نظر گرفته می شود و در نهایت وزن هایی که کمترین GCV را نتیجه می دهند، انتخاب می گردند. محدوده وزن ها را می توان به نسبت میزان همبستگی بین هر یک از متغیرهای پیش بینی کننده با متغیر وابسته تعیین کرد.

۲- از یک تابع کرنل گسسته برای وزن دهی به همسایه ها استفاده می شود. یاکوویتز^۳، تابع کرنل زیر را پیشنهاد داده است [۱]

$$K(Dist(t-i)) = \frac{1/Dist(t-i)}{\sum_{i=1}^k 1/Dist(t-i)} \quad (5)$$

نزدیک ترین همسایه ها براساس وزن محاسبه شده از ۱ تا K رتبه بندی می شوند به نحوی که همسایه با بیشترین وزن، کمترین رتبه را دارد و بر عکس. چنانچه دو یا چند همسایه دارای یک وزن باشند، رتبه همسایه ای که در سری تاریخی فاصله زمانی کمتری با زمان پیش بینی دارد، کمتر در نظر گرفته می شود. اینکه بتوان تعیین کرد کدام تعداد همسایگی به عنوان بهترین تعداد همسایگی است یکی از دغدغه های موجود در این روش است. در نهایت مقدار پیش بینی از رابطه زیر محاسبه می شود

$$D(t) = \sum_{i=1}^k K(Dist(t-i)) \times D(t-i) \quad (6)$$

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 پیش بینی K-NN را نشان می دهد.

۲-۲- روشهای بهبود تخمین های مدل K-NN

به طور کلی برای توسعه روش K-NN به چهار مورد زیر می توان اشاره کرد:

۱- توسعه روشهایی برای تخمین بهترین همسایه ها

۲- توسعه توابع انتقال اطلاعات

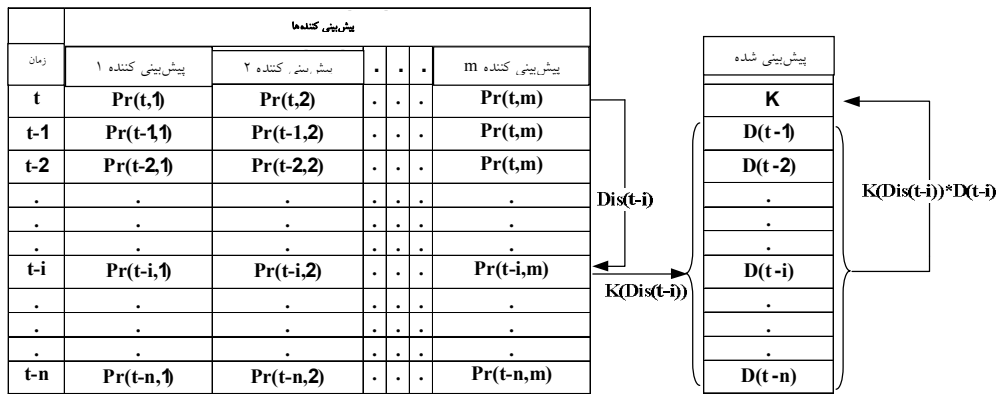
۳- توسعه توابع فاصله سنجی

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

³ Yakowitz

¹ Generalized Cross Validation (GCV)

² Varboton et al.



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

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 جمعی رودخانه در سه ماهه پاییز و مقدار میانگین شاخص اقلیمی نوسانات شمالی^۲ در شش ماهه تابستان و پاییز هستند به جای ابعاد x و y در بعد مکانی جایگزین می‌گردند. واریوگرام تجربی مربوط به این فضای مجازی ترسیم شده و پس از برازش بهترین واریوگرام تئوریک به این واریوگرام تجربی، پارامتر دامنه^۳ واریوگرام استخراج می‌شود که این مقدار، محدوده بهترین همسایگی را به خوبی نمایان می‌سازد.

۴- روشهای بهینه‌سازی^۴
 اساس کار در روشهای بهینه‌سازی بر مبنای یک الگوریتم بهینه‌سازی است که در جهت بهینه کردن تابع هدف حرکت می‌کند. 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 به عنوان ورودی‌های مدل (مقادیر متغیرهای

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. به طور کلی چهار روش معمول زیر برای تخمین بهترین همسایگی مورد استفاده قرار می‌گیرد:

۱- روابط تجربی
 فرمول $K = \sqrt{n}$ به عنوان 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^۱
 این روش با تلفیق مفاهیم زمین‌آماري سعی دارد که بهترین همسایگی را تخمین بزند. به طور کلی مفاهیم زمین‌آماري در بعد مکانی طرح‌ریزی شده و مورد استفاده قرار می‌گیرند هر چند 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 زمین‌آماري را با جایگزینی متغیرهای پیش‌بینی کننده به جای ابعاد

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

¹ Variography

پیش‌بینی کننده) در نظر 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^۱

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 به صورت زیر تعریف می‌شود

$$X_{ij} = \frac{x_{ij} - m_j}{\sigma_j} \quad (7)$$

که در این رابطه

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

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

۲- دامنه مقیاس^۲

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

$$X_{ij} = \frac{x_{ij} - L_j}{U_j - L_j} \quad (8)$$

که در آن

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

۳- حداکثر مقیاس^۳

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

$$X_{ij} = \frac{x_{ij}}{U_j} \quad (9)$$

که در این رابطه

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

۴- نمایه‌ها^۴

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 به شرح زیر است

$$X_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^2 \sum_j x_{ij}^2}} \quad (10)$$

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^۵

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^۶ نام دارد. مقادیر جدید متغیرها که از ترکیبات خطی مقادیر پیشین وزن‌دار شده به دست می‌آیند، امتیازات ترکیبات اصلی^۷ 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 به منظور یافتن بهترین همسایگی‌ها است که دو تابع فاصله‌سنجی اقلیدسی و ماهانالوویس 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 در نظر گرفته شود این تابع فاصله‌سنجی را می‌توان به صورت زیر تعریف نمود

⁴ Profiles

⁵ Principal Component Analysis (PCA)

⁶ Principal components

⁷ Principal component scores

¹ Standardized

² Range scaling

³ Maximum scaling

$$\text{Dist}(t-i) = \frac{\sum_{j=1}^m w_j |x_{j,t} - x_{j,(t-i)}|}{\sum_{j=1}^m w_j |x_{j,t} + x_{j,(t-i)}|} \quad (15)$$

۴- ضریب کوسین^۴

$$\text{Dist}(t-i) = \frac{1}{m} \left(\sum_{j=1}^m w_j \frac{x_{j,t} \times x_{j,(t-i)}}{x_{j,(t-i)}^2} \right) \quad (16)$$

که در این روابط

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

مقادیر وزن‌های w_j براساس نوع مسئله می‌تواند متفاوت باشد. به این معنی که چنانچه تمامی متغیرهای پیش‌بینی کننده دارای اهمیت یکسانی در فرایند پیش‌بینی باشند تمامی w_j ها، مقدار یک را به خود 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 فراکاوشی گسترده شده‌اند. در این تحقیق میزان اهمیت کلیه متغیرهای پیش‌بینی کننده یکسان و برابر یک در نظر 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

$$E = [Z_i] - [\hat{Z}_i] \quad (17)$$

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.

⁴ Cosine coefficient

$$\text{Dist}(t-i) = \sqrt{\sum_{j=1}^m w_j (x_{j,(t-i)} - x_{j,t})^2} \quad (11)$$

که در این رابطه

$x_{j,(t-i)}$ متغیر پیش‌بینی کننده مشاهده شده زدر سری تاریخی $x_{j,t}$ ، $i=1, \dots, n$ متغیر وابسته در زمان t توسط این مقدار تخمین زده می‌شود و w_j وزن‌های مربوطه به هر یک 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 به واسطه دارا بودن فاکتوری که همبستگی میان متغیرهای پیش‌بینی کننده را در نظر می‌گیرد می‌تواند پاسخ‌گویی مناسبی را در پیش‌بینی‌ها داشته باشد [۱۷]. رابطه ۱۲ این تابع فاصله‌سنجی را نشان می‌دهد

$$\text{Dist}(t-i) = \sqrt{(G_{j,t} - G_{j,(t-i)})^T P^{-1} (G_{j,t} - G_{j,(t-i)})} \quad (12)$$

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 پیش‌بینی کننده 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 به خصوص 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 در نظر گرفته شود روابط توابع فاصله‌سنجی به صورت زیر تعریف می‌گردند

۱- تابع فاصله‌سنجی منهتن^۱

$$\text{Dist}(t-i) = \sum_{j=1}^m w_j |x_{j,t} - x_{j,(t-i)}| \quad (13)$$

۲- تابع فاصله‌سنجی کمبرا^۲

$$\text{Dist}(t-i) = \sum_{j=1}^m w_j \left(\frac{x_{j,t} - x_{j,(t-i)}}{x_{j,t} + x_{j,(t-i)}} \right)^2 \quad (14)$$

۳- تابع فاصله‌سنجی لنس - ویلیامز^۳

¹ Manhattan distance function

² Cambera distance function

³ Lance-villiams distance function

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

$$Z_p = KNN_{Z_i}(x_i) + KNN_{E_i}(x_i) \quad (18)$$

که در این رابطه

$KNN_{Z_i}(x_i)$ نتیجه اجرای روش K-NN در مرحله آزمایش بر روی مقادیر متغیر وابسته، $KNN_{E_i}(x_i)$ نتیجه اجرای روش K-NN در مرحله آزمایش بر روی مقادیر خطای مرحله آموزش و Z_p مقدار پیش بینی شده متغیر وابسته در مرحله آزمایش است.

به طور خلاصه در این بخش، راهکارهای بهبود نتایج در روش K-NN ارائه گردید و همان طور که بیان شد این راهکارها در چهاربخش تخمین بهترین همسایگی، انتقال اطلاعات، توابع فاصله سنجی و نهایتاً روش برونابی بود. از روشهای بهبود تخمین بهترین همسایگی می توان به روابط تجربی، روش سعی و خطا، واروگرافی و روش فراکوشی اشاره کرد. توابع انتقال اطلاعات را نیز می توان در پنج رابطه استاندارد کردن، دامنه مقیاس، حداکثر مقیاس، نمایه ها و نهایتاً آنالیز مؤلفه های اصلی خلاصه نمود. از توابع فاصله سنجی که به طور معمول در علوم اقتصاد و ریاضیات برای کلاسه بندی و پیش بینی مقادیر سری های زمانی استفاده می شوند می توان به چهار تابع منهن، کمبرا، لنس-وليامز و ضریب کوسین اشاره نمود. در قسمت آخر این بخش روشی برای رفع مشکل برونابی روش K-NN ارائه گردید که در این روش با استخراج مقادیر خطای پیش بینی ها در مرحله آموزشی و سپس افزودن آنها در مراحل آزمایش، عملیات برونابی صورت می گیرد. 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^۱ است که به صورت زیر ارائه می شود

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (obs_i - for_i)^2}{n}} \quad (19)$$

معیار سنجش دیگری که در این تحقیق برای ارزیابی روشها به کار گرفته شد معیار میانگین حجم خطا^۲ است که از رابطه زیر به دست می آید

$$\%VE = \frac{\sum_{i=1}^n \frac{|obs_i - for_i|}{obs_i}}{n} \times 100 \quad (20)$$

¹ Root Mean Square Error (RMSE)

² Volume error

که در این رابطه
obs_i مقادیر مشاهده ای، for_i مقادیر پیش بینی شده، n تعداد مقادیر است.

آخرین معیار سنجش که در پیش بینی های هیدرولوژیکی دارای ارزش بالایی است معیار ضریب همبستگی^۳ بین مقادیر مشاهده ای و پیش بینی شده است. بالا بودن این معیار برای یک مدل بیان می کند که مدل، یکنواختی را دارد که بتواند تغییرات ناگهانی رخدادها را به خوبی پیش بینی کند. این معیار سنجش را می توان از رابطه زیر تخمین زد

$$Corr\% = \frac{Cov(obs, for)}{\sigma_{obs} \times \sigma_{for}} \quad (21)$$

که در این رابطه

obs سری زمانی مشاهده ای، for سری زمانی پیش بینی شده، Cov(obs, for) کواریانس بین مقادیر مشاهده ای و مقادیر پیش بینی شده، σ_{obs} انحراف معیار مقادیر مشاهده ای، σ_{for} انحراف معیار مقادیر پیش بینی شده است.

در بخش بعدی تمامی راهکارهای بهبود روش K-NN به همراه روش کلاسیک این روش در پیش بینی آورد حوضه بالادست سد زاینده رود مورد استفاده قرار گرفته و نتایج آن مورد ارزیابی قرار می گیرد.

۲-۳- مطالعه موردی

منطقه مورد مطالعه، محدوده حوضه بالادست سد زاینده رود 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 میلی متر 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 زاینده رود است. متوسط میزان جریان ورودی به سد زاینده رود در همین ۳۰ ساله آماری برابر ۱۴۳۹ میلیون مترمکعب در هر سال است. عزمی و همکاران نشان داده اند که در مطالعه موردی این تحقیق، دو متغیر پیش بینی کننده میزان آورد تجمعی رودخانه در سه ماهه پاییز و متغیر پیش بینی کننده میانگین شاخص اقلیمی نوسانات شمالی در شش ماهه تابستان و پاییز می تواند بر روی پیش بینی مجموع آورد حوضه بالادست سد زاینده رود در سه ماهه بعد از پاییز (نقطه شروع پیش بینی) تأثیرگذار باشد [۹].

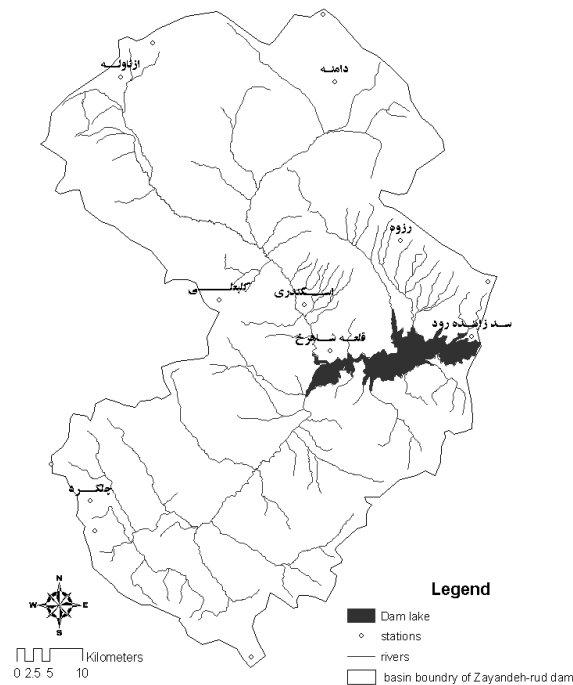
³ Correlation

۳- نتایج و بحث

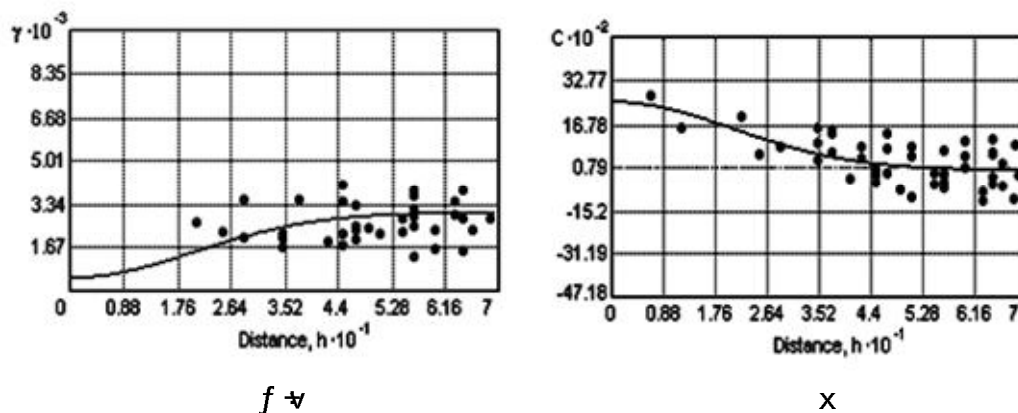
به دلیل سالانه بودن مقیاس پیش‌بینی‌ها که باعث محدود شدن تعداد اعضا در مجموعه سری تاریخی مورد استفاده در جستجوی الگوهای مشاهداتی مشابه وضع کنونی می‌شود، کلیه حالت‌های مدل‌های اجرا شده در این تحقیق با روش صحت‌سنجی متقاطع^۱ صورت گرفت. به طور خلاصه در این تحقیق ابتدا توسط سه روش روابط تجربی، روش سعی و خطا 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 به روش

واریوگرافی که نتایج آن در شکل ۳ آمده است، بهترین تعداد همسایگی عدد ۵ به دست آمد. همچنین برای اطمینان از استخراج بهترین همسایگی از روش سعی و خطا، همسایگی‌های ۴، ۵، ۶، ۷ 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.

^۱ Cross-validation



شکل ۲- منطقه مورد مطالعه، حوضه بالادست سد زاینده‌رود و ایستگاه‌های هواشناسی واقع شده در حوضه



شکل ۳- 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 همسایگی ۵، مورد استفاده قرار گرفتند. لازم به ذکر 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 ۱۷ درصد) 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 در حین 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.

جدول ۱- 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 ۲۲ درصد 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 ۲۹ درصد 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 ۱۳ درصد 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 به این 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 به خود 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 به دلیل 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.

جدول ۲- نتایج روش K-NN به ازای تعداد همسایگی ۵ و تابع فاصله‌سنجی ماهانالوبیس با تأثیر و بدون تأثیر توابع انتقال

| حالت‌های مختلف | RMSE | %VE | CORR |
|-----------------------------------|------|-----|------|
| بدون تأثیر 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 |

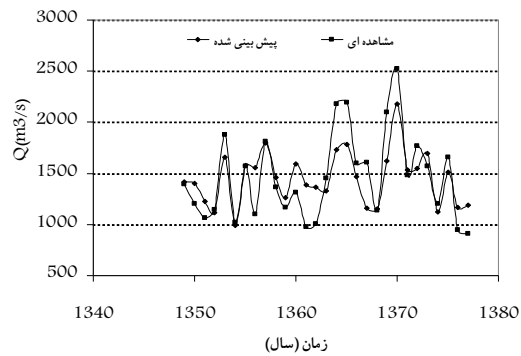
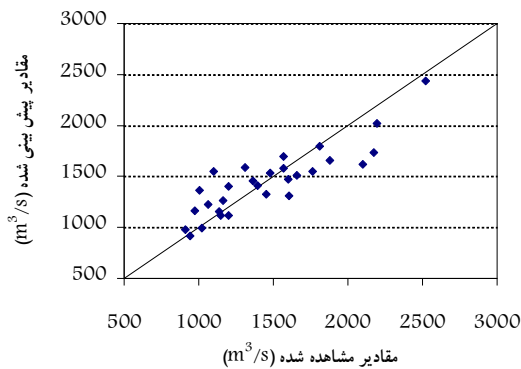
نتایج تأثیر روش پیشنهادی برونابی را می‌توان در شکل‌های ۴ و ۵ ملاحظه کرد. همان‌طور که در این شکل‌ها نیز مشخص است اعمال روش برونابی به‌خصوص در سال‌های ۱۳۷۰ (از نوع مقادیر حدی حدی حداکثر) و سال‌های ۱۳۷۷ تا ۱۳۷۹ (از نوع مقادیر حدی حداقل) توانسته است تأثیر مطلوبی بر روی نتایج پیش‌بینی‌ها داشته باشد. اثر این روش باعث شده است که مقادیر RMSE و %VE از ۳۹ و ۱۳ درصد که برای ترکیب منتخب اجزای روش K-NN بدون تأثیر روش برونابی است به مقادیر ۲۶ و ۸/۵ درصد کاهش یافته و نیز مقدار شاخص CORR از ۸۹ درصد به ۹۴ درصد 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 برآزش RMSE، %VE و CORR به ترتیب ۴۵، ۵۹ و ۱۷ درصد بهبود عملکرد داشته است که این ارقام ضرورت اعمال راهکارهای ذکر شده را به‌منظور استخراج نتایج دقیق‌تر نشان می‌دهد.

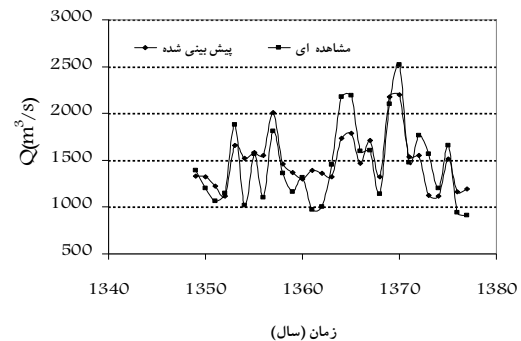
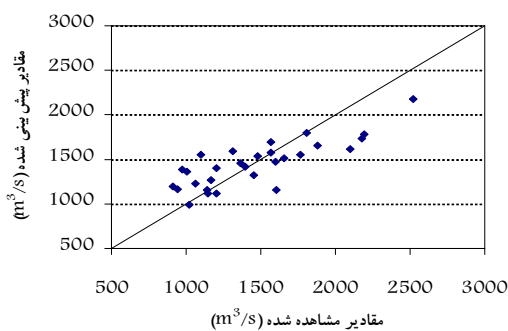
در بخش نهایی این تحقیق برای برطرف نمودن مشکل برونابی روش K-NN بر روی ترکیب منتخب K-NN، روش پیشنهادی برونابی اعمال شد که نتایج آن در جدول ۳ ارائه شده است. نتایج، بیان‌کننده تأثیر مثبت این روش بر روی نتایج نهایی پیش‌بینی است. اگر سری تاریخی طولانی‌تر و دارای تنوع مقادیر حدی و غیرحدی باشد آنگاه تأثیر این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

جدول ۳- نتایج اجزا منتخب روش K-NN بدون تأثیر و با تأثیر روش برونابی پیشنهادی

| نوع رگرسیون | کل داده‌ها | | | | |
|-------------------------|---------------|--------------|------|-----|------|
| | بیشترین مقدار | کمترین مقدار | RMSE | %VE | CORR |
| معمولی | ۴۰ | ۰/۵ | ۳۹ | ۱۳ | ۸۹ |
| بالحاظ کردن روش برونابی | ۲۵ | ۰/۵ | ۲۶ | ۸/۵ | ۹۴ |



شکل ۴- نمودار نتایج روش K-NN با ترکیب منتخب بدون اعمال روش برونابی پیشنهادی



شکل ۵- نمودار نتایج روش K-NN با ترکیب منتخب با اعمال روش برونابی پیشنهادی

۴- نتیجه‌گیری

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 ماهانالویس 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 به ترتیب ۴۵، ۵۹ و ۱۷ درصد بهبود عملکرد داشته است که این ارقام ضرورت اعمال راهکارهای ذکر شده را برای استخراج نتایج دقیق‌تر نشان می‌دهد.

روش K-NN به دلیل سادگی و عدم پیچیدگی بالا می‌تواند همواره یکی از بهترین گزینه‌ها برای انجام پیش‌بینی در علوم مختلف به خصوص 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 پیش‌بینی‌کننده 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

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 properprocessing 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 accuracy 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.