ترکیب مدل محاسبات نرم مبتنی بر الگوریتم یادگیری ماشین و تحلیل مولفه-های اصلی جهت پیش‌بینی بارندگی

لاله پرویز[۱]

تاریخ ارسال:۱۳۹۹/۰۴/۲۷
تاریخ پذیرش:۱۳۹۹/۰۹/۲۳

مقاله پژوهشی

چکیده

تاثیر تغییرات بارندگی بر منابع آب، تولیدات کشاورزی نیاز به روش کارآمدی جهت پیش‌بینی بارندگی را آشکار می‌سازد. در این تحقیق یکی از روش‌های محاسبات نرم در راستای پیش‌بینی بارندگی با رویکرد کاهش داده توسعه داده شد. داده‌های ورودی مدل متوسط دمای هوا، دمای نقطه شبنم، متوسط فشار سطح دریا، متوسط فشار ایستگاه، میانگین رطوبت نسبی و میانگین سرعت باد در ایستگاه‌های تبریز، اهر و جلفا بودند. روش مورد استفاده در این تحقیق شامل رگرسیون بردار پشتیبان، Epsilon و Nu ،می‌باشد. در تمام ایستگاه‌های مورد مطالعه استفاده از رگرسیون بردار پشتیبان Nu نسبت به Epsilon منجر به کاهش خطا شد به طوری که مقادیر UII با رگرسیون بردار پشتیبان Nu در ایستگاه‌های تبریز، اهر و جلفا به‌ترتیب ۱۹/۱۹، ۵/۸۸ و۱۵/۷۸ درصد کاهش داشت. نتایج بیانگر محدویت استفاده از رویکرد کاهش داده برای داده‌هایی با فاکتور KMO پایین‌تر از ۰/۵ است که شامل ایستگاه‌های تبریز و اهر بودند. تحلیل مولفه‌های اصلی در هر دو نوع رگرسیون بردار پشتیبان عملکرد مدل را افزایش داد به طوری که در ایستگاه جلفا با بکارگیری تحلیل مولفه‌های اصلی مقادیر dدر رگرسیون بردار پشتیبان Epsilon و Nu ۱۶/۶ و ۱۷/۵ درصد افزایش یافت. اجرای چرخش وریماکس در پیش‌پردازش داده‌های ورودی به رگرسیون نسبت به تحلیل مولفه‌های اصلی قوی‌تر عمل کرد. در این راستا مقادیرRRMSE وRMSE در ایستگاه جلفا با استفاده از رگرسیون بردار پشتیبان Epsilon و با اجرای چرخش به‌ترتیب ۶/۶۶ و ۶/۴۵ درصد کاهش داشت. بنابراین تحلیل مولفه‌های اصلی ابزار مناسبی جهت ارتقاء عملکرد روش‌های محاسبات نرم با رعایت قیود می‌باشد.

واژه‌های کلیدی: بارندگی، تحلیل مولفه‌های اصلی، رگرسیون بردار پشتیبان، فاکتور KMO،

[۱]دانشیار، دانشکده کشاورزی دانشگاه شهید مدنی آذربایجان، تبریز، ایران
تلفن تماس نویسنده اول۰۹۱۴۴۱۴۶۲۴۶آدرس پست الکترونیکی نویسنده اولlaleh_parviz@yahoo.com

مقدمه

بارندگی یکی از اجزاء مهم چرخه هیدرولوژیکی می‌باشد که متغیر در زمان و مکان است. آگاهی در مورد میزان بارندگی در مورد تامین اطلاعاتی در زمینه بهره-برداری از مخزن، مدیریت منابع آب، کشاورزی، هشدار سیل، پیش‌بینی‌های اقلیمی و آب و هوایی کمک شایانی خواهد کرد (Sehad et al., 2017). پیش‌بینی بارندگی به‌علت وابستگی بارندگی به پارامترهای زیادی مانند دما، رطوبت، سرعت باد از پیچیدگی زیادی برخوردار است (Du et al., 2017). در واقع تغییرات زیاد بارندگی در دامنه مقیاس وسیعی از زمان و مکان، پیش‌بینی آن را مشکل ساخته است (Hamidi et al., 2015). بارندگی مولفه مهمی در تصمیم‌گیری محققین است،بنابراین نیاز به برآورد آن با روشی کارآمد می‌باشد (Zhang et al., 2020).

همچنین در سال‌های اخیر تغییرات جهانی زندگی افراد را تحت تاثیر قرار داده است به طوری که گرم شدن جهانی منجر به تغییراتی در دما و بارش منطقه‌ای شده است. بنابراین پیش‌بینی اقلیمی کوتاه مدت یکی از موضوعات مهم در علم هواشناسی است و برآورد بارندگی با روشی کارآمد دارای اهمیت بالایی است (Chen and Zhu 2013).در بسیاری از تحقیقات از عامل دما برای پیش‌بینی بارندگی استفاده شده است ولی جهت پیش‌بینی بارش یک مدل کاملا پیچیده با استفاده از چندین متغیر هواشناسی لازم است (Moon and Kim 2020).

استفاده از روش‌های محاسبات نرم در سال‌های اخیر جهت تخمین و پیش‌بینی بارندگی کاربرد چشمگیری داشته‌است (Shenify et al., 2016). از انواع ماشین بردار پشتیبان در پیش‌بینی بارندگی در هند استفاده شد. داده‌های ورودی این تحقیق شامل دمای هوا، ساعات آفتابی، رطوبت نسبی، سرعت باد بودند. صحت داده‌ها و نتایج کاملا وابسته به نوع روش مورد استفاده بود (Samui et al., 2011). از مدل رگرسیون بردار پشتیبان در پیش‌بینی بارندگی در بنگلادش استفاده شد. در این تحقیق توابع کرنل مختلف و سه نوع بارندگی (بارندگی کل، بیشینه بارندگی، متوسط بارندگی) در نظر گرفته شد. روش مورد استفاده در این تحقیق تا ۹۹/۹۲ درصد پیش-بینی‌ها را انجام داد (Hasan et al., 2015). از ماشین بردار پشتیبان با مقیاس چند زمانه در پیش‌بینی کوتاه مدت بارندگی استفاده شد. نتایج بیانگر بهبود عملکرد ماشین بردار پشتیبان بود (Jiajia et al., 2017). با اعمال تغییراتی در ماشین بردار پشتیبان یعنی بکارگیری الگوریتم بهینه‌سازی ازدحام ذرات (PSO) به پیش‌بینی بارندگی پرداخته شد. داده‌های مورد استفاده شامل فشار اتمسفری، فشار سطح دریا، جهت باد، سرعت باد، دما، رطوبت نسبی و بارندگی ساعتی بودند. عملکرد قوی ماشین بردار پشتیبان در پیش‌بینی بارندگی اثبات شد (Du et al., 2017).

ضابط پیشخانی و همکاران (۱۳۹۵) به مقایسه الگوسازی بارندگی ماهانه با مدل‌های SVM و ANFIS در شهر گنبد کاووس پرداختند. پارامتر موثر بر بارندگی براساس نتایج تحلیل حساسیت، رطوبت نسبی بود. SVM از کارایی بالایی در برابر ANFIS برخوردار بود. مقایسه‌ای بین عملکرد شبکه عصبی مصنوعی، ماشین بردار پشتیبان و درخت تصمیم در برآورد روزهای بارانی کوتاه انجام گرفت. نتایج نشان دادند که روش ماشین بردار پشتیبان دارای خطای کمی در طبقه‌بندی روزهای بارانی (بدون باران، بارندگی کم و متوسط بارندگی) نسبت به سایر روش‌ها بود(Ingsrisawang et al., 2008). عملکرد بهتر ماشین بردار پشتیبان در برآورد بارندگی روزانه در مقایسه با تکنیک‌های شبکه عصبی و درخت تصمیم نیز اثبات شد (Ortiz-Garcı´a et al., 2014). مقایسه‌ای بین عملکرد دو روش ماشین بردار پشتیبان و شبکه عصبی مصنوعی در دو ایستگاه هواشناسی همدان (فرودگاه و نوژه) جهت پیش‌بینی بارندگی انجام گرفت. دوره آماری شامل داده‌های ماهانه از جولای ۱۹۷۶ تا جون ۲۰۰۱ و آپریل ۱۹۶۱ تا نوامبر ۱۹۹۶ به‌عنوان دوره واسنجی ایستگاه-های همدان و نوژه در نظر گرفته شدند. تقسیم‌بندی

داده‌ها به دو قسمت ۷۰ و ۳۰ درصد بود. براساس آماره-های خطا مانند خطای ریشه متوسط مربعات و میانگین خطای مطلق، ماشین بردار پشتیبان دارای عملکرد بهتری در مقایسه با شبکه عصبی مصنوعی بود (Hamidi et al., 2015). در مطالعه‌ای از سه روش ریاضی شبکه عصبی مصنوعی، برنامه‌ریزی ژنتیک و ماشین بردار تصمیم با تبدیل موجک جهت پیش‌بینی بارندگی در صربستان در دوره زمانی ۱۹۶۴-۲۰۱۲ استفاده شد. نتایج براساس برخی از آماره‌های خطا حاکی از عملکرد بهتر روش ماشین بردار تصمیم با تبدیل موجک نسبت به سایر روش‌ها بود (Shenify et al., 2016). مقایسه بین جنگل تصادفی و ماشین بردار پشتیبان درپیش‌بینی بارندگی (استخراجی از رادار) در طی دوره ۲۰۱۲-۲۰۱۵ حاکی از عملکرد بهتر ماشین بردار پشتیبان بود (Yu et al., 2017). از الگوریتم‌های یادگیری ماشین مانند شبکه عصبی مصنوعی و ماشین بردار پشتیبان در پیش‌بینی بارندگی و دمای ماهانه در یونان استفاده شد. نتایج حاکی از عملکرد بهتر الگوریتم‌های یادگیری ماشین نسبت به روش‌های کلاسیک بود (Papacharalampous et al., 2018).

با توجه به کارایی بالای روش‌های محاسبات نرم در برآورد بارندگی، سعی در توسعه روش‌های بیان شده می‌باشد. یکی از جنبه‌های توسعه استفاده از تحلیل مولفه‌های اصلی در کاهش تعداد داده‌های ورودی است. مدل رگرسیون خطی در تعیین غلظت ذرات معلق با قطر کمتر از ۱۰ میکرون (PM_{10}) در مالزی با استفاده از تحلیل مولفه‌های اصلی توسعه یافت. تعداد متغیرهای ورودی برابر با هفت در نظر گرفته شد که داده‌های هواشناسی شامل متوسط دما، رطوبت نسبی و سرعت باد بودند. نتایج نشان دادند که عملکرد مدل رگرسیون خطی در تعیین PM10 با استفاده از تحلیل مولفه‌های اصلی به دلیل کاهش پیچیدگی داده‌های ورودی ارتقاء یافت (UlSaufie et al., 2011). از شبکه عصبی مصنوعی و رگرسیون خطی چندگانه با استفاده از تحلیل مولفه‌های اصلی در برآورد بارندگی در سه منطقه در جنوب آمریکا استفاده شد. تعداد مولفه‌های در نظر گرفته شده ده بود. براساس تحلیل مولفه‌های اصلی تعداد کمی از مدل‌های آب و هوایی در زمستان برای در نظر گرفتن واریانس بارش لازم است. مناطق انتخاب شده مناطقی هستند که به شدت تحت تاثیر تغییر اقلیم هستند. عملکرد بهتر شبکه عصبی نسبت به رگرسیون خطی چندگانه مشاهده شد (Santos et al., 2016).

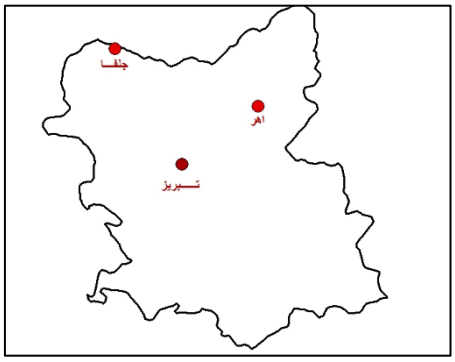
هدف تحقیق توسعه یکی از انواع روش‌های محاسبات نرم با رویکرد کاهش داده برای پیش‌بینی بارندگی است. بدین منظور عملکرد رگرسیون بردار پشتیبان Epsilon و Nu مورد بررسی قرار گرفتند. بررسی تاثیر اعمال چرخش بر کارایی تحلیل مولفه‌های اصلی و در نهایت بر عملکرد دو نوع رگرسیون بردار پشتیبان نیز مورد مطالعه قرار گرفت.

مواد و روش ها

منطقه مورد مطالعه

ایستگاه‌های مورد مطالعه تحقیق در استان آذربایجان شرقی (اهر، تبریز، جلفا) واقع هستند که موقعیت مکانی ایستگاه‌ها در شکل۱ آورده شده است. طول دوره آماری ایستگاه‌های مطالعاتی ۱۹۸۷ تا ۲۰۱۷ بوده است که سال‌های ۱۹۸۷ تا ۲۰۰۸ جهت واسنجی مدل مورد استفاده قرار گرفتند. متغیرهای ورودی مورد استفاده شامل متوسط دمای هوا، دمای نقطه شبنم، متوسط فشار سطح دریا، متوسط فشار ایستگاه، میانگین رطوبت نسبی و میانگین سرعت باد بودند.براساس سیستم اقلیم‌نمای دومارتن اقلیم ایستگاه‌های مطالعاتی نیمه‌خشک می‌باشد.پایش بارندگی در استان آذربجان شرقی به عنوان یکی از پتاسیل‌های تولید کشاورزی کشور، از اهمیت چشمگیری در زمینه عملکرد محصول برخوردار است.

شکل (۱): موقعیت ایستگاه‌های مطالعاتی



شکل (۱): موقعیت ایستگاه‌های مطالعاتی

همچنین در سال‌های اخیر با توجه به گرمایش جهانی و تاثیر آن بر نوسانات سطح آب دریاچه ارومیه، بررسی و پیش‌بینی بارندگی در خود استان و استان‌های همجوار دارای اهمیت زیادی است.

در این تحقیق تاثیر رویکرد کاهش داده (تحلیل مولفه‌های اصلی) در مورد داده‌های ورودی رگرسیون بردار ماشین مورد بررسی قرار گرفت.

تحلیل مولفه های اصلی

در این روش p متغیر اصلی همبسته به p مولفه غیرهمبسته (متعامد) تبدیل می‌شود. با استفاده از تحلیل مولفه‌های اصلی متغیرهای اصلی ورودی به متغیرهای جدید (بدون همبستگی می‌باشند) تبدیل می‌شوند. ترکیب خطی از متغیرهای ورودی یا اصلی، مولفه‌های جدید را تشکیل می‌دهد. اگر در p متغیر اصلی تنها مقداری همبستگی معنی‌دار وجود داشته باشد، امکان استفاده از این روش وجود دارد. پس لازم است در ابتدا قبل از انجام تحلیل مولفه‌ها، مناسب بودن داده‌ها برای انجام تحلیل کنترل شود. یکی از روش‌های کنترل محاسبه فاکتور KMO[1] است که بصورت رابطه ۱ تعریف می‌شود.

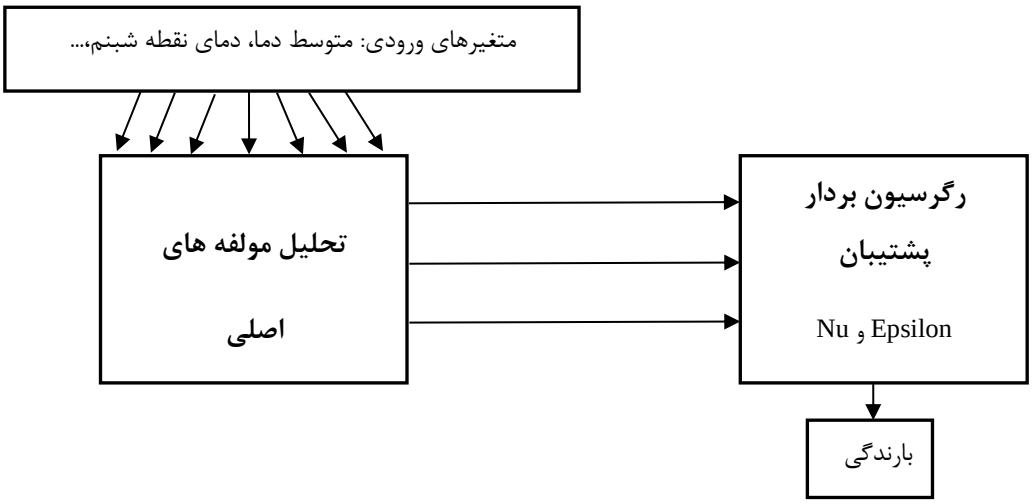
[1]Kaiser-Meyer-Olkin

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \quad i \neq j \qquad (۱)$$

r_{ij}: ضریب همبستگی بین متغیرهای j و i، a_{ij}: ضریب همبستگی جزیی بین متغیرهای j و i، p: تعداد متغیرها.

در صورتی که فاکتور KMO بزرگتر از ۰/۵ باشد، نشان دهنده امکان استفاده از تحلیل مولفه‌های اصلی بر روی داده‌ها است (شیخ الاسلامی و همکاران، ۱۳۹۳). زمانی که مساله با حجم اطلاعات زیادی روبه رو است، کاربرد تحلیل مولفه‌های اصلی مفید می‌باشد. در این تحلیل اطلاعاتی از داده‌های اولیه از بین نمی‌رود چرا که در تشکیل مولفه‌ها از اطلاعات تمام متغیرها استفاده می‌شود. مساله دیگری که صحت تحلیل مولفه‌های اصلی را افزایش می‌دهد، اجرای چرخش مناسب بر روی ماتریس ضرایب مولفه‌ها است. جهت تفسیر ساده مولفه‌ها از چرخش مولفه‌ها استفاده می‌شود. چرخش مولفه‌ها به دو دسته عمودی و مایل تقسیم می‌شود. در چرخش عمودی استقلال بین مولفه‌ها حفظ می‌شود و به همین دلیل این نوع چرخش بیشتر مورد استقبال است (سیفی و همکاران، ۱۳۸۹).

در این تحقیق سعی در ترکیب تحلیل مولفه‌های اصلی با رگرسیون بردار پشتیبان است به اینصورت که با استفاده از تحلیل مولفه‌های اصلی، متغیرهای ورودی به چندین مولفه تبدیل می‌شوند و سپس مولفه‌ها وارد رگرسیون بردار پشتیبان می‌شوند. شکل ۲ مراحل انجام عملیات را نشان می‌دهد.



شکل(۲): طراحی ترکیب تحلیل مولفه‌های اصلی و رگرسیون بردار پشتیبان

ماشین بردار پشتیبان

جهت در نظر گرفتن ارتباط غیرخطی پیچیده بین پارامترهای هواشناسی در زمان و مکان خاص با توجه به پیشرفت‌هایی در زمینه تکنولوژی کامپیوتری و روش‌های هوشمند، مهارت‌هایی در زمینه ماشین هوشمند نیز توسعه داده شد. ماشین بردار پشتیبان به طور گسترده در زمینه مسایل یادگیری ماشین استفاده شد. مفاهیم اولیه ماشین بردار پشتیبان در ادامه توضیح داده می‌شود.

اگر دسته نمونه‌ها $(x_i, y_i)_{1\le i\le N}$ به عنوان نمونه-های آموزشی در نظر گرفته شود N تعداد کل بردارها است. بردارهای آموزشی $x_i \in R^D$ بردارهای ورودی هستند و D ابعاد فضای ورودی است. y_i نمایانگر خروجی متناظر x_i است. جداسازی ابرصفحه بصورت

$w.x+b=0$ است. w بردار وزن و b نمایانگر اریبی است. معادله ۲ نشان دهنده ابرصفحه طبقه‌بندی شده است.

$$y_i(w.x_i+b)\geq 1 \quad (۲)$$

به منظر بیشینه‌سازی فاصله‌ها، نیاز به بیشینه‌سازی $\|w\|^{-1}$ است که معادل کمینه‌سازی $\frac{1}{2}\|w\|^2$ می باشد. نوع پایه‌ای ماشین بردار پشتیبان در معادلات۳و ۴آورده شده است.

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad (۳)$$

$$y_i(w.x_i+b)-1\geq 0 \quad (۴)$$

با توجه به معادله ۴ می‌توان ابرصفحه $w.x_i+b=0$ را با بیشترین حاشیه بدست آورد. جهت برابری قیود در معادله ۴، آنها می‌توانند با بکارگیری لاگرانژین بدون محدودیت شوند که در این حالت حاشیه نرم معرفی می‌شود.

$$\min_{w,b} \frac{1}{2}\|w\|^2 - C\sum_{i=1}^{N}\ell_0/1[y_i(w.x_i+b)-1] \quad (۵)$$

$\ell_0/1$: تابع هزینه،C: ثابتی بزرگتر از صفر است.

در ادامه محاسبات نیاز به این است که در تمام نمونه‌ها قیود در نظر گرفته شود. در این حالت مساله دوگان تابع هدف بصورت مساله برنامه‌ریزی درجه دوم (بیشینه‌سازی) است که بصورت معادله ۶ نشان داده می‌شود.

$$\max_{\alpha}\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_iy_j\alpha_i\alpha_j(x_i.x_j) \quad (۶)$$

$$\sum_{i=1}^{N}\alpha_i y_i=0 \qquad (۷)$$
$$0\leq\alpha_i\leq C, i=1,...,N$$

این یک مساله بحرانی از تابع درجه دوم با قیود نابرابر است که دارای راه حل واحد است. α_i ضرب لاگرانژین برای هر نمونه آموزشی است. به هنگام رویارویی با مسایل بردار پشتیبان غیرخطی، ماشین بردار پشتیبان یک تابع کرنل معرفی می‌کند تا داده‌ها به فضای ابعادی بالا انتقال پیدا کنند.

$$\max_{\alpha}\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_iy_j\alpha_i\alpha_jK(x_i.x_j) \quad (۸)$$

$K(x_i.x_j)$ تابع کرنل.

در این حالت قیود مشابه معادله ۷ می‌باشند. ماشین بردار پشتیبان براساس مطالعه تابع کرنل استوار است. بنابراین تابع کرنل دارای اهمیت زیادی است و انتخاب تابع کرنل در تعمیم توانایی مدل نقش مستقیمی دارد. یکی از معروفترین تابع کرنل، تابع کرنل گاوسی یا پایه شعاعی(RBF) است که در مسایل غیرخطی بصورت رابطه ۹ تعریف می شود.

$$K(x_i.x_j)=\exp\left\{-g\left\|x_j-x_i\right\|^2\right\} \quad (۹)$$

g: پارامتر کرنل جهت اندازه‌گیری عرض تابع کرنل.

براساس تحلیل انجام گرفته، ایده ماشین بردار پشتیبان این است که وقتی نمونه‌های غیرخطی طبقه-بندی می‌شوند، نمونه‌های فضای اصلی با یک نگاشت غیرخطی به فضای با ابعادبالا برده می‌شوند(Du et al., 2017).

آماره‌های ارزیابی عملکرد مدل

جهت بررسی عملکرد مدل رگرسیون بردار پشتیبان با ترکیب تحلیل مولفه‌ها از برخی آماره‌ها استفاده شد که رابطه ریاضی آماره‌ها در ادامه آورده شده است.

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i-F_i)^2} \quad (۱۰)$$

Root Mean Square Error

$$RRMSE=\frac{RMSE}{\overline{O}} \quad (۱۱)$$

Re*lative Mean Square Error*

$$GMER=EXP\left[\frac{1}{N}\sum_{i=1}^{N}\ln(\frac{F_i}{O_i})\right] \quad (۱۲)$$

Geometric Mean Error Ratio

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{O_i-F_i}{O_i}\right|\times 100 \quad (۱۳)$$

Mean Absolute Percentage Error

$$UI = \frac{\left[\sum_{i=1}^{N}(O_i - F_i)^2\right]^{05}}{\left[\sum_{i=1}^{N}O_i^2\right]^{0.5} + \left[\sum_{i=1}^{N}F_i^2\right]^{0.5}} \quad (۱۴)$$

$$UII = \frac{\left[\sum_{i=1}^{N}(O_i - F_i)^2\right]^{05}}{\left[\sum_{i=1}^{N}O_i^2\right]^{0.5}} \quad (۱۵)$$

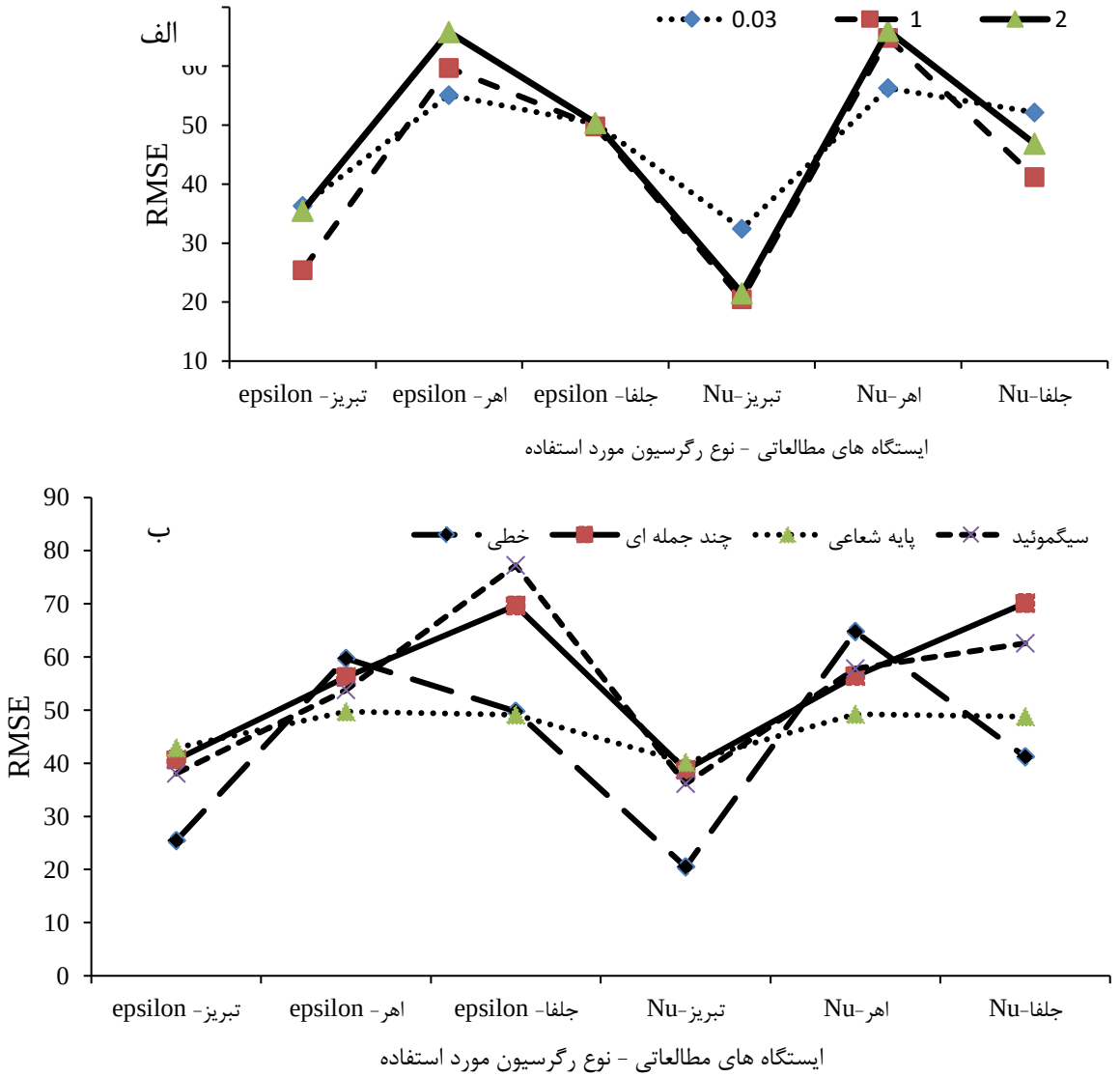
$$d = 1 - \frac{\sum_{i=1}^{N}|Fi - Oi|}{\sum_{i=1}^{N}\left(|Fi - \overline{O}| + |Oi - \overline{O}|\right)} \quad (۱۶)$$

مقادیر کمینه RMSE، MAPE، RRMSEنشان دهنده عملکرد بهتر مدل است. d: ضریب همسانی اصلاح شده که بین صفر و یک است. حالت بهینه آماره نزدیکی به یک است. نسبت میانگین هندسی خطا (GMER) بزرگتر از یک بیانگر بیش برآورد و کوچکتر از یک نمایانگر کم برآورد است. UIو UII به ترتیب جهت ارزیابی دقت و کیفیت پیش‌بینی بکار می‌روند. عملکرد بهتر مدل همراه با نزدیکی UI و UII به صفر همراه است (Pannkkong et al., 2016; Zaynoddin et al., 2018).

نتایج

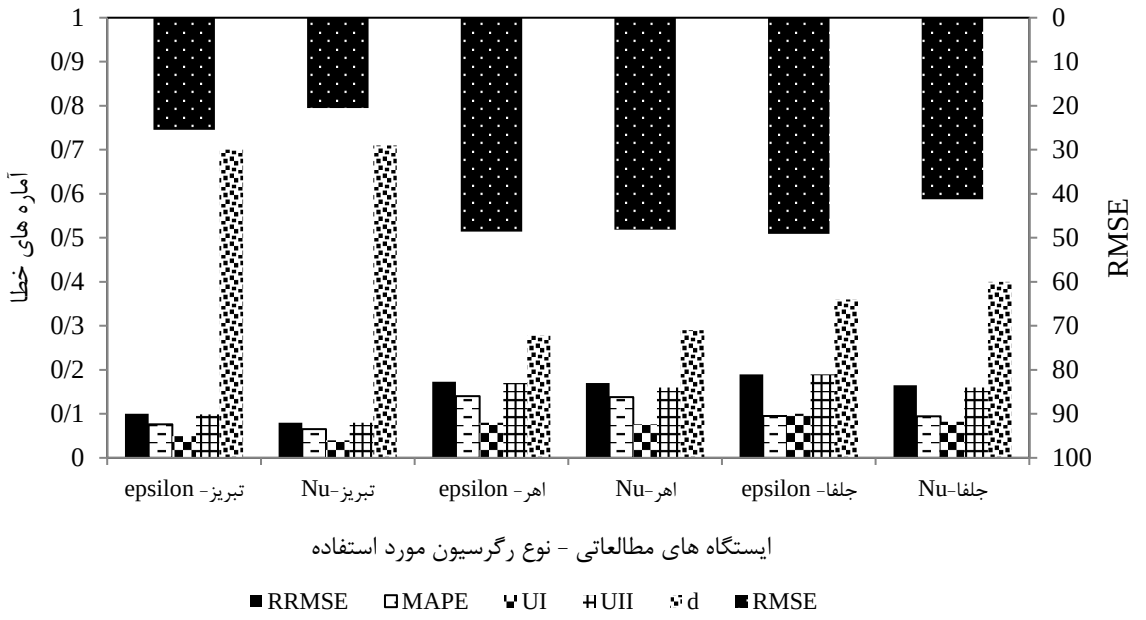
یکی از انواع روش‌های محاسبات نرم، رگرسیون بردار پشتیبان از نوع Epsilon و Nu است که در این تحقیق جهت پیش‌بینی بارندگی با تحلیل مولفه‌های اصلی توسعه داده شد. بدین منظور داده‌های ایستگاه-های اهر، تبریز و جلفا به دو دسته واسنجی و صحت-سنجی تقسیم شدند (۷۰٪ و ۳۰٪). دوره‌های صحت-سنجی و واسنجی به ترتیب شامل ۱۹۸۷-۲۰۰۸ و ۲۰۰۹-۲۰۱۷ بودند. متوسط دمای هوا، دمای نقطه شبنم، متوسط فشار سطح دریا، متوسط فشار ایستگاه، میانگین رطوبت نسبی و میانگین سرعت باد به عنوان داده‌های ورودی مدل تعریف شدند.

در ابتدا از داده‌های موجود جهت اجرای رگرسیون بردار پشتیبان از نوع Epsilon و Nu استفاده شد. در این میان تحلیل حساسیت یکی از گام‌های مهم در مدلسازی است که در دو نوع رگرسیون مورد استفاده تحقیق،این مساله مدنظر قرار گرفت. تحلیل حساسیت رگرسیون بردار پشتیبان در مورد تابع کرنل و پارامتر تنظیم کننده مدل (C) تعریف شد که شکل ۳ نتایج تحلیل حساسیت را نشان می‌دهد.



شکل(۳): تحلیل حساسیت رگرسیون بردار پشتیبان الف) پارامتر تنظیم کننده مدل ب) تابع کرنل

شکل ۲-الف نمایانگر تحلیل حساسیت پارامتر تنظیم کننده رگرسیون بردار پشتیبان Epsilon و Nuبرای تابع کرنل خطی است. در هر دو نوع رگرسیون کمینه خطا در ایستگاه‌های جلفا و تبریز مربوط به C=1 است ولی در ایستگاه اهر کمینه خطا مربوط به C=0.03 می‌باشد. شکل ۳-ب تحلیل حساسیت تابع کرنل رگرسیون بردار پشتیبان Epsilon و Nuبرای C=1 می‌باشد. توابع کرنل مورد استفاده در این تحقیق خطی، چند جمله ای، سیگموئید و پایه شعاعی هستند. در ایستگاه تبریز در هر دو نوع رگرسیون کمینه خطا مربوط به تابع خطی است. در اهر در هر دو نوع رگرسیون کمینه خطا مربوط به تابع پایه شعاعی است. در جلفا در رگرسیون بردار پشتیبان Epsilon و Nu به ترتیب کمینه خطا مربوط به توابع پایه شعاعی و خطی می‌باشد. بعد از انجام تحلیل حساسیت رگرسیون بردار پشتیبان اقدام به پیش‌بینی بارش با هر دو نوع رگرسیون شد که نتایج باآماره‌های خطا در شکل ۴ نشان داده شده است.

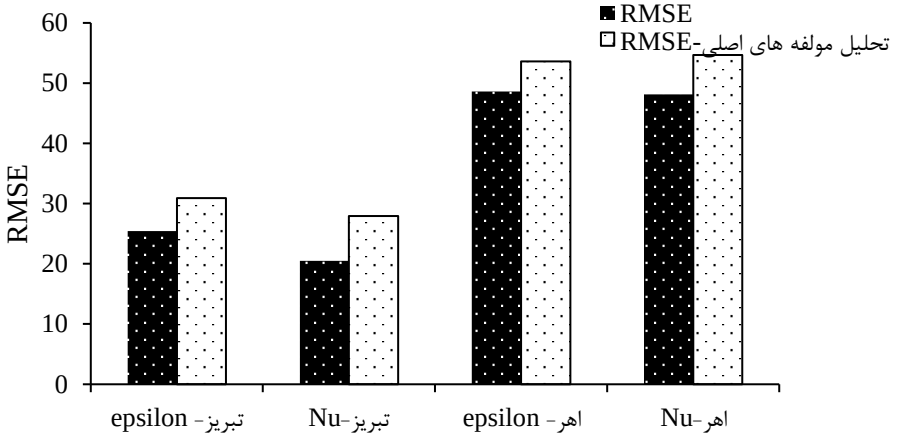


شکل(۴): مقادیر آماره‌های حاصل از پیش‌بینی بارندگی با رگرسیون بردار پشتیبان Epsilon و Nu

با توجه به شکل ۴ جهت حرکت آماره‌های خطا به سمت رگرسیون بردار پشتیبان Nu به حالت بهینه نزدیک شده است، به عنوان نمونه در ایستگاه تبریز میزان کاهش آماره‌های RMSE، RRMSE، MAPE، UI، UII از رگرسیون بردار پشتیبان Epsilon به Nu ۱۹/۴۵، ۲۰، ۱۴/۴۷، ۱۸/۳۶، ۱۹/۱۹ درصد بوده است. به‌طور متوسط در سه ایستگاه مطالعاتی میزان کاهش آماره‌های RMSE، RRMSE، MAPE، UI، UIIو میزان افزایش dاز رگرسیون بردار پشتیبان Epsilon به Nu۱۰/۷۷، ۱۰/۳۶، ۴/۵، ۱۳/۵۳، ۱۲/۸۵، ۴/۶۳ درصد بوده است. مقادیر نسبت میانگین هندسی خطا کمتر از یک می‌باشد که بیانگر کم برآورد است. عملکرد بهینه رگرسیون بردار پشتیبان در مسایل مربوط به پیش‌بینی بارندگی توسط محققینی به اثبات رسیده است(Ingsrisawang et al., 2008; Hamidi et al., 2015; Papacharalampous et al., 2018). ماشین بردار پشتیبان براساس کمینه‌سازی ریسک ساختاری عمل می‌کند در حالی که عملکرد شبکه‌های عصبی مصنوعی براساس کمینه‌سازی ریسک تجربی است، بنابراین ماشین بردار پشتیبان نسبت به شبکه عصبی مصنوعی از عملکرد قوی برخوردار است. استفاده از کمینه‌سازی ریسک تجربی منجر به بیش برازش در مسایل می‌شود که مساله در کمینه موضعی قرار می‌گیرد (Hamidi et al., 2015). با توجه به عملکرد دقیق رگرسیون بردار پشتیبان، نوع پارامترهای دخیل در رگرسیون از اهمیت چشمگیری برخوردار هستند، به طوری که تاثیر دو پارامتر Epsilon و Nu در نتایج پیش‌بینی بارندگی مشاهده شد چرا که پارامترهای بیان شده تاثیر خود را در طبقه‌بندی، جداسازی کلاس‌ها و کنترل تعداد بردارهای پشتیبان در رگرسیون نشان می‌دهند. تفاوت بین Epsilon و Nu در چگونگی پارامتره کردن مساله آموزش است. Nu بر تعداد بردارهای پشتیبانی کنترل دارد.

ابزار مورد استفاده جهت توسعه رگرسیون بردار پشتیبان در این تحقیق تحلیل مولفه‌های اصلی می‌باشد. پردازش اولیه در زمینه تحلیل مولفه‌های اصلی، محاسبه فاکتور KMO است که مقادیر این فاکتور برای ایستگاه‌های اهر، تبریز و جلفا به‌ترتیب ۰/۳، ۰/۴۱،

۰/۵۵ برآورد شد. مقادیر KMO در ایستگاه‌های تبریز و اهر به‌علت پایین‌بودن از ۰/۵ نشان‌دهنده مناسب نبودن داده‌های این ایستگاه‌ها برای انجام تحلیل مولفه-ها است. برای بررسی تاثیر این فاکتور در نتایج، هر دو نوع رگرسیون با تحلیل مولفه‌ها نیز اجرا شدند که نتایج مقادیر خطا در شکل ۵ آورده شده است. در ایستگاه تبریز و اهر تعداد مولفه‌های ایجاد شده سه بود.



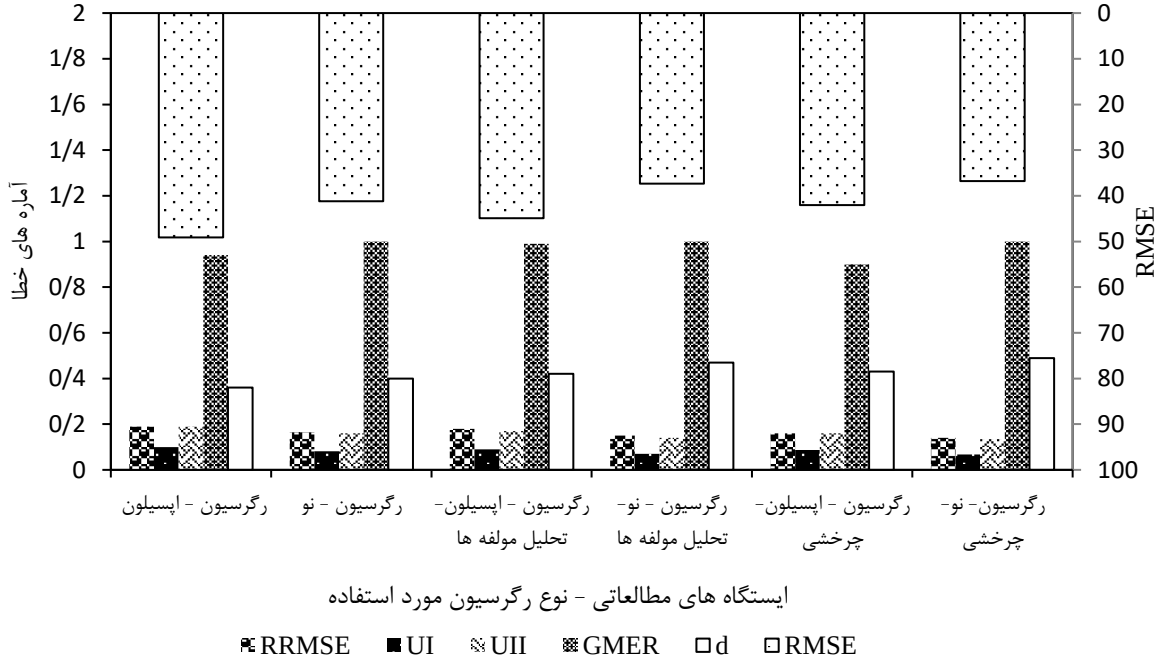
شکل(۵): مقادیر RMSE در ایستگاه هایی با KMO کمتر از ۰/۵

با توجه به شکل ۵ مشخص است که مقادیر خطا با استفاده از تحلیل مولفه‌ها در ایستگاه‌های تبریز و اهر نه تنها بهبود نیافته است بلکه میزان خطا با اعمال تحلیل مولفه‌ها افزایش نیز یافته است. بنابراین کارایی فاکتور KMO در تشخیص مناسب بودن داده‌ها برای تحلیل مولفه‌های اصلی کاملا آشکار می‌شود. در ادامه به دلیل مناسب بودن فاکتور KMO در ایستگاه جلفا، تحلیل مولفه‌های اصلی و چرخشی در این ایستگاه اجرا شد که نتایج ضرایب مولفه‌ها در جدول ۱ آورده شده است. در این حالت علاوه بر تحلیل مولفه‌های اصلی از اجرای چرخش بر روی ماتریس ضرایب مولفه‌ها نیز استفاده شد. در این تحقیق چرخش وریماکس که یکی از انواع چرخش‌های عمودی است بکار گرفته شد.

جدول(۱): بردارهای ویژه داده‌های هواشناسی در ایستگاه جلفا

داده‌های هواشناسی		متوسط دمای هوا	دمای نقطه شبنم	متوسط فشار سطح دریا	متوسط فشار ایستگاه	میانگین رطوبت نسبی	میانگین سرعت باد
تحلیل مولفه‌های اصلی	مولفه اول	-۰/۹۳	-۰/۶	۰/۹۷	۰/۸۹	۰/۷	-۰/۰۶
	مولفه دوم	۰/۰۲۴	۰/۶۳	۰/۰۵۵	۰/۰۷۴	۰/۴۶	۰/۷۲
	مولفه سوم	۰/۲	-۰/۴۲	۰/۱	۰/۲۷	-۰/۵۱	۰/۶۶
چرخشی	مولفه اول	-۰/۸۲	-۰/۰۲۳	۰/۷۵	۰/۶۲	۰/۹۸	-۰/۰۳۱
	مولفه دوم	۰/۴۶	۰/۹۵	-۰/۶۲	-۰/۶۵	۰/۱۲	۰/۰۵۷
	مولفه سوم	۰/۱۶	۰/۱۵	۰/۱۰۸	۰/۲۴	-۰/۰۳	۰/۹۸

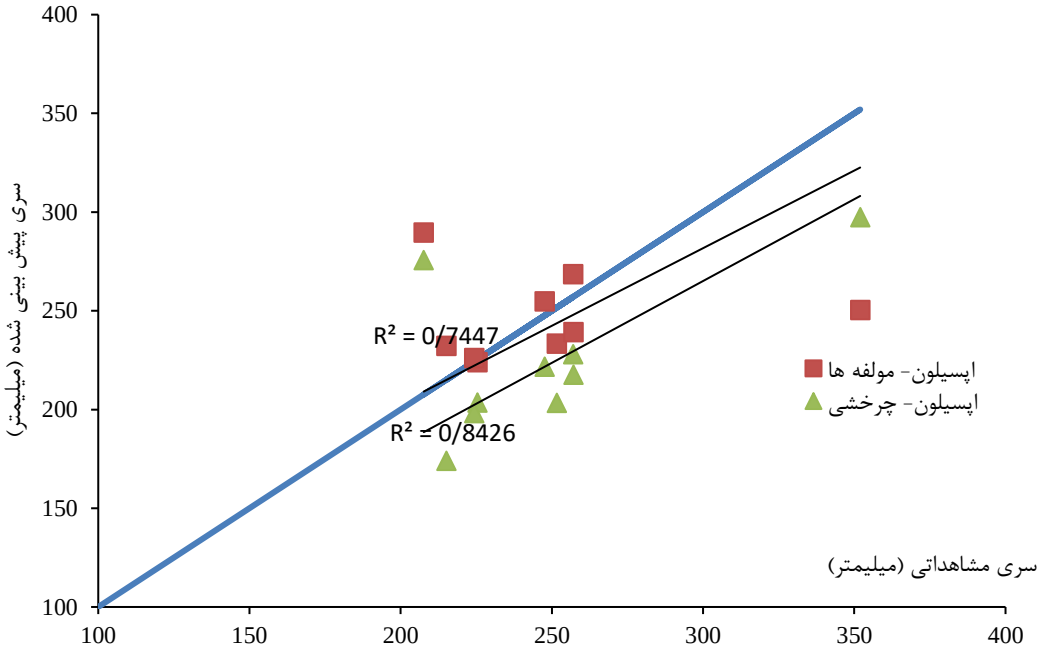
قابل ذکر است که در هر دو تحلیل، سه مولفه انتخاب شده‌اند چرا که سه مولفه اول دارای مقادیر ویژه بزرگتر از یک بوده‌اند و سه مولفه در مجموع ۹۴/۲۴درصد از پراکندگی داده‌های اصلی را بیان کردند. با افزایش در تعداد مولفه‌ها مقادیر ویژه کاهش پیدا کرده‌اند. مقادیر ویژه مولفه اول، دوم و سوم به‌ترتیب ۳/۴۹، ۱/۱۴، ۱/۰۱ بود. در تحلیل مولفه‌های اصلی در مولفه اول تاثیرگذارترین داده‌ها متوسط فشار سطح دریا، متوسط دمای هوا، متوسط فشار ایستگاه، در مولفه دوم میانگین سرعت باد و دمای نقطه شبنم، در مولفه سوم میانگین سرعت باد، میانگین رطوبت نسبی و دمای نقطه شبنم بوده است. در حالت چرخشی در مولفه اول تاثیرگذارترین داده‌ها میانگین رطوبت نسبی، متوسط دمای هوا، در مولفه دوم دمای نقطه شبنم، متوسط فشار ایستگاه، در مولفه سوم میانگین سرعت باد بوده است. پس تحلیل‌ها نشان می‌دهند که بیشینه ضریب تابع دما نیست و داده‌های دیگری هم در تخمین بارندگی موثر هستند که در نظرگرفتن داده‌های کامل هواشناسی در برآورد بارندگی از نتایج تحقیقات Moon and Kim (2020) نیز بوده است. در نهایت عملکرد تحلیل مولفه-های اصلی و چرخشی با هر دو نوع رگرسیون Epsilon و Nuمورد بررسی قرار گرفت که نتایج در شکل ۶ آورده شده است.



شکل(۶): مقادیر آماره‌های حاصل از پیش‌بینی بارندگی با رگرسیون بردار پشتیبان Epsilon و Nu در ایستگاه جلفا

با توجه به شکل ۶ میزان اماره‌ها از رگرسیون بردار پشتیبان Epsilonبه Nu عملکرد بهتری پیدا می‌کنند. در تحقیق C̆esák and Langhammer(2016)کارایی عملکرد رگرسیون بردار پشتیبان Nu در رابطه با سری‌های زمانی هیدرولوژیکی اثبات شد.انجام تحلیل مولفه‌های اصلی منجر به بهبود نتایج می‌شود. از رگرسیون بردار پشتیبان Epsilon به رگرسیون بردار پشتیبان-Epsilon - تحلیل مولفه‌ها میزان RMSE، RRMSE، UI، UII ۸/۵۵، ۵/۲۶، ۱۰، ۱۰/۵۲ کاهش و میزان ۱۶/۶۶d درصد افزایش یافت. از رگرسیون بردار پشتیبان Nu به رگرسیون بردار پشتیبان- Nu- تحلیل مولفه‌ها میزان RMSE، RRMSE، UI، UII ۹/۴۸، ۹/۰۹، ۱۴/۶۳، ۱۲/۵ درصد کاهش و میزان ۱۷/۵d درصد افزایش یافت. از رگرسیون بردار پشتیبان-Epsilon- تحلیل مولفه‌ها به رگرسیون بردار پشتیبان- Epsilon- چرخشی میزان

RMSE، RRMSE، UI، UII ۶/۴۵، ۳/۱۱،۳۳/۱۱، ۵/۸۸ کاهش و میزان ۲/۳۸d درصد افزایش یافت. از رگرسیون بردار پشتیبان- Nu- تحلیل مولفه‌ها به رگرسیون بردار پشتیبان-Nu - چرخشی میزان RMSE، RRMSE، UI، UII ۱/۳۶، ۲/۶،۸۵/۶۶، ۳/۵۷ درصدکاهش و میزان d ۴/۲۵ درصد افزایش یافت. مقادیر GMER در بیشتر مواقع از یک کم بوده است که نمایانگر کم‌برآورد می‌باشد. استفاده از تحلیل مولفه-ها منجر به کاهش خطا و افزایش ضریب همسانی اصلاح شده است. بهبود عملکرد مدل‌ها با استفاده از تحلیل مولفه‌های اصلی در تحقیقات Santos et al., (2016) و UlSaufie et al., (2011) نیز مشاهده شد. از آنجایی که داده‌های اقلیمی معمولا به همدیگر وابسته هستند و همخطی شدیدی بین آنها وجود دارد،بکارگیری این نوع سری‌های زمانی در مدل‌های رگرسیونی منجر به خطای زیادی خواهد شد. برای کاهش تعداد داده‌های ورودی و استفاده درست از آنها و کاهش خطا بهتر است تعداد داده‌های ورودی کم شده و با روش ریاضی مناسبی تعدادی سری زمانی جدید تولید شود تا درصد بالایی از واریانس بین داده‌ها را توجیه کند. تحلیل مولفه‌های اصلی از جمله این گزینه‌ها است (ناظم السادات و شیروانی،۱۳۸۴). در حالت دیگر مقایسه‌ای بین داده‌های مشاهداتی و پیش‌بینی شده با رگرسیون Epsilon انجام گرفت که نتایج در شکل ۷ آورده شده است.



شکل(۷): پراکنش بارندگی مشاهداتی و پیش بینی شده

با توجه به شکل ۷ که پراکنش بارندگی مشاهداتی و پیش‌بینی‌شده را با تحلیل مولفه‌ها و چرخشی نشان می‌دهد حاکی از افزایش ضریب همبستگی خط برازشی بر نمودار داده‌های پیش‌بینی‌شده و مشاهداتی با اعمال چرخش است. اگر مقایسه‌ای بین میانگین بارندگی پیش‌بینی شده و مشاهداتی در دوره صحت‌سنجی انجام گیرد به اینصورت که اختلاف مقادیر پیش‌بینی شده در حالت‌های مختلف با مشاهداتی مورد مقایسه قرار گیرد کمترین اختلاف مربوط به رگرسیون بردار پشتیبان-Nu–چرخشی با مقدار مشاهداتی است (میزان ۱/۱۳ میلیمتر).بنابراین در چندین گام میزان پیش‌بینی‌ها به مقادیر واقعی نزدیک شدند ۱- استفاده از رگرسیون

بردار پشتیبانNu ۲- انجام تحلیل مولفه‌های اصلی بر روی داده‌های ورودی ۳- اجرای چرخش بر روی ضرایب.

نتیجه گیری

بارندگی به‌عنوان یکی از مولفه‌های موثر در کشاورزی، مدیریت منابع آب است که نیاز به برآورد آن با روشی دقیق ضروری می‌باشد. در این تحقیق با بکارگیری رگرسیون بردار پشتیبان Epsilon وNu با رویکرد کاهش داده پیش‌بینی بارندگی توسعه داده شد. از خصوصیات مهم ماشین بردار پشتیبان توانایی در مدل کردن فرآیندهای غیر خطی بدون اطلاع در مود توزیع آماری کلاس‌ها است. خصوصیت مهم دیگر ماشین بردار پشتیبان عملکرد خوب آن در مورد داده-هایی با ابعاد بالا و دسته کوچکی از الگوی آموزشی می-باشد. عملکرد رگرسیون بردار پشتیبان Nu نسبت به Epsilon براساس آماره‌های خطا بهتر بود. این مساله تاثیر پارامتر را در عملکرد رگرسیون نشان می‌دهد چرا که با پارامترهای بیان شده مکان قرارگیری بردارهای پشتیبان اطراف ابر صفحه تنظیم می‌شود. پارامتر Nu برای تعیین نسبت تعداد بردارهای پشتیبانی مورد نظر در حل مساله با توجه به تعداد کل نمونه‌های موجود در مجموعه داده‌ها استفاده می‌شود. با این حال در Epsilon هیچ کنترلی در مورد تعداد بردارهای داده از دسته داده که بردار پشتیبان نامیده می‌شود نیست چرا که می‌تواند کم و زیاد باشد. فقط کنترل در مورد میزان خطایی است که در مدل در نظر گرفته می‌شود.

همچنین رگرسیون بردار پشتیبان Nu برای شبیه‌سازی نمونه آموزشی کوچک قابل اعتماد است. رگرسیون بردار پشتیبان Nu تفسیر معنی‌داری دارد. دلیل این مساله این است که Nu نشان دهنده یک حد بالا در کسری از نمونه‌های آموزشی است که بد پیش بینی شده‌اند و یک حد پایین در کسری از نمونه‌هایی است که بردار پشتیبان هستند. فاکتور KMO فاکتور موثر در تعیین اجرای تحلیل مولفه‌ها بر روی داده‌ها می‌باشد. مقادیر بالای فاکتور KMO می‌تواند ناشی ازکم بودن ضریب همبستگی جزئی بین داده‌ها باشد. تحلیل مولفه‌ها و اجرای چرخش بر روی ضرایب مولفه‌ها نیز منجر به بهبود مقادیر پیش‌بینی شده شد که سهم اجرای چرخش نسبت به تحلیل مولفه‌ها بیشتر بود. دلیل بهبود عملکرد مدل با استفاده از تحلیل مولفه‌های اصلی کاهش تعداد داده‌های ورودی و در نهایت کاهش پیچیدگی حاکم بر مدل می‌باشد.اساس تحلیل مولفه-های اصلی براساس واریانس داده‌ها می‌باشد به طوری‌که اولین مولفه بیشترین واریانس داده‌ها را شامل شود. در واقع با استفاده از تحلیل مولفه‌های اصلی چیدمان داده-های ورودی به مدل براساس قاعده و قانون معین و آماری انجام می‌گیرد. در حالت چرخش با استفاده از روش واریماکس ستون‌های ماتریس عاملی ساده می-شوند که دسترسی به ماتریسی ساده را تسهیل می‌کند. بنابراین ترکیب تحلیل مولفه‌های اصلی و رگرسیون بردار پشتیبان می‌تواند راه‌حل مناسبی در پیش‌بینی بارندگی و تصمیمات مربوط به برنامه‌ریزی منابع آب باشد.

منابع

سیفی، ا.، میرلطیفی، س.م.، و ح. ریاحی. ۱۳۸۹. توسعه مدل ترکیبی رگرسیون چندگانه - تحلیل مولفه‌ها و عامل‌های اصلی (MLR-PCA) در پیش‌بینی تبخیر - تعرق مرجع. نشریه آب و خاک،سال۲۴ بیست و چهارم، شماره ۶، ص ۱۱۹-۱۱۸۶.

شیخ الاسلامی، ن.، قهرمان، ب.، مساعدی، ا.، داوری، ک.، و م. مهاجرپور. ۱۳۹۳. پیش‌بینی تبخیر و تعرق گیاه مرجع (ET_0) با استفاده از روش آنالیز مولفه‌های اصلی (PCA) و توسعه مدل رگرسیون خطی چندگانه (MLR-PCA) (مطالعه موردی: ایستگاه مشهد). نشریه آب و خاک (علوم و صنایع کشاورزی)، سال بیست و هشتم، شماره ۲، ص۴۲۹-۴۲۰.

ضابط پیشخانی، ن.، سیدیان، م.، حشمت پور، ع.، و ح. روحانی. ۱۳۹۵. مقایسه الگو سازی بارندگی ماهانه با مدل های SVMو ANFIS (مطالعه موردی: شهر گنید کاووس). نشریه آب و خاک (علوم و صنایع کشاورزی)، سال سی، شماره ۱، ص ۲۴۶-۲۳۶.

ناظم السادات، ج.، و ا. شیروانی. ۱۳۸۴. پیش بینی دمای سطح آب خلیج فارس با استفاده از رگرسیون چندگانه و تحلیل مولفه های اصلی. علوم و فنون کشاورزی و منابع طبیعی، سال نه، شماره سه، ص ۱-۱۰.

Chen, X., and S. Zhu. 2013. Improved hybrid model based on support vector regression machine for monthly precipitation forecasting. Journal of computers, 8(1): 232-239.

Du, J., Y. Liu, Y. Yu and W. Yan.2017. A Prediction of Precipitation Data Based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) Algorithms. Algorithms, 10(57): 1-15.

Hamidi, O., J. Poorolajal, M. Sadeghifar, H. Abbasi, Z. Maryanaji, H.R. Faridi and L.Tapak.2015. A comparative study of support vector machines and artificial neural networks for predicting precipitation in Iran. Theor Appl Climatol, 119: 723–731.

Hasan, N., N.Chandra Nath and R. Islam Rasel.2015. A support vector regression model for forecasting rainfall. Proceedings of International Conference on Electrical Information and Communication Technology, 554-559.

Ingsrisawang, L., S. Ingsriswang, S.Somchit, P. Aungsuratana and W. Khantiyanan. 2008. Machine learning techniques for short-term rain forecasting system in the northeastern part of Thailand. International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2(5): 1422-1427.

Jiajia, H., CH. Kai, CH. Jinsong, X. Wenwen, T. Li andL.Jun.2017.A multi-time scale SVM method for local short-term rainfall prediction. Meteorology, 43(4): 402–412.

Langhammer, J. and J. Cˇ esák.2016. Applicability of a nu-support vector regression model for the completion of missing data in hydrological time series. Water, 8(560): 1-25.

Moon, S. andY.Kim.2020. An improved forecast of precipitation type using correlation-based feature selection and multinomial logistic regression. Atmospheric Research, 240: 1-14.

Ortiz-Garcı, E.G., S. Salcedo-Sanz and C. Casanova-Mateo. 2014. Accurateprecipitation prediction with support vector classifiers: a studyincluding novel predictive variables and observational data. Atmospheric Research, 139: 128–136.

Papacharalampous, G., H. TyralisandD. Koutsoyiannis.2018.Univariate Time Series Forecasting of Temperature and Precipitation with a Focus on Machine Learning Algorithms: a Multiple-Case Study from Greece. Water Resources Management, 32: 5207–5239.

Pannkkong, W., V.H. Pham and V.N.Huynh. 2016. A hybrid model of ARIMA, ANNs and k-means clustering for time series forecasting. Lecture Notes in Computer Science, 8(4): 30-53.

Samui, P., V.Ravibabu Mandla, A.Krishnaand T. Teja.2011. Prediction of Rainfall Using Support Vector Machine and Relevance Vector Machine, Earth Science India, 4: 188-200.

Sehad, M., M. Lazri andS.Ameur.2017. Novel SVM-based technique to improve rainfall estimation over the Mediterranean region (north of Algeria) using the multisperctral MSG SEVIRI imagery. Advances in Space Research, 59: 1381-1394.

Shenify, M., A.S. Danesh, M. Gocić, R.S. Taher, A.W. Abdul Wahab, A.Gani, SH. Shamshirband and D. Petković.2016. Precipitation Estimation Using Support Vector Machine with Discrete Wavelet Transform. Water Resource Management, 30: 641–652.

Soares dos Santos, T., D. Mendes and R Rodrigues Torres. 2016. Artificial neural networks and multiple linear regression model using principle components to estimate rainfall over South America. Nonlinear Processes Geophysics, 23: 13-20.

UlSaufie, A.Z., A.S. Yahya and N.A. Ramli. 2011. Improving multiple linear regression model using principal componentanalysis for predicting PM10 concentration in Seberang Prai, PulauPinang. International Journal of Environmental Sciences, 2(2): 403-410.

Zaynoddin, M., H. Bonakdari, A.Azari, I.Ebtehaj, B.Gharabaghi and H.Riahi Madavar. 2018. Novel hybrid linear stochastic with non-linear extreme learning machine methods for forecasting monthly rainfall a tropical climate. Journal of Environmental Management, 222: 190-206.

Zhang, CH., H.Wang, J.Zeng, L.MA and L. Guan.2020. Short-term dynamic radar quantitative precipitation estimation based on wavelet transform and support vector machine. Journal of Meteorological Research, 34: 413-426.

Combining soft Computing Model Based on Machine Learning Algorithm and Principal Component Analysis for Precipitation Forecasting

Laleh Parviz[1]

Abstract

The effect of precipitation changes on water resources, agricultural production reveals the need for accurate methods for precipitation forecasting. In this research, one of the soft computing methods was developed in order to forecast precipitation with the data reduction approach. Input data of model was mean air temperature, dew point temperature, mean sea level pressure, mean station pressure, mean relative humidity and mean wind speed at Tabriz, Ahar and Jolfa stations. The method used in this study includes Epsilon and Nu support vector regression. In all studied stations, the use of Nu support vector regression compared to Epsilon reduced the error so that UII values with Nu support vector regression in Tabriz, Ahar and Jolfa stations were decreased 19.19, 5.88 and 15.78%, respectively. The results indicate the limitation of using the data reduction approach for data with KMO factor lower than 0.5, which included Tabriz and Ahar stations. Principal component analysis in both types of support vector regression increased the performance of the model so that in Jolfa station by using principal component analysis d values of Epsilon and Nu support vector regression increased by 16.6 and 17.5%. Execution of Verimax rotation in preprocessing of input data to regression was stronger than principal component analysis. In this regard, RRMSE and RMSE values in Jolfa station using Epsilon support vector regression were decreased 6.66 and 6.45%. Therefore, principal component analysis is a suitable tool to improve the performance of soft computing methods by regarding the relevant constraints.

Keywords: Precipitation, support vector regression, KMO factor, principal component analysis.

[1]Associate Professor, Faculty of Agriculture, Azarbaijan Shahid Madani University, Tabriz, Iran, laleh_parviz@yahoo.com

Research Paper

Combining soft Computing Model Based on Machine Learning Algorithm and Principal Component Analysis for Precipitation Forecasting

Laleh Parviz

[1]Associate Professor, Faculty of Agriculture, Azarbaijan Shahid Madani University, Tabriz, Iran, laleh_parviz@yahoo.com

10.22125/IWE.2020.239917.1391

Received: **July.17.2020**
Accepted: **December.13.2020**
Available online: **March.13.2022**

Keywords: Precipitation, support vector regression, KMO factor, principal component analysis.

Abstract

The effect of precipitation changes on water resources, agricultural production reveals the need for accurate methods for precipitation forecasting. In this research, one of the soft computing methods was developed in order to forecast precipitation with the data reduction approach. Input data of model was mean air temperature, dew point temperature, mean sea level pressure, mean station pressure, mean relative humidity and mean wind speed at Tabriz, Ahar and Jolfa stations. The method used in this study includes Epsilon and Nu support vector regression. In all studied stations, the use of Nu support vector regression compared to Epsilon reduced the error so that UII values with Nu support vector regression in Tabriz, Ahar and Jolfa stations were decreased 19.19, 5.88 and 15.78%, respectively. The results indicate the limitation of using the data reduction approach for data with KMO factor lower than 0.5, which included Tabriz and Ahar stations. Principal component analysis in both types of support vector regression increased the performance of the model so that in Jolfa station by using principal component analysis d values of Epsilon and Nu support vector regression increased by 16.6 and 17.5%. Execution of Verimax rotation in preprocessing of input data to regression was stronger than principal component analysis. In this regard, RRMSE and RMSE values in Jolfa station using Epsilon support vector regression were decreased 6.66 and 6.45%. Therefore, principal component analysis is a suitable tool to improve the performance of soft computing methods by regarding the relevant constraints.

1. Introduction

Precipitation is one of the important components of the hydrological cycle with temporal and spatial variation. Precipitation forecasting is very complex due to dependency of that on many parameters such as temperature, humidity and wind speed. In fact, the different variation of precipitation makes its forecasting difficult. In recent years, the soft computing models have been significantly used for precipitation forecasting. Due to the high efficiency of soft computing models for precipitation forecasting, an attempt is made to develop the expressed models. One aspect of development is the use of principle component analysis in reducing the number of input data. The aim of this study is to develop one of the soft computing models for precipitation forecasting. For this purpose, the Nu and Epsilon support vector regression were investigated. The effect of rotation on the efficiency of principal component analysis and finally on the performance of two types of support vector regression was also studied.

2. Materials and Methods

The stations of this research are Ahar, Tabriz and Jolfa with studied period from 1987 to 2017. Input data of model was mean air temperature, dew point temperature, mean sea level pressure, mean station

pressure, mean relative humidity. In this study, the effect of data reduction (principle component analysis, PCA) on the input data of support vector regression was investigated. PCA is a mathematical method to transform the number of correlated variables into the number of uncorrelated variables which called principle components. PCA can be introduced as a dimensionality reduction method. A linear combination of variables is used in the structure of PCA to extract the maximum variance of variables. Generally, PCA extracts the most important information of data and compresses the size of data. PCA can investigate the structure of observation and variables. The description of the data set is simple with this method. With a loss function, support vector machine (SVM) can be used for regression problems. The mapping of input data x into a higher dimensional feature space is possible with SVM and in this case the solving of a linear regression problem is possible. The basis of SVM is on the statistical learning theory. For regression problems, the non-linear function is learned with the linear learning machine in the form of kernel induced feature space. Also, the parameter is defined to control the capacity of system. RMSE, RRMSE, GMER, MAPE, UI, UII and dare the evaluation criteria for model performance investigation.

3. Results

One of the processes which can impact the model performance is the sensitivity analysis which the minimum error in Tabriz station is related to the linear function. Nu support vector regression (Nu SVR) can decrease the values of error criteria for example at Tabriz station, the RMSE, RRMSE, MAPE, UI and UI decreasing from Epsilon-SVR to Nu-SVR is 19.45%, 20%, 14.47%, 18.36% and 19.9%, respectively. In the tree stations(average), the RMSE, RRMSE, MAPE, UI, UII decreasing and d increasing from Epsilon-SVR to Nu-SVR is 10.77%,10.36%,4.5%,13.53%, 12.85% and 4.63%, respectively. The primary processing in the PCA is KMO factor calculation. The values of this factor for Ahar, Tabriz and Jolfa stations were estimated to be 0.3, 0.41 and 0.55, respectively. KMO values in Tabriz and Ahar stations, due to being lower than 0.5, indicate that the data of these stations are not suitable for PCA. Combination PCA and SVR could improve the forecasts, for example from Epsilon SVR to Epsilon SVR-PCA, the decreasing of RRMSE, RRMSE, UI, UII and d increasing is 8.55%, 5.26%, 10%, 10.52%, 16.66%, respectively.

4. Discussion and Conclusion

Sensitivity analysis can affect on the forecasts of model. Due to the precise performance of SVR, the type of parameters involved in regression problems has more importance, so that the effect of Epsilon and Nu was observed for precipitation forecasts. Nu SVR has better performance relative to Epsilon SVR. The expressed parameters have their effect on the classification, separation of classes and controlling the number of support vectors in regression. The difference between Epsilon and Nu is in how they parameterize the training problem. Nu controls the number of support vectors. An important feature of SVM is the ability to model nonlinear process without knowing the statistical distribution of the classes. Another important feature of SVM is its good performance on high dimensional data and small batch of training pattern. The KMO factor is an effective factor in determining the performance of component analysis on data. High values of the KMO factor can be due to the low partial correlation coefficient between the data. Therefore, combining principal component analysis and support vector regression can be a good solution for rainfall forecasting and water resources planning decisions. In fact, with PCA the arrangement of input data to the model is done based on certain rules and statistics and statistics. In the rotation mode, the matrix columns are simplified using the Varimax method, which facilitates access to a simple matrix.

5. Six important references

1. Chen, X., and S. Zhu. 2013. Improved hybrid model based on support vector regression machine for monthly precipitation forecasting. Journal of computers, 8(1): 232-239.

2. Hamidi, O., J. Poorolajal, M. Sadeghifar, H. Abbasi, Z. Maryanaji, H.R. Faridi and L.Tapak. 2015. A comparative study of support vector machines and artificial neural networks for predicting precipitation in Iran. Theoretical and Applied Climatology, 119:723–731.

3. Moon, S. andY.Kim.2020. An improved forecast of precipitation type using correlation-based feature selection and multinomial logistic regression. Atmospheric Research, 240: 1-14.

.4Soares dos Santos, T., D. Mendes and R Rodrigues Torres. 2016. Artificial neural networks and multiple linear regression model using principle components to estimate rainfall over South America. Nonlinear Processes Geophysics, 23: 13-20.

5. UlSaufie, A.Z., A.S. Yahya and N.A. Ramli. 2011. Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, PulauPinang. International Journal of Environmental Sciences, 2(2): 403-410.

6. Zaynoddin, M., H. Bonakdari, A.Azari, I.Ebtehaj, B.Gharabaghi and H.Riahi Madavar. 2018. Novel hybrid linear stochastic with non-linear extreme learning machine methods for forecasting monthly rainfall a tropical climate. Journal Environmental Management, 222: 190-206.

Conflict of Interest

Authors declared no conflict of interest.

Acknowledgments

We are grateful to Vice President for Research and Technology of Azarbaijan Shahid Madani University for financial Support.