مجله ترویجی علوم و فناوری اطلاعات کشاورزی
دوره هفتم، شماره ۱، شماره پیاپی ۱۳، بهار و تابستان ۱۴۰۳
jaist.areeo.ac.ir

شناسه دیجیتال (DOI): 10.22092/JAIST.2023.363907.1103

کاربرد روش ماشین‌های بردار پشتیبان در طبقه‌بندی داده‌های اقلیمی

سیدمسعود سلیمان‌پور[۱]، پارسا حقیقی[۲]

۱.دانشیار بخش تحقیقات حفاظت خاک و آبخیزداری، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان فارس، سازمان تحقیقات، آموزش و ترویج کشاورزی، شیراز، ایران
***(نویسنده مسئول)**. رایانامه: m.soleimanpour@areeo.ac.ir*

۲. کارشناس بخش تحقیقات حفاظت خاک و آبخیزداری، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان فارس، سازمان تحقیقات، آموزش و ترویج کشاورزی، شیراز، ایران. رایانامه: p.haghighi@areeo.ac.ir

تاریخ دریافت: ۱۴۰۲/۰۸/۰۳ تاریخ ویرایش: ۱۴۰۲/۰۹/۲۵ تاریخ پذیرش: ۱۴۰۲/۱۱/۱۰ تاریخ چاپ: ۱۴۰۳/۰۱/۲۰ صص : ۱۱-۱

چکیده

در یک ساختار اقلیمی، شناسایی، پیش‌بینی و مدیریت بحران اهمیت زیادی دارد. برای درک نظام‌های پیچیده و شبیه‌سازی و پیش‌بینی رفتار آن‌ها از الگوها، به‌عنوان ابزارهایی کاربردی، استفاده می‌شود. ماشین‌های بردار پشتیبان یکی از روش‌های یادگیری نظارت‌شده است که برای طبقه‌بندی و رگرسیون از آن استفاده می‌شود. ماشین‌های بردار پشتیبان می‌توانند الگوهای پنهان را تشخیص داده و به تغییرات پیچیده در داده‌های اقلیمی پاسخ دهند. در این مقاله، ساختار روش ماشین‌های بردار پشتیبان و کاربرد آن‌ها در طبقه‌بندی داده‌های اقلیمی بررسی شده است. ویژگی‌های ساختار ماشین‌های بردار پشتیبان به انتخاب نوع تابع هسته مرتبط است. بنابراین، در انتخاب نوع تابع هسته باید دقت زیادی شود. ازسوی دیگر، در پیش‌بینی اقلیمی، تعیین مؤلفه اصلی مرحله بسیار مهمی است تا بااستفاده از تعداد مؤلفه بهینه بتوان بین داده‌های پیش‌بینی‌کننده و پیش‌بینی‌شونده بهترین برازش را انجام داد.

کلیدواژه‌ها: ماشین‌های بردار پشتیبان، طبقه‌بندی داده‌های اقلیمی، توابع هسته، تعیین مؤلفه اصلی.

مقدمه

ایران با مشکلات بی‌سابقه اقلیمی ازجمله خشک شدن دریاچه‌ها و رودخانه‌ها، توفان‌های گردوغبار، دمای بی‌سابقه، خشک‌سالی و سیل روبرو است. نگرانی از تغییرات آب‌وهوایی و تاثیرات آن بر جنبه‌های مختلف زندگی انسان در حال افزایش است. در مدیریت خشک‌سالی و سیلاب، پیش‌بینی‌های اقلیمی بسیار بااهمیت است. بنابراین، لزوم توجه به روش‌های نوین در پیش‌بینی اقلیمی و بررسی کارکرد روش‌ها بر کسی پوشیده نیست. پیش‌بینی فرآیندهای آب و هوایی، به‌ویژه پیش‌بینی بارندگی، ابزار مناسبی در اختیار مدیران بخش‌های مختلف (کشاورزی و منابع طبیعی و غیره) قرار می‌دهد تا برای بهینه‌سازی هزینه‌ها، امکانات و بهره‌وری حداکثری از آن‌ها، مقابله با خشک‌سالی و کاهش خسارت ناشی از آن، سیاست‌های آینده را برنامه‌ریزی کنند. بارندگی مبحثی کلیدی در چرخه آبشناسی است که بیش‌ازپیش، هم در سطح منطقه و هم در سطح جهان، در حال تغییر است (نسوبوگا و همکاران[1]، ۲۰۱۴) و درزمینه مدیریت منابع آب یکپارچه، پیش‌بینی نیازهای محصول و ارزیابی محیط‌زیست به شکلی گسترده استفاده می‌شود (هیان و همکاران[2]، ۲۰۰۷). بنابراین، برای مدیریت منابع آب، کشاورزی و بسیاری از کاربران، مؤلفه‌های آب و هوایی بسیار جالب‌توجه است (گابین و همکاران[3]، ۲۰۱۳). همچنین، موقعیت جغرافیایی ایران در منطقه کناره حاره، کشور ما را در کمربند خشک جهان قرار داده است. باتوجه به این موقعیت جغرافیایی، میانگین بارش ایران بسیار کمتر از میانگین جهانی آن است (صفوی، ۱۳۸۸). یادگیری ماشینی[4] به‌عنوان یکی از شاخه‌های وسیع و پرکاربرد هوش مصنوعی، شیوه‌ها و روش‌هایی را تنظیم و ابداع می‌کند که رایانه‌ها و سامانه‌ها براساس آن‌ها می‌توانند یاد بگیرند. استفاده از یادگیری ماشینی در کاهش هزینه‌های عملیاتی و بهبود سرعت عمل تجزیه‌وتحلیل داده‌ها بسیار به‌صرفه است.

یادگیری ماشینی به دو دسته کلی «یادگیری نظارت‌شده[5]» و «یادگیری نظارت‌نشده[6]» تقسیم می‌شود. در روش‌های یادگیری ماشینی نظارت‌شده، مجموعه‌ای از بردارهای ورودی مانند X و بردارهای خروجی متناظر با آن‌ها مانند t به رایانه ارائه می‌شود. هدف آن است که با استفاده از این داده‌های آموزشی در ورودی X جدید، ماشین بتواند t را پیش‌بینی کند. ازجمله روش‌های یادگیری نظارت‌شده می‌توان به روش‌های طبقه‌بندی[7] مانند شبکه‌های عصبی[8]، درخت تصمیم[9]، بیز ساده[10]، نزدیک‌ترین همسایگی[11]، ماشین‌های بردار پشتیبان[12] و روش‌های رگرسیون مانند رگرسیون خطی[13]، رگرسیون غیرخطی[14] و رگرسیون بردار پشتیبان[15] اشاره کرد (استوارت و همکاران[16]، ۲۰۱۰). در روش‌های یادگیری نظارت‌نشده، یادگیری ماشین تنها ازطریق داده‌های ورودی انجام می‌شود. یعنی، مجموعه داده‌ها تنها شامل متغیرهای ورودی است و هیچ خروجی متناسبی با ورودی‌ها وجود ندارد. بنابراین، در یادگیری نظارت‌نشده، الگوریتم یادگیری، خود، الگو و ساختار میان داده را پی‌می‌گیرد. درواقع، یادگیری نظارت‌نشده روشی است که از آن برای یافتن الگوهای میان داده‌ها استفاده می‌شود. به‌عبارت‌دیگر، با یادگیری نظارت‌نشده می‌توان ساختار و الگوهای پنهان میان داده‌ها را یافت. ازجمله روش‌های یادگیری نظارت‌نشده می‌توان به روش‌های خوشه‌بندی[17] مانند کی- میانگین[18]، دی بی اسکن[19]

1.Nsubuga et al
2. Haiyun et al
3. Guobin et al
4. Machine Learning
5. Supervised Learning
6. Unsupervised Learning
7. Classification
8. Artificial Neural Network (ANN)
9. Decision Tree
10. Naive Bayes
11. K Nearest Neighbor
12. Support Vector Machines (SVM)
13. Linear Regression
14. Non-Linear Regression
15. Support Vector Regression
16. Stuart J at al
17. Clustering
18. K-Mediods
19. DBSCAN

و روش‌هـای کاهـش ابعاد[1] مانند تحلیل مؤلفـه اصلی[2] و آنالیز تشخیصی خطی[3] اشاره کرد (هیستی در همکاران[4]، ۲۰۰۱).

روش ماشـین‌های بردار پشـتیبان، ابزاری جدید و قدرتمند برای تدوین راه‌حل‌هایی برای طبقه‌بندی و رگرسیون است. این روش بر تئوری یادگیری آماری استوار است و در آن، برمبنای مجموعـه‌ای از داده‌های آموزشـی، یادگیری انجام می‌شـود. تحقیقات نشـان می‌دهند که استفاده از این روش یادگیری، در مقایسه با سـایر روش‌های داده‌محور پیشین ازجمله شبکه‌های عصبـی مصنوعـی، حتی در شـرایطی که ابعـاد داده‌ها زیاد و تعداد نمونه‌ها کم باشـد، دقت الگو، نتایج را بهبود می‌بخشـد. به‌طورکلی، از ماشین‌های بردار پشتیبان، هم در دسته‌بندی و هم در رگرسیون استفاده می‌شود. گرچه از مبنای تئوری این روش برای طبقه‌بندی اسـتفاده می‌شـود، اما اخیراً با گسـترش روش ماشـین‌های بردار پشـتیبان و مقبولیت آن در اسـتفاده، نسبت بـه روش‌های قدیمی‌تـر ازجمله شـبکه‌های عصبی مصنوعی، تئـوری جدیـد و جامع‌تـری در محدوده مسـائل رگرسـیون نیز گسـترش‌یافته اسـت. در سـال ۱۹۹۵، وپنیک[5] استفاده از ماشین‌های بردار پشتیبان، به‌عنوان ابزار رگرسیون، را برای اولین بار مطرح کرد. از روش ماشـین‌های بردار پشتیبان در پیش‌بینی و طبقه‌بندی شـاخص‌های اقلیمی اسـتفاده شده است. خاشعی سیوکی و همکاران (۱۳۹۷)، سهندرا[6] و همکاران (۲۰۱۳) و لو و کین[7] (۲۰۱۴) کارایی روش ماشـین‌های بردار پشتیبان را در طبقه‌بندی و پیش‌بینی مؤلفه‌های بارش و دما بررسی کرده و آن را با سایر روش‌ها مقایسه کرده‌اند. نتایج پژوهش آن‌ها گویای دقـت بالای روش ماشـین‌های بردار پشـتیبان در طبقه‌بندی و پیش‌بینی اقلیمی است.

ماشین‌های بردار پشتیبان

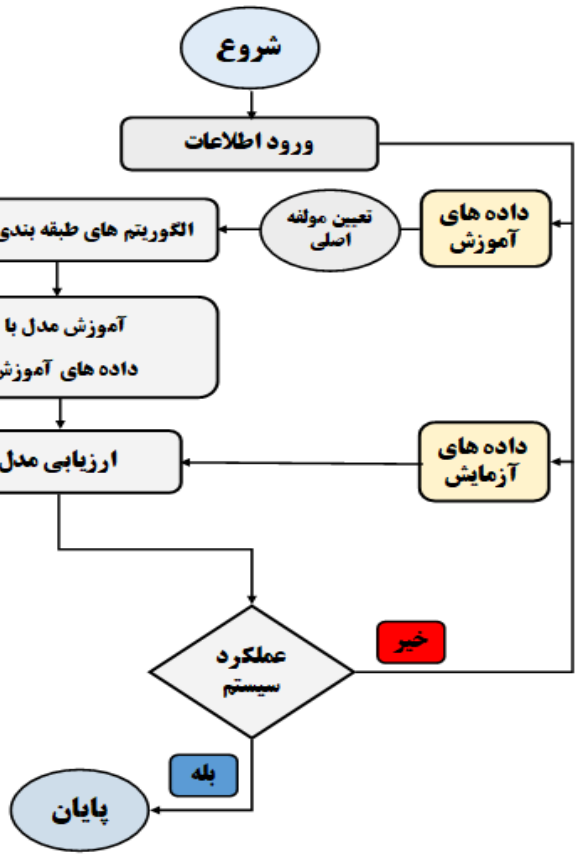
ماشـین‌های بردار پشـتیبان یکـی از روش‌هـای یادگیری بانظارت اسـت که از آن برای طبقه‌بندی و رگرسـیون استفاده می‌شـود. به‌طورکلی، در یادگیری نظارت‌شـده، با استفاده از مجموعـه‌ای از داده‌هایی کـه خروجی مطلـوب آن‌ها از قبل مشخص است، الگو را آموزش می‌دهند. وقتی یک داده جدید که خروجی آن مشخص نیست به الگو ارائه شد، الگو می‌تواند خروجـی مطلوب را تولید کند. ولادیمیـر وپنیک، از محققین روسـی، در سـال ۱۹۶۵ نظریه آماری یادگیـری را به‌صورت مسـتحکم‌تری بنا نهاد و بر این اساس ماشین‌های بردار پشتیبان را ارائه داد. ماشـین‌های بردار پشـتیبان دارای ویژگی‌های زیر هستند (شین و همکاران[8]، ۲۰۰۵):

- طراحی طبقه‌بند با حداکثر تعمیم؛
- رسیدن به نقطه بهینه کلی تابع؛
- تعییـن خـودکار سـاختار و جاینـگاری بهینـه بـرای طبقه‌بندی‌کننده؛
- الگوکـردن توابع تمایز غیرخطی با اسـتفاده از هسـته‌های غیرخطی و مفهوم حاصل‌ضرب داخلی در فضاهای هیلبرت (شین و همکاران، ۲۰۰۵).

ماشین‌های بردار پشـتیبان الگوریتمی است که نوع خاصی از الگوهـای خطی را می‌یابد که حداکثر حاشـیه ابرصفحه را به دسـت می‌دهند. حداکثرکردن حاشـیه ابرصفحه به حداکثر شـدن تفکیک بین طبقات منجر می‌شود. به نزدیک‌ترین نقاط آموزشـی به حداکثر حاشیه ابـر صـفحه، بردارهای پشتیبان اطلاق می‌شـود. اگر داده‌ها به‌صـورت خطی از یکدیگر مجزا باشـند تنها از این بردارها (نقاط) برای مشخص کردن مرز بین طبقات استفاده می‌شود. ماشین‌های بردار پشتیبان به ماشین‌های خطی آموزش می‌دهند که چگونه سطح بهینه‌ای تولید کنندکه بـدون خطا و با حداکثر فاصله میان صفحه و نزدیک‌ترین نقاط آموزشـی بردارهای (پشـتیبان)، داده‌ها را تفکیک کنند (مین

1. Dimensionality Reduction
2. Principal Component Analysis
3. Linear Discriminant Analysis
4. Hastie at al
5. Vapnik
6. Sahindra
7. Lu & Qin
8. Shin at al
9. Min at al

و همکاران[9]، ۲۰۰۵). برای الگوسازی در ماشین‌های بردار پشتیبان، ابتدا داده‌ها به دو بخش آموزش و آزمایش تقسیم شده و اطلاعات پیش‌پردازش می‌شوند. سپس، با استفاده از معیارهای خطا، نوع الگوریتم مناسب برای طبقه‌بندی اطلاعات انتخاب و الگو ارزیابی می‌شود. در ادامه، برای بررسی کارایی الگو، داده‌های آزمایشی به‌عنوان ورودی الگو در نظر گرفته می‌شوند. اگر دقت الگو کافی باشد روند خروجی و گزارش نهایی تکمیل می‌شود، در غیر این صورت باید مؤلفه‌های ورودی و نوع الگوریتم طبقه‌بندی ماشین‌های بردار پشتیبان بازبینی‌شده و اجرای الگو تکرار شود. در بحث پیش‌بینی اقلیمی نیز می‌توان از ماشین‌های بردار پشتیبان استفاده کرد. پس از تعیین مؤلفه‌های اصلی، داده‌های مؤلفه‌های اقلیمی به دو دسته آموزش و آزمایش تقسیم می‌شوند. پس از انتخاب نوع تابع هسته و الگوریتم‌های طبقه‌بندی ماشین‌های بردار پشتیبان، خروجی الگو ارزیابی خواهد شد. در شکل ۱ روند کلی الگوسازی ماشین‌های بردار پشتیبان ارائه‌شده است.



شکل ۱. فرآیند الگوسازی در ماشین‌های بردار پشتیبان

ساختار ماشین‌های بردار پشتیبان

ماشین بردار پشتیبان، یک الگوریتم یادگیری با ناظر است که نمونه داده‌ها را به‌صورت نقاطی در فضا نشان می‌دهد و با استفاده از یک خط[1] از هم جدا می‌کند. این جداسازی به‌گونه‌ای است که نقاط داده‌ای که در یک‌طرف خط هستند مشابه به هم و در یک گروه قرار می‌گیرند. نمونه داده‌های جدید نیز بعد از اضافه شدن به همان فضا، دریکی از دسته‌های موجود قرار خواهند گرفت. فاصله بین دسته‌های مختلف داده‌ها را حاشیه[2] گویند و برای تعیین محلِ مرز بین دو دسته‌بندی، از

1.Hyperplane
2. Margin

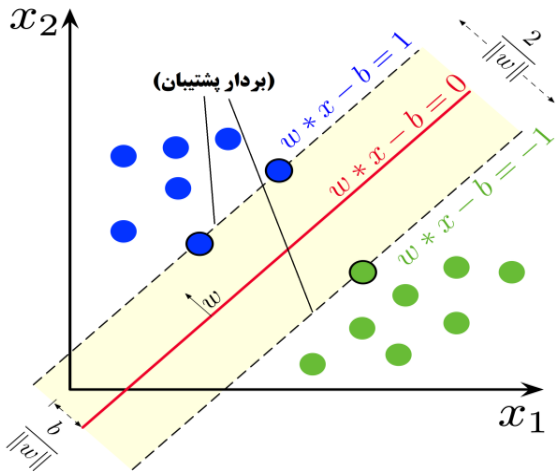
بردار پشتیبان استفاده می‌شود. توابع هسته، با اضافه کردن ابعاد بیشتر به مسئله، مسائل تفکیک‌ناپذیر را به مسائل تفکیک‌پذیر تبدیل می‌کند تا دسته‌بندی بهتری حاصل آید. در ادامه ساختار ماشین‌های بردار پشتیبان شرح داده‌شده است.

۱) بردارهای پشتیبان[1]

بردار پشتیبان در ماشین‌های بردار پشتیبان نقاط داده‌هایی است که برای تصمیم‌گیری در فرآیند طبقه‌بندی استفاده می‌شود. این بردارها به‌طور خاص در مرز جداکننده بین دو کلاس قرار دارند و در تعیین حاشیه جداکننده نقش اصلی را برعهده دارند. فرمول رسم خط جداکننده در ماشین‌های بردار پشتیبان به‌صورت رابطه ۱ است:

$$w^T * x - b = 0 \quad (1)$$

که در آن، w بردار وزن است که برابر با جمع ضرایب نمونه‌های پشتیبان، x بردار ویژگی‌ها و b عبارت است از تعداد منفی ضرایب نمونه‌های پشتیبان تقسیم بر تعداد نمونه‌های پشتیبان. بردار پشتیبان در ماشین‌های بردار پشتیبان، نقاط حاشیه دسته‌بندی را تعیین می‌کند و به‌عنوان نقاط حاشیه اصلی در ساخت مکان مرز بین دسته‌بندها شناخته می‌شود. در شکل ۲ حاشیه بین دودسته و بردار پشتیبان نشان داده‌شده است:



شکل ۲. حداکثر حاشیه بین دو دسته و بردار پشتیبان

۲) حاشیه[2]

در ماشین‌های بردار پشتیبان، به فاصله بین دسته‌های مختلف داده‌ها در فضای ویژگی، حاشیه می‌گویند. حاشیه بزرگ‌تر نشان‌دهنده جداپذیری بهتر بین دسته‌ها است. در ماشین‌های بردار پشتیبان، هدف اصلی آن است که بین داده‌های دو کلاس مختلف، یک خط جداکننده بیابند. این خط جداکننده[3] باید طوری باشد که فاصله بین خط و نزدیک‌ترین نقطه از هر دو کلاس (که به آن‌ها بردار پشتیبان یا support vectors می‌گویند) حداکثر شود (وین و همکاران[4]، ۲۰۱۵). رابطه ۲ مربوط به حاشیه در ماشین‌های بردار پشتیبان است؛ که در آن، w بردار وزن و ||w|| نرم آن است.

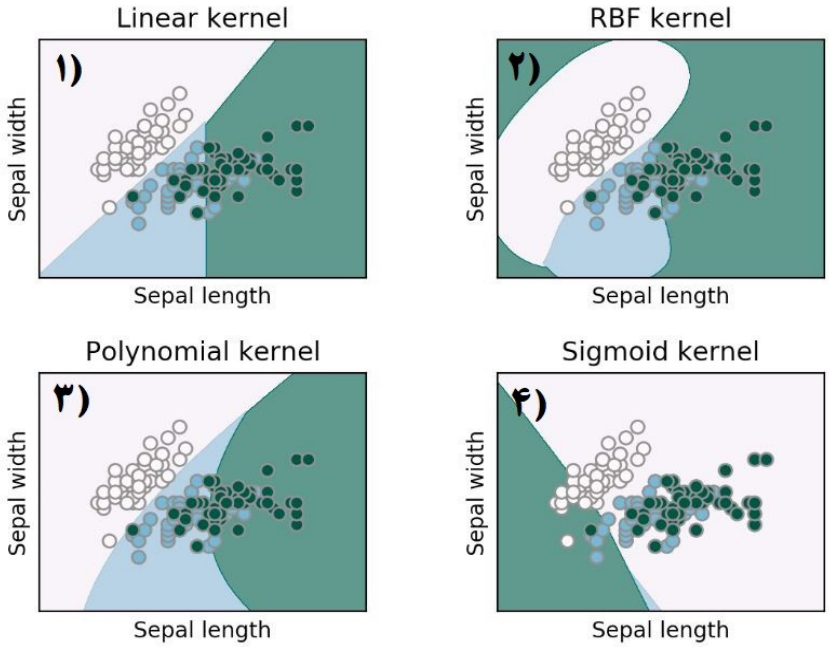
$$\text{Margin} = 2 / \|w\| \quad (2)$$

۳) توابع هسته[5]

همان‌طور که قبلاً اشاره شد، ماشین‌های بردار پشتیبان ممکن است برای تقسیم‌بندی داده‌های غیرخطی استفاده شود، درحالی‌که برای داده‌های خطی طراحی‌شده است. برای اینکه طبقه‌بندی بر روی داده‌های غیرخطی کار کند

1.Support Vectors
2. Margin
3. Hyper Plane
4. Wen, at al
5. Kernel Functions

از مؤلفه‌ای به نام توابع هسته استفاده می‌شود. در ماشین‌های بردار پشتیبان، از توابع هسته برای انتقال داده‌ها به فضای بالاتر استفاده می‌شود. در این فضا، دسته‌بندی بهتری انجام می‌شود. تابع هسته غیرخطی عمل می‌کند و امکان می‌دهد تا داده‌های ورودی به فضای با بعد بالاتر منتقل شوند، به‌طوری‌که در این فضا، داده‌های جدید قابل‌تفکیک باشند. از انواع توابع هسته در ماشین‌های بردار پشتیبان می‌توان به تابع خطی[1]، چندجمله‌ای[2]، تابع پایه شعاعی[3]، گوسین[4]، سیگموید[5]، تابع لاپلاس[6] و هسته شبکه عصبی[7] اشاره کرد (میر عربی و همکاران، ۲۰۱۹). در شکل ۳ انواع طبقه‌بندی مجموعه داده‌ها با استفاده از کرنل‌های مختلف ارائه شده است.



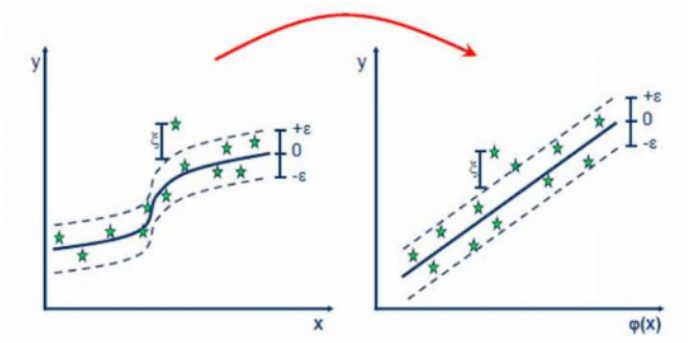
شکل ۳. طبقه‌بندی مجموعه داده‌ها با استفاده از کرنل‌های: (۱) خطی، (۲) تابع پایه شعاعی، (۳) چندجمله‌ای و (٤) سیگموید

رگرسیون بردار پشتیبان[8]

شکل پرکاربرد ماشین‌های بردار پشتیبان، رگرسیون بردار پشتیبان است. ماشین‌های بردار پشتیبان برای انواع الگوهای پیش‌بینی در طبقه‌بندی و رگرسیون استفاده می‌شود. زمانی که از رگرسیون استفاده می‌شود به آن رگرسیون بردار پشتیبان می‌گویند. در الگوی مسئله رگرسیونی، ورودی‌ها (مقادیر بارش، دما و تبخیر و تعرق) به یک فضای چندبعدی ترسیم می‌شوند و رگرسیون بردار پشتیبان ازطریق اصل به حداقل رساندن ریسک ساختاری[9] آموزش داده می‌شود. رگرسیون بردار پشتیبان روشی ناپارامتریک است که واپنیک[10] (۱۹۹۵) توسعه داده است. عملکرد رگرسیون بردار پشتیبان به‌شدت به هسته آن بستگی دارد. در ابتدا رگرسیون بردار پشتیبان خطی ارائه‌شده بود که بعدها برای مسائل غیرخطی تعمیم داده شد. واپنیک برای طبقه‌بندی داده‌های غیرخطی از نگاشت غیرخطی کمک گرفت. برای تعمیم رگرسیون بردار پشتیبان به مسائل غیرخطی، ابتدا داده‌ها با کمک یک نگاشت غیرخطی به فضای

1.Linear Kernel
2. Polynomial Kernel
3. Radial Basis Function
4. Gaussian Kernel
5. Sigmoid Kernel
6. Laplacian Kernel
7. Neural Network Kernel
8. Support Vector Regression (SVR)
9. Structural Risk Minimization (SRM)
10. Vapnik

خطی نگاشت می‌یابند. سپس در فضای جدید رابطه خطی بین داده‌ها در فضای ویژگی و خروجی به دست می‌آید که این رابطه خطی در فضای ویژگی، معادل رابطه غیرخطی بین ورودی و خروجی در فضای اصلی است. نگاشت ورودی‌ها به فضای ابعاد بالاتر با استفاده از یک تابع هسته انجام می‌شود تا در آن فضا خط جداساز بهتر و دقیق‌تر عمل کند. شکل ۴ انتقال داده‌ها از یک ساختار غیرخطی به فضای خطی را نشان می‌دهد (تریپاتی[1] و همکاران، ۲۰۰۶).



شکل ٤. انتقال داده‌ها از یک ساختار غیرخطی به فضای خطی

انتخاب متغیرهای پیش‌بینی کننده غالب

کارن پیرسون (۱۹۰۱) برای اولین بار فن تجزیه به مؤلفه‌های اصلی را شرح داده است. در روش تحلیلی تحلیل مؤلفه اصلی سعی می‌شود، تعداد مؤلفه‌ها تا حد امکان کاهش یابد و در چند مؤلفه اصلی و اثرگذار خلاصه شود. در تحلیل عاملی بعد از به دست آوردن ماتریس همبستگی، ابتدا باید مشخص کنیم که برای تحلیل عاملی از کدام الگو استفاده کنیم. روش تحلیل مؤلفه‌های اصلی دو کارکرد مهم دارد. اولین کارکرد این روش آن است که عامل‌ها را به‌صورت مستقیم و بدون برآورد اشتراکات، از ماتریس همبستگی تعیین می‌کند. در این روش برای تبیین حداکثر مقدار واریانس متغیرها، ترکیب خطی آن‌ها برآورد می‌شود. بدین‌صورت که اولین مؤلفه، بیشترین واریانس متغیرها را تبیین می‌کند. سپس مؤلفه دوم بیشترین مقدار واریانس باقی‌مانده در متغیرها را بعد از مؤلفه اول توضیح می‌دهد و به همین ترتیب تا آخر ادامه می‌یابد. کارکرد دیگر تحلیل مؤلفه‌های اصلی این است که مجموعه‌ای از متغیرهای سنجیده شده را به مجموعه‌ای از ترکیب خطی با حداکثر مقدار واریانس تبدیل می‌کند (گوچیچ و همکاران[2]، ۲۰۱۶). در پیش‌بینی متغیرهای اقلیمی نیز می‌توان با استفاده از این روش، مؤلفه‌های اصلی را انتخاب و در الگوسازی از تعداد مؤلفه‌های بهینه استفاده کرد. برایر پیش‌بینی اقلیمی بارش و دما می‌توان از ۲۶ متغیر پیش‌بینی کننده بزرگ‌مقیاس با نام داده‌های بزرگ‌مقیاس استفاده کرد. داده‌های NCEP برای نمونه شامل میانگین فشار سطح دریا، رطوبت نسبی در ارتفاع ۸۵۰ هکتوپاسکال، میانگین دما در ارتفاع ۲ متری، رطوبت نسبی در ارتفاع ۵۰۰ هکتوپاسکال و غیره است. با تعیین ارتباط بین داده‌های مشاهداتی ایستگاه محلی (بارش و دما) و دادهای بزرگ‌مقیاس NCEP می‌توان در شرایط مختلف، میزان بارش و دما را برای آینده پیش‌نمایی کرد.

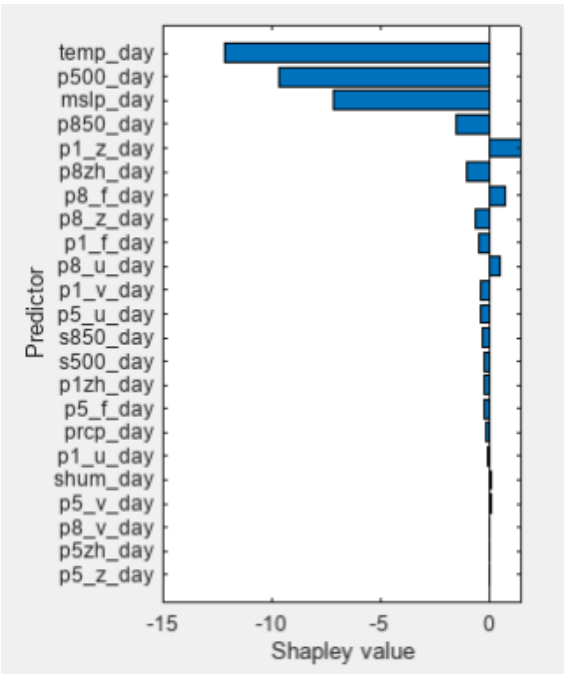
برای اجرای الگوهای ماشین‌های بردار پشتیبان می‌توان از ابزارهای متفاوتی مثل کتابخانه Scikit Learn و بسته e۱۰۷۱ استفاده کرد. کتابخانه Scikit Learn متن‌باز، مفید، پرکاربرد و قدرتمند در زبان برنامه‌نویسی پایتون است که برای اهداف یادگیری ماشین استفاده می‌شود. این کتابخانه برای یادگیری ماشین و الگوسازی آماری داده‌ها همچون طبقه‌بندی، رگرسیون،

1.Tripathi
2. Gocic, at al

خوشه‌بندی و کاهش ابعاد، ابزارهای کاربردی زیادی را فراهم می‌آورد. سادگی، قابلیت استفاده، پوشش گسترده الگوریتمی، ارزیابی الگو، تنظیم مؤلفه، پشتیبانی از تقسیم‌بندی داده، پشتیبانی از تعداد زیادی مسئله و غیره ازجمله ویژگی‌های مهم این کتابخانه است. بسته e۱۰۷۱ در زبان برنامه‌نویسی R نیز برای الگوسازی ماشین بردار پشتیبان استفاده می‌شود. در این بسته توابع کمکی[1] نیز وجود دارد. در ساخت ماشین بردار پشتیبان در زبان R و پایتون از رویکرد مشابهی استفاده شده است.

یافته‌ها

در این مطالعه برای طبقه‌بندی داده‌های اقلیمی با روش ماشین‌های بردار پشتیبان از داده‌های ماهانه دمای متوسط ایستگاه سینوپتیک شیراز از سال ۱۹۶۱ تا ۲۰۱۷ استفاده شد و از خروجی الگو CanESM5 در قالب داده‌های NCEP استفاده‌شده است. داده‌های NCEP شامل ۲۳ مؤلفه بزرگ‌مقیاس اقلیمی است که به‌صورت روزانه از سال ۱۹۶۱ تا ۲۰۱۷ مربوط به منطقه موردمطالعه در دسترس است. داده‌ها در دو دسته آموزش و آزمایش تفکیک شدند. ۷۰ درصد داده‌ها در دسته آموزش و ۳۰ درصد در دسته آزمایش قرار گرفتند. برای طبقه‌بندی دمای متوسط از ۲۳ متغیر پیش‌بینی‌کننده بزرگ‌مقیاس با عنوان داده‌های بزرگ‌مقیاس استفاده شد. داده‌های NCEP برای نمونه شامل میانگین فشار سطح دریا، رطوبت نسبی در ارتفاع ۸۵۰ هکتوپاسکال، میانگین دما در ارتفاع ۲ متری، رطوبت نسبی در ارتفاع ۵۰۰ هکتوپاسکال و غیره است. با تعیین ارتباط بین داده‌های مشاهداتی ایستگاه محلی (دمای متوسط ایستگاه سینوپتیک شیراز) و داده‌های بزرگ‌مقیاس NCEP می‌توان در شرایط مختلف میزان دما را برای آینده پیش‌نمایی کرد. برای انتخاب متغیرهای پیش‌بینی‌کننده غالب، با استفاده از تحلیل مؤلفه اصلی، از بین ۲۳ مؤلفه بزرگ‌مقیاس NCEP ۵ مؤلفه انتخاب شدند. شکل ۵ تحلیل مؤلفه اصلی مربوط به ۲۳ مؤلفه بزرگ‌مقیاس NCEP را نشان می‌دهد. بیشترین میزان اثرگذاری مربوط به ۵ مؤلفه: دمای سطح در ارتفاع ۲ متری (temp_day)، میانگین فشار سطح دریا (mslp_day)، ارتفاع ژئوپتانسیلی تراز ۵۰۰ hPa (p500_day)، ارتفاع ژئوپتانسیلی تراز ۸۵۰ hPa (p850_day) و تاوایی سطح (p1_z_day) است.



شکل ۵. تحلیل مؤلفه اصلی ۲۳ مؤلفه داده‌های بزرگ‌مقیاس NCEP

1.Helper

ماشـین‌های بردار پشتیبان برای تقسـیم‌بندی داده‌ها از انواع مختلف توابع هسـته اسـتفاده می‌کنند. در این پژوهش از چهار نوع توابع خطی، تابع پایه شعاعی، چندجمله‌ای و سیگموید برای طبقه‌بندی اطلاعات استفاده شـد. داده‌ها در دو مرحله آموزش و آزمایش تحت ۴ نوع توابع هسـته متفاوت با هم مقایسه شدند. نتایج ارائه‌شـده در جدول ۱ نشـان می‌دهد که انواع توابع هسته در طبقه‌بندی داده‌های دمای متوسـط ایستگاه سینوپتیک شیراز دقت قابل‌قبولی دارند.

جدول ۱. ارزیابی دقت ٤ نوع توابع هسته در طبقه‌بندی دمای متوسط ایستگاه سینوپتیک شیراز

نوع الگو	مرحله آموزش			مرحله آزمایش		
	R^2	*RMSE*	*MAE*	R^2	*RMSE*	*MAE*
خطی	۰/۹۹	۰/٦۸	۰/٤٦	۰/۹۹	۰/۷۲	۰/۵۲
تابع پایه شعاعی	۰/۹۷	۰/۸۵	۰/۷۹	۰/۹۸	۰/۷۸	۰/٦۸
چندجمله‌ای	۰/۹۸	۰/۷۲	۰/۷۲	۰/۹۸	۰/٦۸	۰/٤٦
سیگموید	۰/۹۸	۰/۹۱	۰/۸۷	۰/۹۷	۰/۹۸	۰/۸٦

جمع‌بندی و توصیه‌ها

پیش‌بینی‌هـا کمک می‌کنند تـا تغییـرات آب و هوایی و الگوهای طبیعت بهتر درک شـوند. درنتیجه، برای حفاظت از منابع طبیعی و جامعه بشری در برابر تغییرات آب و هوائی شدید، می‌توان راهبردهای مناسـبی را برنامه‌ریـزی کرد. پژوهش‌های مرتبـط با پیش‌بینـی تغییرات اقلیـم، درک و عملکـرد بهتری از تغییـرات آینده را به دسـت می‌دهند. با اسـتفاده از داده‌های جمع‌آوری‌شـده و الگوریتم‌های پیچیده، پژوهشگران می‌توانند الگوهای تغییرات آب‌وهوایی را شبیه‌سـازی کـرده و نتایج را برآورد کنند. این نتایج به سـازمان‌ها، سیاست‌گذاران و جامعه کمک می‌کنـد تا در حوزه‌هـای مختلف تصمیمات مناسـبی بگیرند. در پیش‌بینی، طبقه‌بندی اطلاعات اهمیت زیادی دارد. با طبقه‌بندی اطلاعات می‌توان الگوها و روابط مختلف در داده‌ها را شناسـایی و به تحلیل و پیش‌بینی بهتری دست یافت. در طول فرآیند طبقه‌بندی، اطلاعات مشـابه گروه‌بندی شـده و به‌عنوان چارچوب اصلی برای پیش‌بینی استفاده می‌شود. نتایج مربوط به الگوسازی تحت الگوهای ماشین‌های بردار پشتیبان نشان‌دهنده دقت مناسـب این الگو ها در پیش بینی مؤلفه های اقلیمی است.

بـرای بهبود عملکـرد الگوهای ماشـین‌های بردار پشـتیبان در پیش‌بینی اقلیمی، در ادامه چند توصیه ارائه‌شده است:

- ویژگی‌های یک ساختار ماشین‌های بردار پشتیبان به انتخاب نوع تابع هسـته وابسته اسـت. بنابراین، در انتخاب نوع تابع هسته دقت شود؛
- تعیین مؤلفه اصلی، فرآیندی مهم در تحلیل داده‌ها اسـت و کمک می‌کند تا مؤلفه‌های اصلی تحقیق شناسایی و عوامل غیرضـروری حذف شـوند. بنابراین، بایـد در تعیین مؤلفه اصلی در پیش‌بینی اقلیمی تلاش شـود تا با تعداد مؤلفه‌های بهینه بهترین برازش و پیش‌بینی به دست آید؛
- الگوهای مختلف ماشـین‌های بردار پشـتیبان مقایسـه شده و عملکـرد آن‌ها ارزیابی شـود تا بهینه‌ترین سـاختار برای پیش‌بینی مؤلفه‌های اقلیمی شناسایی شوند.

منابع

خاشعی سیوکی، عباس، شهیدی، علی، پور رضا بیلندی، محسن، امیرآبادی زاده، مهدی و جعفر زاده، احمد. (۱۳۹۷). بررسی عملکرد روش‌های ANN و SVR در ریزمقیاس نمایی بارش روزانه مناطق خشک. *تحقیقات آب‌وخاک ایران (علوم کشاورزی ایران)*، ۴۹(۴).

صفوی، ح. (۱۳۸۸). هیدرولوژی مهندسی. *انتشارات ارکان دانش*. چاپ اول، ۶۰۷ ص.

Chen, H. Wu, W. & Liu, H. B. (2016). Assessing the relative importance of climate variables to rice yield variation using support vector machines. *Theoretical and Applied Climatology,* 126, 105-111.

Gocic, M. Shamshirband, S. Razak, Z. Petković, D. Ch, S. & Trajkovic, S. (2016). Long-term precipitation analysis and estimation of precipitation concentration index using three support vector machine methods. *Advances in Meteorology, 2016.*

Guobin, F. Stephen, P. C. Francis, H.S.C. Jin, T. Hongxing, Z. Andrew, J. F. Wenbin, L. Sergey, K. 2013.Modelling runoff with statistically downscaled daily site, gridded and catchment rainfall series. *Journal of Hydrology 492* (2013) 254-265.

Haiyun Panda D K. Mishra A. Jena S K. James B K. Kumar A.2007: *The Influence of Drought and Anthropogenic Effects on Groundwater Levels in Orissa,* India. 140-153.

Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, *the Elements of Statistical Learning,* Springer. (2001). ISBN 0-387-95284-5.

Kalra, A. Ahmad, S. (2012). Estimating ANNual precipitation for the Colorado River Basin using oceanic -atmospheric oscillations. *Water Resources Research*, 48(6)

Lu, Y. & Qin, X. S. (2014). A coupled K -nearest neighbour and Bayesian neural network model for daily rainfall downscaling. *International Journal of Climatology*, 34(11), 3221 -3236.

Min H. Jae, & Lee C. Young (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications,* 28, 603-614.

Mirarabi, A. Nassery, H. R. Nakhaei, M. Adamowski, J. Akbarzadeh, A. H. & Alijani, F. (2019). Evaluation of data-driven models (SVR and ANN) for groundwater-level prediction in confined and unconfined systems. *Environmental Earth Sciences,* 78(15), 1-15.

Nsubuga, F.W.N. Botai, O.J. Olwoch, J.M. Dew Rauten bach, C.J.Yvette, B. & Adebayo, O.A. (2014). The nature of rainfall in the maindrainage sub-basins of Uganda. *Hydrological Sciences Journal,* 59 (2), 278-299. DOI: 10.1080/02626667.2013.804188.

Sachindra, D. A. Huang, F. Barton, A. & Perera, B. J. C. (2013). Least square support vector and multi - linear regression for statistically downscaling general circulation model outputs to catchment streamflows. *International Journal of Climatology*, 33(5), 1087 -1106.

Shin S. Kyung, Lee S. Taik, & Kim J. Hyun (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications,* 28, 127-135.

Stuart J, Russell, Peter Norvig, Artificial Intelligence: A Modern Approach (Third Ed). Prentice Hall. (2010). ISBN 9780136042594.

Tripathi, S. Srinivas, V. V. & Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of hydrology*, 330(3-4), 621-640.

Vapnik, V. (1995). *The nature of statistical learning theory.* NY: Springer-Verlag.

Vapnik, Vladimir, 2000, *The nature of statistical learning theory.* Springer Science & Business Media.

Wen, X. Si, J. He, Z. Wu, J. Shao, H. & Yu, H. (2015). Support-vector-machine-based models for modeling daily reference evapotranspiration with limited climatic data in extreme arid regions. *Water resources management*, 29, 3195-3209.

Journal of Agricultural Information sciences and Technology
Volume 7, Issue 1, Serial Number 13, Spring and Summer 2024
jaist.areeo.ac.ir

The use of support vector machines in classification of climatic data

Seyed Masoud Soleimanpour[1], Parsa Haghighi[2]

1. Associate Professor, Soil Conservation and Watershed Management Research Department, Fars Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization (AREEO), Shiraz, Iran. **(Corresponding author). Email:** *m.soleimanpour@areeo.ac.ir*

2. Masters, Soil Conservation and Watershed Management Research Department, Fars Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization (AREEO), Shiraz, Iran **Email:** *p.haghighi@areeo.ac.ir*

Abstract

Identifying, predicting and managing crisis in a climate structure is of great importance. Models are used as practical tools for understanding complex systems and simulating and predicting their behavior. Support vector machines are one of the supervised learning methods used for classification and regression. Support vector machines are able to detect hidden patterns and respond to complex changes in climate data. In this article, the structure of the support vector machine method and its application in climate data classification are presented. The characteristics of the structure of support vector machines are related to the selection of the kernel function type, so sufficient care must be taken in the selection of the kernel function type and on the other hand, PCI in climate forecasting is an important step in climate forecasting in order to make the best fit between forecasting and predicted data with the optimal number of parameters.

Keywords: Support vector machines, Classification of climatic data, kernel functions, PCI.