دوازدهمین کنفرانس هیدرولیک ایران
گروه مهندسی آبیاری و آبادانی
پردیس کشاورزی و منابع طبیعی
۷ و ۸ آبان‌ماه ۱۳۹۲

مقایسه کارایی الگوریتم‌های ماشین بردار پشتیبان و K نزدیک‌ترین همسایگی برای طبقه‌بندی کیفیت آب زیرزمینی

فرشته مدرسی
دانشجوی دکتری مهندسی منابع آب- دانشگاه تهران
شهاب عراقی‌نژاد
استادیار دانشگاه تهران

چکیده

امروزه در کشور ما آبهای زیرزمینی یکی از منابع مهم آبی به‌شمار می‌روند که برای تأمین نیاز شرب مورد استفاده قرار می‌گیرند. از این‌رو تعیین کیفیت آنها کاملا ضروری است. یکی از راه‌های تعیین کیفیت آب، محاسبه شاخص‌های کیفی است که کاری بسیار زمانبر و نیازمند به نظرات کارشناسی است. راهکار دیگر برای طبقه‌بندی کیفی آب، استفاده از الگوریتم‌های طبقه‌بندی است که می‌توان با یکبار آموزش این الگوریتم‌ها، کیفیت هزاران نمونه آب را به آسانی تعیین نمود. ولی سئوال این است که آیا و تا چه اندازه، نتایج این الگوریتم‌ها با یکدیگر متفاوت است و کدام الگوریتم دارای نتایج صحیح‌تری است. برای پاسخگویی به این سئوالات، در مقاله حاضر، کارایی دو الگوریتم پرکاربرد ماشین بردار پشتیبان و K نزدیک‌ترین همسایگی با استفاده از روش صحت سنجی متقاطع ۵ لایه برای طبقه‌بندی کیفیت آب زیرزمینی بررسی و مقایسه شده است. بدین منظور از آمار کیفیت آب ۱۰۰ چاه مشاهداتی در دشت تهران در سال آبی ۸۳-۱۳۸۲ استفاده شده است که بر اساس دو آلاینده نیترات و کلر و با استفاده از شاخص CCME طبقه‌بندی شده‌اند. نتایج این تحقیق نشان داد که الگوریتم ماشین بردار پشتیبان به خوبی آموزش پذیر است به طوریکه در مراحل واسنجی و صحت سنجی بدون خطا بوده‌است و کارایی بسیار بهتری نسبت به الگوریتم K نزدیک‌ترین همسایگی برای طبقه‌بندی کیفیت آب دارد.

واژه‌های کلیدی: طبقه بندی کیفی، آب زیرزمینی، ماشین بردار پشتیبان، K نزدیک‌ترین همسایه، CCME.

مقدمه

کیفیت آب یکی از معیارهای اساسی در مدیریت منابع آب است و نقش مهمی را در تعیین سیاست‌های بهره برداری از منابع آب ایفا می‌نماید. امروزه در کشور ما منابع آب زیرزمینی به عنوان یکی از منابع اصلی آب در تأمین نیازهای مختلف بخصوص نیاز شرب، به طور گسترده مورد استفاده قرار می‌گیرند. از آنجایی که کیفیت آب برای تأمین نیاز شرب یکی از فاکتورهای اصلی برای تعیین قابلیت استفاده از منبع آبی مورد نظر است، بنابراین لزوم طبقه‌بندی کیفی آب‌های زیرزمینی احساس می‌شود. برای تعیین کیفیت آب، شاخص‌های کیفی متنوعی ارائه شده‌اند. بسیاری از شاخص‌های کیفی نظیر شاخص اتحادیه بهداشت جهانی (National Sanitation Federation (NSF)) و نیز شاخص ISQA (L' Index Simplificat de la Qualitat de l'Aigua) از روش وزن‌دهی به متغیرهای کیفی استفاده می‌کنند. حال آنکه تعریف وزن هر متغیر کاری پیچیده و نیازمند به نظرات کارشناسی است. شاخص کیفی دیگری به نام CCME (Canadian Council of Ministers of the Environment) توسط وزارت محیط زیست، پارک‌ها و زمین کانادا معرفی شد که به جای استفاده از روش وزن‌دهی به متغیرها، از تعریف سه فاکتور با نام‌های گستره یا دامنه (Scope)، فراوانی (Frequency) و بزرگی (Amplitude) برای تعیین کیفیت آب بهره می‌برد. با این وجود، محاسبه کلاس‌های کیفی آب زیرزمینی به کمک این روش نیز برای

هر نمونه آب بسیار زمانبر بوده و با مشکلات خاص خود همراه است. بنابراین، برای رهایی از انجام محاسبات پیچیده برای هر نمونه آب، روش‌های آموزش آماری و ماشینی می‌توانند برای طبقه‌بندی کیفی آب مورد استفاده قرار گیرند. با این روش، فقط کافی است که مقدار متغیرهای کیفی نمونه آب مورد نظر به مدل داده شود تا کلاس کیفی آن بلافاصله تعیین شود.

برای انجام فرآیند طبقه‌بندی داده‌ها، طیف وسیعی از روش‌های آموزش آماری و ماشینی ارائه شده‌اند نظیر: آنالیز جداکننده خطی (LDA)، آنالیز جداکننده درجه دو (QDA)، آنالیز جداکننده تنظیم شده (RDA)، درخت‌های طبقه‌بندی، مدل‌سازی مستقل ملایم از قیاس کلاس‌ها (SIMCA)، K نزدیک‌ترین همسایگی (KNN)، شبکه‌های عصبی (NN) و ماشین‌های بردار پشتیبان (SVM). ولی سئوال این است که آیا و تا چه اندازه، نتایج این الگوریتم‌ها با یکدیگر متفاوت است و کدام الگوریتم دارای نتایج صحیح‌تری است. برای مقایسه این روش‌ها، مطالعات مختلفی در رشته‌های گوناگون صورت گرفته است. به‌طور مثال:

Werther و همکاران در سال ۱۹۹۴ برای طبقه‌بندی طیف‌های جرمی (mass spectra) براساس ۵۴ خصوصیت، از چهار روش KNN ، آنالیز جداکننده (DA)، SIMCA و NN استفاده کردند و نتایج آنها را با یکدیگر مقایسه نمودند. نتایج این تحقیق نشان داد که از یک سو، روش شبکه عصبی دارای بهترین جواب است و از سوی دیگر، تمامی این روش‌ها برای طبقه‌بندی براساس تعداد اندکی خصوصیت کار می‌کنند.

Byvatov و همکاران (۲۰۰۳) دو روش SVM و ANN را برای طبقه‌بندی دوتایی دارو و نادارو (drug/nondrug) مورد ارزیابی و مقایسه قرار دادند. آنها در این تحقیق، یک شبکه عصبی استاندارد پیشخور با تعداد نرون مشخص در لایه مخفی را با SVM مقایسه کردند. نتایج آنها نشان داد که روش SVM دارای صحت کمی بالاتر و خطای استاندارد کمتر برای پیش‌بینی دسته‌ها نسبت به روش ANN است. آنها معتقد بودند که روش SVM دارای دو خصوصیت مفید است که سبب برتری آن نسبت به روش شبکه‌های عصبی می‌شود. این دو روش عبارتند از: ۱- پراکندگی راه حل (sparseness of the solution) : بدین معنا که SVM فقط به بردارهای پشتیبان وابسته است و همه داده‌ها بر تابع طبقه بندی کننده اثر نمی‌گذارند، در حالیکه برای تعداد زیادی از شبکه‌های عصبی این طور نیست. ۲- از آنجایی که SVM از توابع کرنل بهره می‌برد، این روش می‌تواند برای طبقه‌بندی داده‌ها براساس تعداد زیادی از ویژگی‌ها عمل نماید.

Lee and Park در تحقیقی در سال ۲۰۰۶ برای تشخیص جفت پروتئین‌های برهم کنش کننده از سایرین، الگوریتم‌های مختلف طبقه‌بندی را مورد ارزیابی و مقایسه قرار دادند. نتایج آنها نشان داد که اولاً: این الگوریتم‌ها ابزارهای مناسبی برای تشخیص جفت پروتئین‌های برهم کنش کننده است و ثانیا: الگوریتم‌های KNN و درخت تصمیم (Decision Tree) بهترین کارایی را نسبت به سایرین دارا می‌باشند.

Tsuta و همکاران در سال ۲۰۰۹ برای تعیین آفت کش روی برگ اسفناج، دو روش SVM و LDA را مورد مقایسه و ارزیابی قرار دادند. در این تحقیق، آنها سه نوع تصویر فلورسنس را طبقه‌بندی کردند. نتایج این تحقیق نشان داد که نرخ خطای دو روش SVM و LDA به ترتیب برابر است با ۹/۹٪ و ۱۸/۸٪. بنابراین روش SVM بر روش LDA برتری دارد.

Balabin و همکاران (۲۰۱۰) برای طبقه‌بندی بنزین براساس داده‌های طیف‌بینی زیرقرمز نزدیک (Near infrared spectroscopy)، نه روش طبقه‌بندی چند متغیره را مورد ارزیابی و مقایسه قرار دادند که عبارتند از: ۱- LDA، ۲- QDA، ۳- RDA، ۴- SIMCA، ۵- طبقه‌بندی حداقل مربعات جزئی (PLS)، ۶ - KNN، ۷- SVM، ۸- شبکه عصبی احتمالاتی (PNN) و ۹- ANN-MLP . در این تحقیق سه مجموعه داده شامل ۴۵۰، ۴۱۵ و ۳۴۵ طیفی برای طبقه بندی به ۳، ۶ و ۳ کلاس به ترتیب براساس منبع و نوعشان استفاده شدند. نتایج بدست آمده از این تحقیق نشان می‌دهد که سه روش KNN, SVM, PNN نسبت به سایر روش‌ها برای طبقه‌بندی مناسب‌ترند.

Tamouk و Allahakbari در سال ۲۰۱۲ مقایسه‌ای میان ۵ روش طبقه‌بندی شامل: ۱- KNN ، ۲- PNN ، ۳- K نزدیک‌ترین همسایه مرکزی، ۴- نزدیک‌ترین همسایه انطباقی جداکننده (DANN) و ۵- نزدیک‌ترین خط ویژگی(NFL) انجام دادند. در این تحقیق برای مقایسه این روش‌ها از داده‌های عنبیه و قمر مصنوعی استفاده شده است. نتایج این تحقیق نشان داد که دو روش KNN و KNCN دارای صحت بیشتری در نتایجشان هستند.

همان‌طور که در سابقه تحقیقات صورت گرفته مشاهده می‌شود، طیف وسیعی از روش‌های طبقه‌بندی در رشته‌ها و زمینه‌های گوناگون مورد ارزیابی و مقایسه قرار گرفته‌اند. با این وجود، در زمینه طبقه‌بندی کیفیت آب، مطالعات بسیار معدودی درباره کارایی روش‌های طبقه‌بندی صورت گرفته است. تنها تحقیق صورت گرفته در این مورد، تحقیقی است که توسط Chen و همکاران در سال ۲۰۰۴ انجام شد. آنها برای طبقه‌بندی الگوهای فضایی رنگ اقیانوس که نشان دهنده کیفیت آب هستند به پنج کلاس بر اساس چهار باند طیفی، از سه روش طبقه بندی باناظر شامل حداکثر درست‌نمایی (MLH) ، ANN و SVM استفاده کردند. نتایج این تحقیق نشان داد که دو روش ANN و SVM برای تعداد داده آموزشی کم نسبت به روش MLH مناسب‌ترند و از میان این دو روش نیز، روش SVM مناسب‌تر است. به طوری که صحت روش‌های MLH ، ANN و SVM به ترتیب برابر بود با ۷۸/۳ ٪ ، ۸۲/۶ ٪ و ۹۱/۳ ٪ ولی هر سه روش الگوهای مکانی مشابهی را ایجاد کردند.

سابقه تحقیقات صورت گرفته در زمینه بکارگیری الگوریتم‌های طبقه‌بندی نشان می‌دهند که روش ماشین بردار پشتیبان (SVM) و K نزدیک‌ترین همسایگی (KNN) دارای بیشترین کارایی و صحیح‌ترین نتایج بوده‌اند. بنابراین با توجه به این امر و نیز از آنجایی که روش SVM یک روش جدید برای طبقه‌بندی بر اساس بردارهای پشتیبان است و نیز روش KNN یک روش قدیمی و البته قدرتمند برای طبقه‌بندی به شمار می‌رود، در این تحقیق این دو روش برای طبقه‌بندی کیفیت آب زیرزمینی مورد ارزیابی و مقایسه قرار گرفته‌اند. همچنین، به دلیل اینکه شاخص کیفی CCME یک شاخص قابل اعتماد است و به نظریات شخصی کاربران وابسته نیست، در تحقیق حاضر این شاخص برای آموزش دو روش طبقه‌بندی مذکور جهت طبقه‌بندی کیفی آب مورد استفاده قرار گرفته است.

در سایر بخش‌های این مقاله، ابتدا منطقه مطالعاتی معرفی شده است. پس از آن، روش محاسبه شاخص CCME و نیز الگوریتم‌های طبقه‌بندی ارائه شده‌اند و در نهایت، نتایج تحقیق و جمع بندی و مراجع به ترتیب ارائه می‌شوند.

منطقه مطالعاتی

رشد فزاینده جمعیت در استان تهران سبب شده است که در این استان، علاوه بر آب‌های سطحی، از آب‌های زیرزمینی به عنوان یکی از منابع اصلی آب استفاده شود. با این‌حال، دفع فاضلاب از طریق چاه‌های جذبی سبب آلودگی آب‌های زیرزمینی در این استان شده است. بدین دلیل، در تحقیق حاضر، کارایی الگوریتم‌های طبقه‌بندی SVM و KNN برای طبقه بندی کیفیت آب آبخوان اصلی دشت تهران مورد تحقیق و بررسی قرار گرفته است. آبخوان اصلی تهران بین رودخانه‌های کن و سرخه حصار قرار دارد و از شمال به عباس آباد و از جنوب به ارتفاعات کهریزک محدود می‌شود. این آبخوان از نوع آزاد می‌باشد. نیترات و کلر دو آلاینده اصلی این آبخوان هستند که در اثر وجود چاه‌های جذبی فاضلاب وارد این آبخوان شده‌اند. بنابراین، طبقه‌بندی کیفی آب این آبخوان براساس دو آلاینده مذکور صورت گرفت. داده‌های مورد استفاده در این تحقیق از آمار ۱۰۰ چاه مشاهداتی در سراسر آبخوان در دوره ۸۳-۱۳۸۲ به دست آمده است.

شاخص کیفیت آب CCME

این شاخص در سال ۱۹۹۵ توسط وزارت محیط زیست، پارک‌ها و زمین کانادا برای ارزیابی کیفیت منابع آب ایجاد شد. در این روش، متغیرهای کیفی آب نسبت به یک حد معین سنجیده می‌شوند و میزان تجاوز از آن حد مشخص می‌شود. این حد می‌تواند رهنمودهای توصیه شده به منظور حفظ قابلیت استفاده از آب برای مصارف مورد نظر باشد و یا هر استانداردی که برای مصارف مختلف آب مطرح است.

برای محاسبه این شاخص، پس از مشخص نمودن استانداردهای کیفی مدنظر و تعریف متغیرهای مورد نظر، لازم است که سه فاکتور اصلی مورد استفاده در تعیین این شاخص شامل گستره یا دامنه (Scope)، فراوانی (Frequency) و بزرگی (Amplitude) محاسبه شوند. این سه فاکتور به ترتیب با F_1، F_2 و F_3 نشان داده می‌شوند و به صورت زیر محاسبه می‌شوند(CCME,2001):

۱- **فاکتور F_1 (Scope):** این فاکتور تعیین کننده درصدی از متغیرهاست که از استانداردهای مربوط به خود، دست کم یک‌بار در طول دوره زمانی موردنظر تخطی کرده باشند و از رابطه زیر محاسبه می‌شود:

$$F_1 = \left(\frac{Number\ of\ \ failed\ \text{variables}}{Total\ number\ of\ \text{variables}} \right) \times 100 \qquad (۱)$$

۲- **فاکتور F_2 (Frequency):** این فاکتور تعیین کننده درصدی از آزمایش‌های مجزای انجام پذیرفته است که از استانداردهای مربوط به خود تخطی کرده باشند و از رابطه زیر محاسبه می‌شود:

$$F_2 = \left(\frac{Number\ of\ \ failed\ tests}{Total\ number\ of\ tests} \right) \times 100 \qquad (۲)$$

۳- **فاکتور F_3 (Amplitude):** این فاکتور تعیین کننده مقدار تجاوز از استانداردهای مربوطه در آزمایش‌های تخطی یافته است. برای تعیین آن، ابتدا شاخص (Excursion) و پس از آن مقدار nse و در نهایت فاکتور F_3 به صورت زیر محاسبه می‌شود.

$$excursion_i = \left(\frac{FailedTestValue_i}{Objective_j} \right) - 1 \Rightarrow nse = \frac{\sum_{i=1}^{n} excursion_i}{number\ of\ tests} \Rightarrow F_3 = \left(\frac{nse}{0.01 nse + 0.01} \right) \qquad (۳)$$

پس از محاسبه مقدار سه فاکتور مذکور، شاخص کیفیت آب CCME از جمع برداری سه فاکتور بر اساس رابطه زیر محاسبه می‌شود:

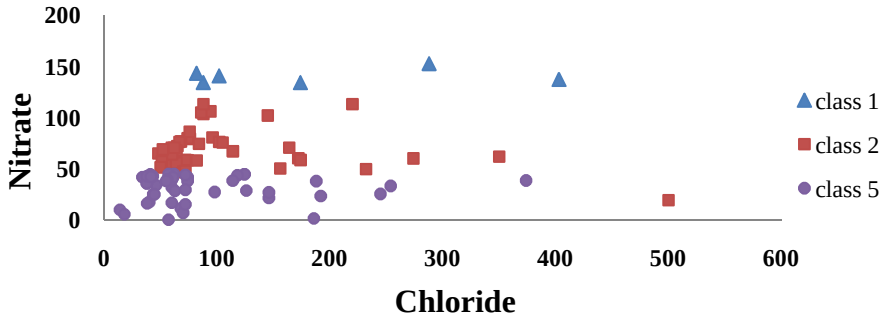
$$CCME = 100 - \left(\frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \right) \qquad (۴)$$

مقدار این شاخص از صفر تا ۱۰۰ متغیر است به طوری‌که مقدار صفر و ۱۰۰ به ترتیب نشان دهنده بدترین و بهترین کیفیت هستند. کلاس‌بندی کیفیت آب بوسیله این استاندارد به صورت جدول شماره ۱ است:

جدول۱- کلاس‌بندی کیفیت آب بر اساس استاندارد CCME

کلاس	ضعیف	بحرانی	متوسط	خوب	عالی
محدوده کلاس	۰-۴۴	۴۵-۶۴	۶۵-۷۹	۸۰-۹۴	۹۵-۱۰۰

با استفاده از این استاندارد، در تحقیق حاضر کیفیت آب ۱۰۰ چاه مشاهداتی بر اساس دو آلاینده نیترات و کلر محاسبه و تعیین شد. بر این اساس، کیفیت آب چاه‌ها در سه کلاس ضعیف، بحرانی و عالی طبقه بندی شد و کیفیت آب هیچ چاهی در دو کلاس متوسط و خوب قرار نگرفت. در شکل ۱، کلاس‌بندی چاه‌های مشاهداتی بر اساس این استاندارد نشان داده شده است.



شکل ۱- کلاس‌بندی چاه‌های مشاهداتی بر اساس استاندارد CCME

الگوریتم ماشین بردار پشتیبان (Support Vector Machine=SVM)

ماشین بردار پشتیبان یکی از روش‌های یادگیری بانظارت است که برای طبقه‌بندی و رگرسیون استفاده می‌شود. الگوریتم SVM، در رده الگوریتم‌های تشخیص الگو، دسته بندی می‌شود. این الگوریتم اولین بار توسط بوسر و گایون در سال ۱۹۹۲ معرفی شد و ساختار آن در سال ۱۹۹۵ توسط وپنیک توسعه یافت (Vapnik,1995).

الگوریتم پایه SVM برای طبقه‌بندی به دو دسته ایجاد شده است. در این الگوریتم، هدف، یافتن یک ابرصفحه بهینه با بیشترین حاشیه بر اساس داده‌های آموزشی است به نحوی که بتواند داده‌ها را به دو دسته تقسیم نماید. فرمول‌بندی کلی الگوریتم SVM برای

مجموعه‌ای از داده‌های جفت شده به صورت $(x_i, y_i), i = 1,2,...,n$ تعریف می‌شود که در آن، x_i برداری با m ویژگی و $y_i \in \{-1,1\}$ کلاس مرتبط با هر x_i است. برای یافتن ابرصفحه بهینه، این الگوریتم، مسئله بهینه سازی زیر را حل می‌نماید:

$$\begin{aligned} &Maximize \quad \sum_{i=1}^{n} \sim_i - \frac{1}{2} \sum_{i,j=1}^{n} \sim_i \sim_j y_i y_j . k(x_i, x_j) \\ &Subject\ to: \quad \sim_i \geq 0 \ , \quad \forall i \in \{1,...,n\} \\ &\qquad \sum_{i=1}^{n} \sim_i y_i = 0 \end{aligned} \qquad (۵)$$

در رابطه ۵، $k(x_i, x_j)$ تابع کرنل نام دارد و برابر است با $(\text{¤}(x_i).\text{¤}(x_j))$ که در آن، تابع ¤ بردارهای آموزشی x_i را به فضایی با ابعاد بالاتر نگاشت می‌کند. توابع کرنل دارای انواع مختلفی نظیر خطی، چندجمله‌ای، نمایی و هذلولی هستند.

پس از تعیین مقدار بهینه پارامتر ~، تابع تصمیم طبقه‌بندی برای j امین عنصر به صورت زیر محاسبه می‌شود:

$$f(x_j) = sign\left(\sum_{i=1}^{n} \sim_i y_i . k(x_i x_j) + b \right) \qquad (۶)$$

همان‌طور که پیش‌تر بیان شد، الگوریتم پایه SVM برای طبقه‌بندی دو دسته‌ای ایجاد شده است. برای طبقه‌بندی چنددسته‌ای از دو رویکرد "مقایسه دوبه دو" و "مقایسه یکی با همه" می‌توان استفاده نمود که در تحقیق حاضر، از رویکرد دوم استفاده شده است.

الگوریتم K نزدیک‌ترین همسایگی (KNN)

الگوریتم KNN یک روش رایج برای طبقه‌بندی است و براساس سنجش فاصله می‌باشد. در این روش ابتدا داده‌های آموزشی و کلاس‌های متناظر با آنها درنظر گرفته می‌شود. سپس برای طبقه‌بندی یک نمونه جدید، فاصله آن تا هر یک از نمونه‌های آموزشی محاسبه می‌شود و K تا از نزدیک‌ترین همسایه‌ها (نمونه‌های آموزشی) انتخاب می‌شوند. در نهایت، نمونه جدید در کلاسی قرار می‌گیرد که اکثر نمونه‌های موجود در K همسایه‌اش در آن کلاس هستند. برای تعیین فاصله هر نمونه جدید از نمونه‌های مشاهداتی از روش فاصله اقلیدسی استفاده می‌شود. در این روش، برای دو نمونه با n ویژگی یعنی $X=(x_1, x_2, \ldots, x_n)$ و $Y=(y_1, y_2, \ldots, y_n)$، فاصله اقلیدسی از رابطه زیر محاسبه می‌شود (Yang Su, 2011) :

$$dist(X,Y) = \sqrt{\sum_{i=1}^{n} c_i (x_i - y_i)^2} \qquad (۷)$$

در این رابطه، c_i وزن هر ویژگی است که در تحقیق حاضر به علت یکسان بودن ارزش آلاینده‌ها در تعیین شاخص کیفیت آب CCME، این ضریب برای هردو آلاینده یکسان درنظر گرفته شده است. در این الگوریتم، تعیین مقدار پارامتر K از اهمیت ویژه‌ای برخوردار است. بدین دلیل در تحقیق حاضر از روش صحت سنجی متقاطع (Cross Validation) برای تعیین K بهینه استفاده خواهد شد.

نتایج و بحث

همان‌طور که پیش‌تر بیان شد، هدف از انجام این تحقیق، مقایسه کارایی دو روش SVM و KNN برای طبقه‌بندی کیفی آب زیرزمینی است. بدین منظور، داده‌های کیفی ۱۰۰ چاه مشاهداتی به ۵ گروه ۲۰ تایی تقسیم شد. سپس، ۵ مجموعه داده ایجاد شد که در هر کدام از آنها، ۴ گروه ۲۰ تایی برای آموزش و یک گروه دیگر برای آزمودن الگوریتم‌ها مورد استفاده قرار گرفت. در واقع یک صحت سنجی متقاطع ۵ لایه (5-fold cross validation) برای ارزیابی اثر مقدار داده‌ها (وجود یا عدم وجود مقادیر بیشینه و کمینه در داده‌های آموزشی) بر کارایی الگوریتم‌ها صورت گرفت. در ادامه، ابتدا نتایج بدست آمده از هر الگوریتم ارائه می‌شود و در نهایت، نتایج آنها با یکدیگر مقایسه خواهد شد.

نتایج روش ماشین بردار پشتیبان (SVM)

همان‌طور که در توضیح روش SVM بیان شد، توابع کرنل مختلفی برای آموزش این الگوریتم وجود دارد. همچنین، برای تنظیم برازش هر یک از این توابع بر داده‌های آموزشی، پارامتری به نام C وجود دارد که لازم است مقدار آن توسط کاربر تعیین شود. با توجه به اینکه در تحقیق حاضر از رویکرد "مقایسه یکی با همه" برای آموزش این الگوریتم استفاده شده است و اینکه کیفیت داده‌های مشاهداتی بر اساس استاندارد CCME در سه کلاس ۱، ۲و ۵ قرار گرفتند(کیفیت آب هیچ چاهی در کلاس‌های ۳ و ۴ قرار نگرفت)، این الگوریتم سه بار (یک‌بار برای هر کلاس) آموزش دید و در هر بار، تمامی توابع کرنل با مقادیر مختلف C آموزش دیدند. در نهایت، ساده‌ترین تابع کرنل با کم‌ترین مقدار C به ازای کمترین مقدار خطا برای هر کلاس به عنوان مناسب‌ترین تابع و پارامتر انتخاب شدند. به‌طور مثال، اگر توابع کرنل خطی و درجه دو برای کلاس ۱ دارای خطای یکسان بودند، تابع خطی انتخاب و اگر مقادیر C برابر ۱۰۰ و ۱۰۰۰ خطای یکسان داشتند، مقدار ۱۰۰ انتخاب شده است. تمام این فرآیند برای هر یک از ۵ مجموعه داده انجام شده است. نوع تابع کرنل بهینه و مقدار C بهینه برای هر کلاس به ازای هر یک از مجموعه داده‌ها در جدول شماره ۲ آورده شده است.

جدول ۲- نوع تابع کرنل و مقدار C بهینه برای هر کلاس به ازای هر یک از مجموعه داده‌ها

مجموعه داده‌ها	کلاس ۱		کلاس ۲		کلاس ۵	
	نوع تابع کرنل	مقدار C	نوع تابع کرنل	مقدار C	نوع تابع کرنل	مقدار C
۱	خطی	۱۰	درجه ۲	۱۰۰۰۰	درجه ۲	۱۰۰۰
۲	خطی	۱۰	درجه ۲	۱۰۰	درجه ۲	۱۰۰۰
۳	خطی	۱۰	درجه ۲	۱۰۰	درجه ۲	۱۰۰۰
۴	خطی	۱۰	درجه ۲	۱۰۰۰۰۰	درجه ۲	۱۰۰۰
۵	خطی	۱۰	درجه ۲	۱۰۰	درجه ۲	۱۰۰۰

همان‌طور که در جدول ۲ مشاهده می‌شود، نوع تابع کرنل بهینه برای هر کلاس به ازای تمامی ۵ مجموعه داده یکسان است به طوری که تابع خطی متعلق به کلاس ۱ و تابع درجه ۲ متعلق به کلاس‌های ۲ و ۵ است. این امر نشان می‌دهد که برای آموزش این الگوریتم، ساختار کلی داده‌های هر کلاس در تعیین نوع تابع بهینه آن کلاس اثر دارند و این الگوریتم نسبت به مقدار داده‌ها حساس نیست. همچنین، مقدار C بهینه برای کلاس‌های ۱ و ۵ به ازای تمامی مجموعه داده‌ها یکسان است و به ترتیب برابر با ۱۰ و ۱۰۰۰ می‌باشد؛ ولی برای کلاس ۲ مقدار C بهینه برای تمامی مجموعه داده‌ها یکسان نیست. از آنجایی که میزان نامنظمی در پراکندگی داده‌ها بر مقدار پارامتر C اثر می‌گذارد و با افزایش نامنظمی، مقدار آن افزایش می‌یابد، می‌توان نتیجه گرفت که داده‌های کلاس ۱ دارای نامنظمی کمتری نسبت به داده‌های کلاس ۵ هستند ولی میزان این نامنظمی در تمامی مجموعه داده‌ها یکسان است. همچنین، میزان نامنظمی در پراکندگی داده‌های کلاس ۲ در مجموعه داده‌های مختلف، متفاوت است و گروه ۴ دارای بیشترین نامنظمی می‌باشد.

براساس نوع تابع و مقادیر C موجود در جدول ۲، الگوریتم SVM اجرا شد و نرخ خطای (٪) مدل برای هر مجموعه داده در مراحل واسنجی و صحت سنجی از تقسیم تعداد داده‌های اشتباه طبقه‌بندی شده به تعداد کل داده‌ها محاسبه شد که نتایج آن در جدول ۳ آورده شده است.

جدول ۳- نرخ خطای الگوریتم SVM در مراحل واسنجی و صحت سنجی (٪)

مجموعه داده‌ها	واسنجی	صحت سنجی
۱	۰	۰
۲	۰	۱۵
۳	۰	۵
۴	۰	۰
۵	۰	۵

همان‌طور که در جدول ۳ مشاهده می‌شود، نرخ خطای الگوریتم SVM در مرحله واسنجی برای تمامی مجموعه داده‌ها برابر صفر

است که نشان می‌دهد این الگوریتم در آموزش پذیری بسیار منعطف است و به خوبی آموزش می‌پذیرد. همچنین دو مجموعه ۱ و ۴ در مرحله صحت سنجی نیز بدون خطا هستند که این امر کارایی بسیار خوب این الگوریتم را برای این دو مجموعه نشان می‌دهد. همچنین، نرخ خطا در این مرحله برای سایر مجموعه‌ها نیز کوچک و برابر با ۵٪ (یک خطا) برای مجموعه‌های ۳ و ۵ و ۱۵٪ (سه خطا) برای مجموعه ۲ است. به طور کلی از نتایج بدست آمده می‌توان نتیجه‌گیری کرد که الگوریتم SVM با انواع تقسیم‌بندی داده‌ها سازگاری بالایی دارد و می‌تواند به خوبی واسنجی و صحت سنجی شود.

نتایج روش K نزدیک‌ترین همسایگی (KNN)

همان‌طور که در توضیح روش KNN بیان شد، مقدار پارامتر K در کارایی این الگوریتم از ارزش بالایی برخوردار است. بنابراین، در تحقیق حاضر ابتدا با استفاده از روش صحت سنجی متقاطع مقدار K بهینه تعیین شد و سپس براساس این مقدار، الگوریتم برای هر یک از مجموعه داده‌های ۵ گانه، واسنجی و صحت سنجی شد. مقدار K بهینه و نرخ خطای حاصل از اجرای مدل در هر یک از مراحل واسنجی و صحت سنجی در جدول ۴ آورده شده است.

جدول ۴- مقدار K بهینه و نرخ خطای اجرای الگوریتم KNN در مراحل واسنجی و صحت سنجی(٪)

مجموعه داده‌ها	مقدار بهینه پارامترK	واسنجی	صحت سنجی
۱	۲	۰	۵
۲	۲	۰	۱۵
۳	۱۰	۸/۷۵	۲۵
۴	۲۲	۱۰	۲۵
۵	۳	۶/۲۵	۱۵

همان‌طور که در جدول ۴ مشاهده می‌شود، مقدار بهینه پارامتر K برای مجموعه ۱ و۲ برابر با ۲، برای مجموعه ۳ برابر با ۱۰، برای مجموعه ۴ برابر با ۲۲ و برای مجموعه ۵ برابر با ۳ است. مقادیر بزرگ K برای مجموعه‌های ۳ و ۴ نشان دهنده نامنظمی پراکندگی داده‌ها در این دو مجموعه است که این امر سبب الگوپذیری سخت الگوریتم می‌شود. در نتیجه مشاهده می‌شود که این دو مجموعه داده در مراحل واسنجی و صحت سنجی نیز دارای بیشترین تعداد خطا هستند به طوری‌که نرخ خطا در مرحله واسنجی برای مجموعه ۳ برابر با ۸/۷۵٪ (هفت خطا) و برای مجموعه ۴ برابر با ۱۰٪ (هشت خطا) و در مرحله صحت سنجی برای هر دو گروه برابر با ۲۵٪ (۵ خطا) است. در حالیکه گروه ۱ و ۲ در مرحله واسنجی، بدون خطا و در مرحله صحت سنجی به ترتیب دارای ۵٪ (یک خطا) و ۱۵٪ (سه خطا) با کمترین تعداد K هستند، می‌توان گفت که این الگوریتم برای دو مجموعه مذکور دارای بهترین کارایی بوده است. به طور کلی تنوع نتایج حاصل از این الگوریتم برای این ۵ مجموعه داده نشان می‌دهد که این الگوریتم به نحوه تقسیم‌بندی داده‌ها نسبتا حساس است به طوری‌که بهترین و بدترین جواب با یکدیگر اختلاف قابل توجهی دارند و به ترتیب متعلق به مجموعه‌های ۱ و ۴ هستند.

مقایسه نتایج و نتیجه‌گیری

بررسی و مقایسه نتایج حاصل از این تحقیق نشان می‌دهد که:

۱) برای طبقه‌بندی کیفیت آب به جای استفاده مستقیم از شاخص‌های کیفی، می‌توان با آموزش الگوریتم‌های طبقه‌بندی نظیر SVM و KNN براساس شاخص‌های کیفی مناسب، تعداد زیادی از نمونه‌های آب را با سرعت بالا و هزینه اندک طبقه‌بندی کیفی نمود.

۲) برای بررسی کارایی الگوریتم‌های باناظر اگر به جای یک نوع از تقسیم‌بندی داده‌ها از چند نوع تقسیم‌بندی (روش صحت سنجی متقاطع) استفاده شود، کارایی الگوریتم‌ها بهتر ارزیابی می‌شود.

۳) الگوریتم SVM با داشتن هیچ خطایی در مرحله واسنجی و تعداد خطای کم و یا بدون خطا (در مجموعه داده‌های ۱و۴) در مرحله صحت سنجی دارای کارایی بسیار خوب برای طبقه‌بندی داده‌های کیفیت آب است؛ ولی الگوریتم KNN برای طبقه‌بندی داده‌های کیفت آب دارای تعداد خطای زیادی در هر دو مرحله واسنجی و صحت سنجی است و می‌توان گفت که کارایی چندان زیادی برای این امر ندارد.

٤) هر دو الگوریتم دارای کمترین مقدار بهینه برای پارامترهای C و K به ازای مجموعه داده‌های ۲ و ۵ هستند که این امر نشان دهنده پراکندگی منظم‌تر و الگوپذیرتر داده‌ها در این دو مجموعه است؛ در حالیکه این پارامترها به ازای مجموعه ۴ دارای بیشترین مقدار بهینه هستند که نشان دهنده پراکندگی نامنظم داده‌ها در این مجموعه است و این امر سبب آموزش پذیری سخت الگوریتم‌ها می‌شود.

٥) الگوریتم SVM دارای بهترین نتایج به ازای مجموعه داده‌های ۱ و ۴ است در حالیکه الگوریتم KNN بهترین نتایج را به ازای مجموعه‌های ۱ و ۲ دارد. در نتیجه می‌توان گفت که هر دو الگوریتم به‌طور مشترک دارای بهترین نتیجه به ازای مجموعه ۱ هستند. بنابراین می توان نتیجه گرفت که در مجموعه ۱، داده‌ها به بهترین نحو توزیع شده‌اند به گونه‌ای که هر دو الگوریتم در مراحل واسنجی و صحت سنجی بهترین نتایج را به ازای این مجموعه داده داشته‌اند.

٦) باوجود آموزش‌پذیری سخت الگوریتم‌ها درشرایط نامنظمی پراکندگی داده‌ها (مجموعه داده‌های ۴)، الگوریتم SVM برای چنین شرایطی دارای بهترین کارایی می‌باشد و این امر یک مزیت مهم برای این الگوریتم است. در صورتی که الگوریتم KNN دارای بدترین کارایی در چنین شرایطی است.

۷) الگوریتم SVM دارای نتایج تقریبا مشابهی به ازای مجموعه داده‌های مختلف است که نشان می‌دهد این الگوریتم نسبت به نحوه پراکندگی داده‌ها حساس نیست و این امر نیز مزیت دیگر این الگوریتم است. در حالیکه الگوریتم KNN نسبت به نحوه تقسیم‌بندی داده‌ها حساس بوده و نتایج کاملا متفاوتی را ایجاد می‌نماید. بنابراین می‌توان گفت که الگوریتم SVM قابل اعتمادتر از KNN است.

مراجع

1. Balabin, R. M., Safieva, R. Z., and Lomakina, E. I. (2010). "Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques." *J. Analytica Chimica Acta.*, 671, 27-35.
2. Byvatov, E., Fechner, U., Sadowski, J., and Schneider, G. (2003). "Comparison of Support Vector Machine and Artificial Neural Network systems for drug/nondrug classification." *J. Chem. Inf. Comput. Sci.*, 43, 1882- 1889.
3. Canadian Council of Ministers of the Environment (CCME): 2001, "Canadian Water Quality Guide-lines for the Protection of Aquatic Life: CCME Water Quality Index 1.0.", Technical Report, *Canadian Council of Ministers of the environment winnipeg*, MB, Canada.
4. Chen, X., Li, Y. S., Liu, Z., Yin, K., Li, Z., Wai, O. W. H., and King, B. (2004). "Integration of multi-source data for water quality classification in the Pearl River estuary and its adjacent coastal waters of Hong Kong." *J. Continental Shelf Research.*, 24, 1827–1843.
5. Horton, R. K. (1965). "An index number system for rating water quality." *J. Water Poll. Cont. Fed.*, 37(3), 300-306.
6. Lee, M. S., and Park, S. S. (2006). "Comparative analysis of classification methods for protein interaction verification system." *Lecture Notes in Computer Science, Advances in Information Systems.*, 4243, 227-236.
7. Tamouk, J., and Allahakbari, F. (2012). "A comparison among accuracy of KNN, PNN ,KNCN, DANN and NFL." *International Journal of Computer Science Issues.*, 9(3), 319-322.
8. Tsuta, M., Masry, G. E., Sugiyama, T., Fujita, K., and Sugiyama, J. (2009). "Comparison between linear discrimination analysis and support vector machine for detecting pesticide on spinach leaf by hyper spectral imaging with excitation-emission matrix." *proceedings, European Symposium on Artificial Neural Networks-Advances in Computational Intelligence and Learning*. Bruges (Belgium), pp. 337-342.
9. Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.
10. Werther, W., Lohninger, H., Stancl, F., and Varmuza, K. (1994). "Classification of mass spectra: A comparison of yes/no classification methods for the recognition of simple structural properties." *J. Chemometrics and Intelligent Laboratory Systems*., 22(1), 63-76.
11. Yang Su, M. (2011)," Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers", *J. Expert Systems with Applications*., 38, 3492–3498.