پیش بینی کیفیت آب‌های سطحی با استفاده از روش درخت تصمیم

محمدتقی ستاری[۱]، مهدی عباسقلی نایب زاد[۲] و رسول میرعباسی نجف آبادی[۳]*

تاریخ دریافت ۱۳۹۲/۰۲/۲۵
تاریخ پذیرش: ۱۳۹۲/۱۱/۰۶

چکیده

با توجه به کمبود منابع آب‌های سطحی در ایران توجه به کیفیت آب و اتخاذ تمهیداتی در راستای جلوگیری از آلودگی منابع آب شیرین ضروری است. توسعه پایدار کشاورزی بدون در نظر گرفتن کیفیت شیمیایی آب‌های سطحی غیرممکن است. کنترل کیفیت آب از موضوعات مورد توجه در برنامه‌ریزی آبیاری اراضی کشاورزی می باشد. از آنجائیکه پایش و ارزیابی کیفیت آبهای سطحی پرهزینه و زمان بر می باشد، بنابراین، یافتن روشی ارزان، آسان و نسبتا دقیق که در آن با حداقل پارامترهای هیدروشیمیایی بتوان طبقه کیفیت آب را پیش‌بینی نمود، بسیار مفید می باشد. درخت تصمیم جزو روش‌های نوین داده کاوی بوده که با بهره‌گیری از یک ساختار درختی داده‌ها را طبقه‌بندی نموده و ضمن استخراج الگوها و قوانین موجود در بین داده‌ها به منظور پیش‌بینی به کار می رود. در این مقاله با استفاده از روش تصمیم‌گیری درختی کیفیت آب برخی از رودخانه‌های واقع در دامنه‌های جنوبی کوه سهند در محل ایستگاه‌های هیدرومتری چکان، قیرمیزی گؤل، شیشوان، تازه کند مراغه و مغانجیق مورد بررسی قرار گرفت و برای هریک از رودخانه طبقه کیفیت آب با استفاده از قوانین اگر-آنگاه توسعه داده شد. نتایج بدست آمده از مدل نشان داد که روش تصمیم‌گیری درختی قادر است با استفاده از کمترین تعداد پارامتر هیدروشیمیایی طبقه کیفیت آب را با دقت بسیار بالایی تعیین کند.

واژه‌های کلیدی: پارامترهای هیدروشیمیایی، داده کاوی، درخت تصمیم، کوه سهند، کیفیت آبهای سطحی.

[۱] استادیار گروه مهندسی آب دانشکده کشاورزی، دانشگاه تبریز، ۰۹۱۴۴۰۱۵۸۰۲
[۲] کارشناسی ارشد مهندسی عمران، دانشگاه آزاد مراغه، ۰۹۳۹۲۰۸۵۸۴۲
[۳] دکترای منابع آب، گروه مهندسی آب دانشکده کشاورزی، دانشگاه شهرکرد، ۰۹۱۳۳۳۳۳۲۷۵ نویسنده مسئول: mirabbasi_r@yahoo.com

مقدمه

ارزیابی کیفیت آب‌های سطحی در هر منطقه جهت توسعه اراضی کشاورزی، طراحی و بهره‌برداری از سیستم‌های آبیاری و انتخاب الگوی کشت مناسب، ضروری می‌باشد. برای ارزیابی کیفیت آب‌های سطحی در دوره‌های زمانی مختلف از آب رودخانه نمونه‌برداری گردیده و پس از انتقال نمونه‌ها به آزمایشگاه، غلظت یون‌های موجود در آب، هدایت الکتریکی، pH و ... اندازه‌گیری می‌شود و نهایتا بر اساس مقاصد مختلف کیفیت آب طبقه‌بندی می‌گردد. استفاده از انواع مختلف دیاگرام‌ها، به دلیل سهولت، یکی از ابزارهای متداول در طبقه‌بندی کیفیت آب می‌باشد. دیاگرام‌های زیادی توسط محققین برای طبقه‌بندی کیفیت آب برای مقاصد گوناگون توسعه داده شده‌اند. یکی از این دیاگرام ها که برای طبقه بندی کیفیت آب آبیاری ارائه شده، دیاگرام United States Salinity Laboratory با نام اختصاری دیاگرام USSL می باشد (آزمایشگاه شوری ایالات متحده، ۱۹۵۴). هزینه‌های بالای آزمایشات هیدروشیمی در اندازه‌گیری تعداد زیادی پارامتر و همینطور فاصله زمانی زیاد بین نمونه‌برداری و اخذ نتایج از آزمایشگاه جزو محدودیت‌های اساسی در تعیین کیفیت آب‌های سطحی می‌باشد. برای حل این مشکلات می‌توان از مدل‌هایی به منظور طبقه‌بندی و پیش‌بینی کیفیت آب سطحی بر اساس مشاهدات تاریخی و نتایج برگرفته از کلاس‌بندی دیاگرام USSL استفاده نمود. میرعباسی و همکاران (۲۰۰۸) با استفاده از منطق فازی اقدام به ارزیابی کیفیت آب آبیاری نمودند. یحیی و همکاران (۲۰۱۲) با استفاده از روش تحلیل رگرسیون، گسترش آلودگی و تغییرات فصلی کیفیت آب رودخانه ایندوس در پاکستان را در دوره قبل و پس از بارش های موسمی سالهای ۲۰۰۸ و ۲۰۰۹ مورد بررسی قراردادند. نتایج نشان دادکه با استفاده از روش رگرسیون می‌توان گسترش آلودگی و تغییرات فصلی را مدل نمود و این روش باعث صرفه جویی در زمان و هزینه‌های آزمایشگاهی می‌شود. رحمانی و همکاران (۱۳۸۷) کیفیت آب رودخانه‌های جاری در دشت همدان-بهار را برای مقاصد آبیاری بر مبنای دیاگرام ویلکاکس مورد مطالعه قراردادند. بدین منظور پس از شناسایی سرشاخه‌های رودخانه‌ها، یون‌های سدیم،کلسیم، منیزیم، اسیدیته (pH) و هدایت الکتریکی (Ec) اندازه‌گیری و نسبت جذب سدیم (SAR) محاسبه شد. سپس با استفاده از دیاگرام ویلکاکس کیفیت آب رودخانه ها ارزیابی گردید. نتایج نشان داد که کیفیت آب با حرکت به سمت پایین دست مسیر رودخانه کاهش می یابد. گلجان و همکاران (۱۳۸۸) جهت بررسی و طبقه بندی کیفیت آب رودخانه‌های شهرستان نور از دیاگرام پایپر استفاده نموده و اثرات زیست محیطی کیفیت آب رودخانه ها را مورد بررسی قرار دادند. علیایی و همکاران (۱۳۸۹) از مدل شبکه عصبی مصنوعی پرسپترون چند لایه برای مدل سازی شاخص های کیفی آب رودخانه مرادبیک در همدان استفاده نمودند. شاخص های کیفی مورد بررسی شامل اکسیژن مورد نیاز بیولوژیکی (BOD) و اکسیژن محلول (DO) بودند. نتایج حاصل، کارایی مدل شبکه عصبی مصنوعی را بعنوان تکنیکی برتر برای شبیه سازی تغییرات شاخص های کیفی آب نشان داد. حاجیان نژاد و رهسپار (۱۳۸۹) تاثیر رواناب ها و پساب تصفیه خانه فاضلاب را بر کیفیت آب رودخانه زاینده رود مورد بررسی قرار دادند. نتایج نشان داد که کیفیت آب رودخانه زاینده رود حتی درشرایطی که بارندگی صورت نمی گیرد از رواناب های شهر اصفهان متاثراست. سلاجقه و همکاران (۱۳۹۰) اثر تغییرات کاربری اراضی و آثار آن برکیفیت آب رودخانه‌های حوضه آبریزکرخه را مورد مطالعه قراردادند. نتایج نشان دادکه تغییرات کاربری اراضی در این حوضه باعث کاهش شدید کیفیت آب رودخانه‌ها شده است.

پیشرفت های جدید در شیوه های مدل سازی مبتنی بر روش های داده کاوی نظیر تصمیم گیری درختی یک جایگزین مناسب برای طبقه بندی کیفیت منابع آب می‌باشد. روش های داده کاوی با استفاده از تکنیک‌های متعدد نسبت به آموزش داده‌ها جهت استخراج روابط، الگوها، قواعد و نظم پنهان درون داده ها اقدام می کند. درخت تصمیم بعنوان یکی از روش‌های طبقه بندی و پیش بینی قادر است با استفاده از داده‌های مشاهداتی تاریخی، قواعد اگر-آنگاه را طوری تولید کند که بتوان طبقه کیفیت آب را پیش بینی نمود. استفاده از روش‌های داده کاوی باعث کاهش چشمگیر هزینه های نمونه برداری و آزمایشگاهی گردیده و این امکان را برای کارشناسان مهیا خواهد نمود تا در مدت زمان خیلی کوتاه و با استفاده از تعداد کمتری پارامتر هیدروشیمیایی، طبقه کیفیت آب

را تعیین نمایند. سانتوس و همکاران (۲۰۰۵) از یک چارچوب درختی برای انتخاب پارامترهای بیوشیمیائی مورد بررسی در سنجش از دور کیفیت آب سیلاب ها استفاده نمودند.

هدف از این مطالعه، طبقه بندی کیفیت آب‌های سطحی در ایستگاه‌های هیدرومتری چکان، قیرمیزی گؤل، شیشوان، تازه کند مراغه و مغانجیق واقع در دامنه های جنوبی کوه سهند با استفاده از روش تصمیم گیری درختی و توسعه قوانین اگر-آنگاه بر اساس دیاگرام USSL جهت انجام طبقه بندی کیفیت آب می باشد.

مواد و روش

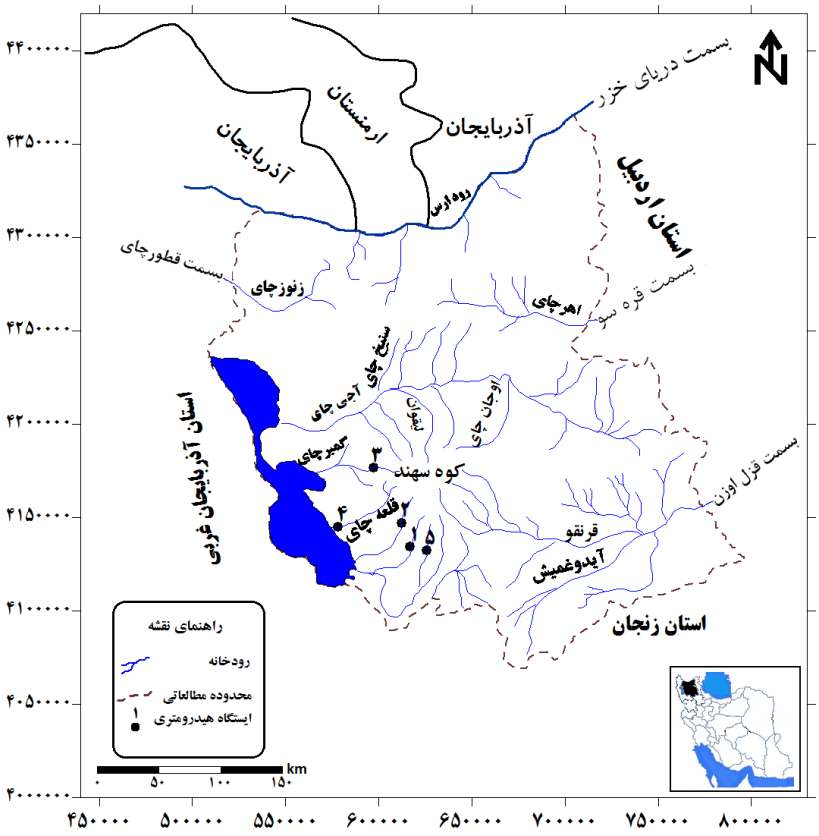
منطقه مورد مطالعه و داده های مورد استفاده

منطقه مورد مطالعه دامنه های جنوب کوه سهند می باشد. موقعیت جغرافیایی این منطقه مابین ۳۷ درجه و ۴۵ دقیقه تا ۳۸ درجه و ۴۵ دقیقه عرض شمالی و ۴۶ درجه تا ۴۷ درجه طول شرقی واقع شده است. در شکل ۱ موقعیت جغرافیایی منطقه مورد مطالعه به همراه رودخانه ها و محل ایستگاه های هیدرومتری نشان داده شده است. مشخصات ایستگاه های هیدرومتری منتخب در جدول ۱ ارائه شده است. جهت مدل سازی از پارامترهای هیدروشیمیائی و دبی ماهانه رودخانه ها در ایستگاه های مورد مطالعه، استفاده گردید. این داده ها از شرکت آب منطقه ای استان آذربایجان‌شرقی اخذ و پس از بررسی صحت و سقم آنها و بازسازی داده های گمشده مورد استفاده قرار گرفتند.

داده‌کاوی و روش تصمیم‌گیری درختی

داده‌کاوی به بررسی و تجزیه و تحلیل مجموعه بزرگی از داده‌ها به منظور کشف الگوها و قوانین پنهان و معنی‌دار درون داده‌ها اطلاق می‌شود. داده‌کاوی به دو نوع هدایت شده و غیرهدایت شده تقسیم بندی می‌شود. داده‌کاوی هدایت شده، دارای متغیر هدف خاص و از پیش تعیین شده است که به دنبال الگویی خاص می‌گردد، در حالیکه هدف داده‌کاوی غیر هدایت شده، یافتن الگوها یا تشابهات بین گروه‌هایی از اطلاعات، بدون داشتن متغیر هدف خاص و یا مجموعه‌ای از دسته‌ها و الگوهای از پیش تعیین شده می‌باشد. یک مدل داده کاوی اساساً به الگوریتم یا مجموعه‌ای از قوانینی گفته می‌شود که مجموعه‌ای از ورودی‌ها را با هدف یا مقصد خاصی مرتبط می‌نماید. داده کاوی شامل مراحل مختلفی می باشد که عبارتند از: ۱- تعیین اطلاعات گذشته، ۲- پیرایش داده ها و پردازش اولیه. در این مرحله خطاهای داده ها تصحیح و داده های اشتباه جایگزین می شوند، ۳- یکپارچه سازی داده ها، ۴- انتخاب مجموعه متغیرهای هدف، ۵- یافتن ویژگی‌های مورد استفاده و تعیین ویژگی‌های جدید، ۶- نمایش داده‌ها بصورتی که بتوان در داده کاوی استفاده نمود، ۷- انتخاب عملیات داده‌کاوی (طبقه بندی، خوشه بندی، پیش‌بینی و غیره)، ۸- انتخاب روش داده‌کاوی (شبکه های عصبی، درخت تصمیم و نظایر آن)، ۹- انجام داده کاوی و جستجو برای یافتن الگوی مناسب، ۱۰- ارزیابی و تحلیل الگوی به دست آمده و حذف الگوهای نامناسب، ۱۱- تفسیر نتایج داده ها و استنتاج اطلاعات با ارزش.

طبقه بندی شامل بررسی ویژگی‌های یک شیء جدید و تخصیص آن به یکی از مجموعه‌های از قبل تعیین شده می‌باشد. عمل طبقه بندی با تعریف درستی از دسته‌ها و مجموعه‌ای از ویژگی‌ها که حاوی موارد از پیش دسته‌بندی شده هستند، انجام می شود. این عمل شامل ساختن مدلی است که بتوان از آن برای طبقه بندی داده‌های دسته‌بندی نشده، استفاده نمود. درختان تصمیم گیری به منظور پیشگویی یا کلاسه بندی داده ها براساس مجموعه قوانین تصمیم ایجاد می شوند. کلاسه بندی داده ها با درختان تصمیم گیری یک فرایند دو مرحله‌ای می باشد. در مرحله اول که به آن مرحله آموزش گفته می شود، مدلی براساس یک الگوریتم کلاسه بندی منطبق با داده کاوی مربوط به مجموعه آموزشی ساخته می شود. مجموعه آموزشی به صورت تصادفی از پایگاه داده انتخاب می شود. در مرحله دوم یادگیری از طریق یک تابع $y=f(X)$ انجام می شود که می تواند برچسب کلاس هر رکورد x از پایگاه داده را پیش بینی کند. مرحله یادگیری خود طی دو گام اساسی رشد و هرس انجام می شود. در طول فرآیند آموزش الگوریتم درخت تصمیم می بایست به صورت مکرر موثرترین روش جهت تقسیم کردن مجموعه رکوردها به فرزندان را بیابد. مرحله هرس برای جلوگیری از پردازش بیش از حد و بزرگ شدن درخت تصمیم که باعث پیچیدگی و افزایش تعداد قوانین اگر-آنگاه می شود، صورت می گیرد.



شکل (۱): موقعیت جغرافیایی محدوده مطالعاتی و ایستگاه‌های هیدرومتری منتخب

جدول (۱): موقعیت مکانی ایستگاه‌های هیدرومتری منتخب

ردیف	نام رودخانه	نام ایستگاه	عرض جغرافیایی دقیقه-درجه	طول جغرافیایی دقیقه-درجه	ارتفاع (متر)
۱	چکان چای	چکان	۳۷-۵۱	۴۶-۴۹	۱۷۰۰
۲	صوفی چای	تازه کند	۳۷-۲۱	۴۶-۱۹	۱۵۵۰
۳	قنبر چای	قیرمیزی گؤل	۳۷-۴۴	۴۶-۰۶	۱۰۸۰
۴	قلعه چای	شیشوان	۳۷-۲۷	۴۵-۵۳	۱۲۹۰
۵	مغانجیق	مغانجیق	۳۷-۲۰	۴۶-۲۵	۱۶۵۰

در روش درخت تصمیم معیار مورد استفاده برای ایجاد شاخه ها و جداسازی، بی نظمی یا آنتروپی می باشد. اگر ویژگی هدف دارای C مقدار مختلف باشد، آنتروپی S نسبت به این دسته بندی C گانه بصورت زیر تعریف می شود (کوینلن، ۱۹۹۳):

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log p_i \qquad (۱)$$

که در آن، p_i نسبتی از S است که به دسته i تعلق دارد. توجه شود که لگاریتم در مبنای ۲ در نظر گرفته می‌شود. در این حالت حداکثر آنتروپی میتواند $\log_2 C$ باشد. در این مقاله جهت توسعه مدل درختی کیفیت آب سطحی و استخراج قواعد اگر-آنگاه از نرم افزار See5 که مبتنی بر الگوریتم C5.0 بوده و توسط کوینلن (۲۰۰۰) توسعه داده شده، استفاده گردیده است.

نتایج و بحث

در این تحقیق ابتدا با استفاده از دیاگرام USSL و بر اساس پارامترهای هیدروشیمیایی موثر در ارزیابی کیفیت آب، طبقه و کلاس کیفیت آب رودخانه ها مشخص شدند. براساس نمونه برداری‌های انجام یافته در این ۶ رودخانه

در مجموع برای هر یک از رودخانه‌ها دبی و ۱۲ پارامتر هیدروشیمیایی شامل یون کلسیم (Ca)، یون منیزیم (Mg)، یون کلر(Cl)، بی کربنات (HCo3)، درصد سدیم (Na%)، اسیدیته (pH)، سولفات (SO_4)، مجموع آنیون‌ها (Sum A)، مجموع کاتیون‌ها (Sum C)، کل نمک محلول (TDS)، نسبت سدیم جذبی (SAR) و هدایت الکتریکی(Ec) مورد توجه قرار گرفت. ۱۳ صفت و ویژگی هیدروشیمیایی فوق الذکر بعنوان ورودی مدل درخت تصمیم و کلاس کیفیت آب بعنوان خروجی و ویژگی هدف در نظر گرفته شد. کلاس کیفیت آب، که براساس دیاگرام USSL تعیین می شود و در اینجا بعنوان ویژگی هدف مورد استفاده قرار گرفت، دارای ماهیتی گسسته می باشد. در زیر نتایج بدست آمده از تحلیل مدل جهت تعیین کلاس کیفیت آب برای ۶ رودخانه مورد مطالعه ارائه گردیده است. برای هر ایستگاه از بین نمونه های مورد بررسی ۷۵ درصد داده ها برای آموزش و ۲۵ درصد برای آزمون (تست) در نظر گرفته شده است.

رودخانه چکان چای

در این رودخانه جمعا تعداد ۲۳۱ مورد نمونه برداری از آب در فواصل زمانی مختلف در محل ایستگاه چکان انجام گرفته و با توجه به نتایج آنالیز هیدورشیمیایی، کلاس کیفیت آب با استفاده از نمودار USSL تعیین شده است. از بین ۲۳۱ مورد به تفکیک تعداد ۱۷۲ نمونه (۷۵ درصد از کل داده ها) برای آموزش مدل و مابقی یعنی ۵۹ نمونه (۲۵ درصد از کل داده ها) برای آزمون در نظر گرفته شد. خلاصه نتایج مدل تصمیم درختی C5.0 برای ایستگاه چکان در جدول ۲ ارائه گردیده است.

نتایج ارزیابی نمونه‌های آموزشی در ایستگاه چکان نشان می‌دهد که در این مدل تصمیم درختی ۳ قانون جهت طبقه‌بندی استفاده شده است. قانون ۱ بیان می‌کند که "اگرEC≤245 باشد، آنگاه کلاس کیفیت آب C1-S1 خواهد بود". از کل نمونه های آموزشی، ۱۴ نمونه با این قانون همخوانی دارد. دقت آماره لاپلاس برای قانون ۱ برابر با ۹۳/۸٪ و مقدار عددی آماره لیفت برابر با ۱۱/۵ می‌باشد. کم بودن نسبی مقدار آماره لیفت نشانگر دقت بالای قانون ۱ می باشد. قانون ۲ بیان می کند که "اگر 245〈EC و EC≤750 باشد، آنگاه کلاس کیفیت آب C2-S1 خواهد بود". از کل نمونه‌های آموزشی، ۱۵۵ نمونه با این قانون همخوانی دارد. دقت آماره لاپلاس برای قانون ۲ برابر با ۹۹/۴٪ و مقدار عددی آماره لیفت برابر با ۱/۱ می باشد. کم بودن مقدار آماره لیفت نشانگر دقت بالای قانون ۲ می‌باشد. قانون ۳ نیز بیان می کند که "اگر EC〉750 باشد، آنگاه کلاس کیفیت آب C3-S1 خواهد بود". از کل نمونه های آموزشی، ۳ نمونه با شرایط این قانون همخوانی دارد. دقت آماره لاپلاس برای قانون ۳ برابر با ۸۰٪ و مقدار عددی آماره لیفت برابر با ۴۵/۹ می باشد. بالا بودن نسبی مقدار آماره لیفت نشانگر دقت نسبتا کم قانون ۳ می باشد. همانگونه که از جدول ۲ استنباط می شود در مجموعه این قوانین مستخرج از نمونه های آموزشی تمامی نمونه ها از قوانین تبعیت کرده، این در حالیست که ۱۷۲ مورد درست طبقه بندی شده است. از اینروی در مجموع خطا برابر با صفر درصد خواهد بود. در این بخش از بین ۱۸۰ نمونه آموزشی ۱۴ نمونه در کلاس C1-S1، ۱۵۵ نمونه در کلاس C2-S1، ۳ نمونه در کلاس C3-S1 توزیع شده است. نمودار پراکنش نمونه های آموزشی در ایستگاه چکان در شکل ۲ ارائه گردیده است. همچنین ماتریس اغتشاش که بیانگر نحوه پراکنش نمونه ها بین کلاس‌های مختلف می باشد، برای ایستگاه چکان در جدول ۳ ارائه گردیده است. در ماتریس اغتشاش نمونه هایی که روی قطر اصلی ماتریس واقع شده اند، نمونه‌هایی هستند که درست و صحیح پیش بینی شده اند. در مورد رودخانه چکان تمامی نمونه های آموزشی در روی قطر اصلی واقع شده است و باعث شده که خطای این بخش صفر درصد باشد. این امر بیانگر دقت بسیار بالای مدل در بخش آموزش می‌باشد.

پس از آموزش مدل و استخراج قوانین مربوطه، برای بررسی صحت نتایج، مدل با ۵۹ مورد نمونه مورد آزمون قرار گرفت. ماتریس اغتشاش مربوط به نمونه های آزمایشی در جدول ۴ ارائه گردیده است. از جدول ۴ ملاحظه می‌شود که از کل مجموع ۵۹ مورد نمونه آزمایشی هیچ کدام از نمونه ها خارج از قطر اصلی ماتریس اغتشاش نمی باشد. بنابراین خطای ناشی از داده های بخش آموزش مدل برابر با صفر درصد خواهد بود. نمودار پراکنش نمونه های آزمون در ایستگاه چکان در شکل ۳ ارائه گردیده است. با توجه به اینکه در این رودخانه بیشترین پراکنش کلاس کیفیت آب مربوط به کلاس C2-S1 می باشد، بنابراین این کلاس بعنوان کلاس پیش‌فرض

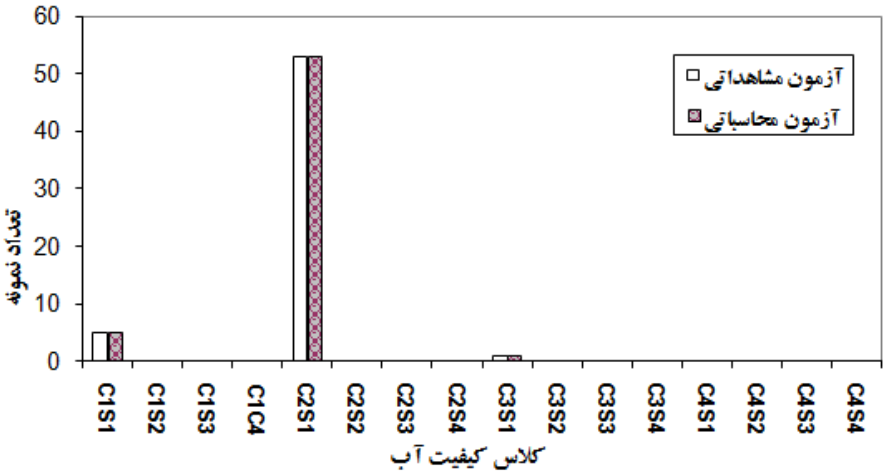
در نظر گرفته شده است. همچنین نتایج بدست آمده نشان می دهد مدل تصمیم درختی C5.0 برای پیش بینی کیفیت آب تنها از پارامتر Ec استفاده کرده است. به عبارت دیگر، از بقیه پارامترهای هیدروشیمایی جهت توسعه مدل بهره نجسته است. لذا برای تخمین کلاس کیفیت آب در ایستگاه چکان با اندازه گیری فقط پارامتر Ec کافی می باشد.

جدول (۲): خلاصه نتایج مدل تصمیم درختی برای ایستگاه هیدرومتری چکان

شماره قانون	اگر ۱	اگر ۲	آنگاه (طبقه کیفیت آب)	تعدادنمونه صادق غیرصادق/	دقت لاپلاس(درصد) /LIFT
۱	EC ≤ 245	-	C1-S1	14/0	93.8/11.5
۲	EC 〉 245	EC ≤ 750	C2-S1	155/0	99.4/1.1
۳	EC 〉 750	-	C3-S1	3/0	80.0/45.9

جدول (۳): ماتریس اغتشاش برای نمونه های آموزش ایستگاه هیدرومتری چکان

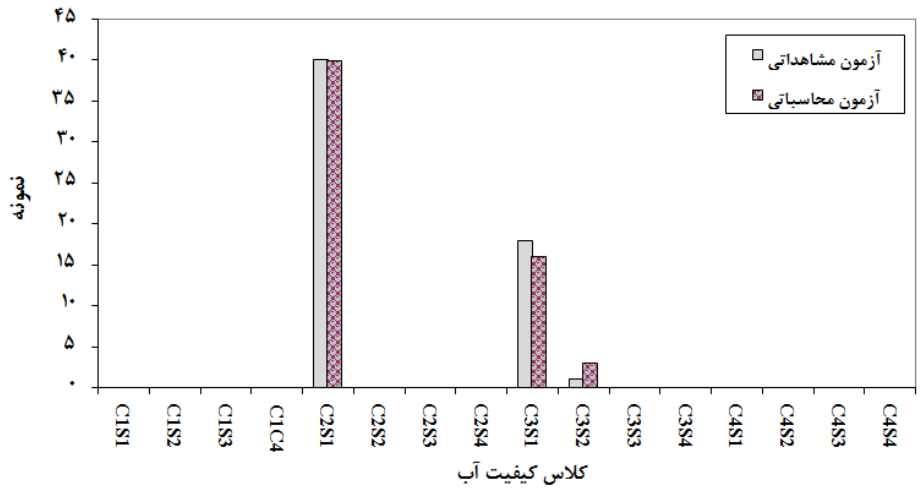
کلاس	C1S1	C2S1	C3S1
C1S1	۱۴		
C2S1		۱۵۵	
C3S1			۳



شکل (۲): نمودار پراکنش نمونه های آموزشی در ایستگاه چکان

جدول (۴): ماتریس اغتشاش برای نمونه های تست ایستگاه هیدرومتری چکان

کلاس	C1S1	C2S1	C3S1
C1S1	۵		
C2S1		۵۳	
C3S1			۱



شکل (۳): نمودار پراکنش نمونه های آزمون در ایستگاه هیدرومتری چکان

رودخانه صوفی چای (تازه کند)

در این رودخانه جمعا به تعداد ۲۴۱ مورد نمونه برداری از آب در فواصل زمانی مختلف در محل ایستگاه هیدرومتری تازه کند انجام گرفته است. خلاصه نتایج مدل تصمیم درختی C5.0 برای رودخانه تازه کند در جدول ۵ ارائه گردیده است. نتایج ارزیابی نمونه های آموزشی در ایستگاه تازه کند نشان می دهد که در این مدل تصمیم درختی ۳ قانون جهت طبقه بندی استفاده شده است. همچنانکه از جدول ۵ ملاحظه می شود از مجموع ۱۸۰ مورد، ۱۷۹ مورد درست طبقه بندی شده است از اینروی در مجموع خطا برابر با ۰/۶ درصد خواهد بود. در این بخش از بین ۱۸۰ نمونه آموزشی ۴۴ نمونه در کلاس C2-S1، ۱۳۲ نمونه در کلاس C3-S1، ۴ نمونه در کلاس C3-S2 توزیع شده است. ماتریس اغتشاش نمونه‌های آموزشی ایستگاه تازه کند در جدول ۶ ارائه گردیده است. در مورد رودخانه صوفی چای تازه کند فقط ۱ نمونه آموزشی در خارج از قطر اصلی واقع شده است و بجای اینکه نمونه در کلاس C1-S1 قرار بگیرد، به اشتباه در کلاس C2-S1 قرار گرفته است.

پس از آموزش مدل و استخراج قوانین مربوطه برای بررسی صحت نتایج، مدل با ۶۱ مورد نمونه مورد آزمون قرار گرفت. ماتریس اغتشاش مربوط به نمونه های آزمایشی در جدول ۷ ارائه گردیده است. از جدول ۷ ملاحظه می شود که از کل مجموع ۶۱ مورد نمونه آزمایشی ۲ مورد از نمونه ها خارج از قطر اصلی ماتریس اغتشاش می باشد. این ۲ مورد بجای اینکه در کلاس C3-S1 قرار بگیرند در کلاس C3-S2 واقع شده اند. بنابراین خطای ناشی از داده‌های بخش آزمایشی مدل برابر با ۳/۳ درصد خواهد بود. در این رودخانه کلاس C3-S1 بعنوان کلاس پیش فرض در نظر گرفته شده است. همچنین نتایج بدست آمده نشان می دهد مدل تصمیم درختی C5.0 برای پیش‌بینی کیفیت آب تنها از پارامترهای SAR و Ec استفاده کرده است.

جدول (۵): خلاصه نتایج مدل تصمیم درختی برای ایستگاه هیدرومتری تازه کند

شماره قانون	اگر		آنگاه (طبقه کیفیت آب)	تعدادنمونه صادق/ غیرصادق	دقت لاپلاس(درصد) /LIFT
	۱	۲			
۱	EC≤750		C2-S1	44/1	95.7/4.0
۲	EC〉750	SAR≤4.651	C3-S1	132/0	99.3/1.4

جدول (۶): ماتریس اغتشاش برای نمونه های آموزش ایستگاه هیدرومتری تازه کند

کلاس	C1S1	C2S1	C3S1	C3S2
C1S1	۰	۱		
C2S1		۴۴		
C3S1			۱۳۲	
C3S2				۴

جدول (۷): ماتریس اغتشاش برای نمونه های تست ایستگاه هیدرومتری تازه کند

کلاس	C2S1	C3S1	C3S2
C2S1	۱۴		
C3S1		۴۵	۲
C3S2			۰

رودخانه قنبرچای

در این رودخانه جمعا به تعداد ۲۱۸ مورد نمونه برداری از آب در فواصل زمانی مختلف در محل ایستگاه قیرمیزی گؤل انجام گرفته است. خلاصه نتایج مدل تصمیم درختی C5.0 برای رودخانه قنبر چای در جدول ۸ ارائه گردیده است. نتایج ارزیابی نمونه‌های آموزشی در رودخانه قنبر چای نشان می دهد که در این مدل تصمیم درختی ۳ قانون جهت طبقه بندی استفاده شده است. همچنانکه از جدول ۸ ملاحظه می‌شود در مجموعه این قوانین مستخرج از نمونه های آموزشی تمامی نمونه ها از قوانین تبعیت کرده و این در حالیست که ۱۶۴ مورد درست طبقه بندی شده است. در این بخش از بین ۱۶۴ نمونه آموزشی ۲۷ نمونه در کلاس C1-S1، ۱۲۱ نمونه در کلاس C2-S1 و ۳ نمونه در کلاس C3-S1 توزیع شده است. ماتریس اغتشاش نمونه های آموزشی رودخانه قنبرچای در جدول ۹ ارائه گردیده است. در مورد ایستگاه قیرمیزی گؤل تمامی نمونه ها روی قطر اصلی واقع شده است و باعث شده که خطای این بخش صفر درصد باشد. پس از آموزش مدل و استخراج قوانین مربوطه برای بررسی صحت نتایج مدل با ۵۴ مورد نمونه مورد آزمون قرار گرفت. ماتریس اغتشاش مربوط به نمونه های آزمایشی در جدول ۱۰ ارائه گردیده است. از جدول ۱۰ ملاحظه می شود که از کل مجموع ۵۴ مورد نمونه آزمایشی فقط ۱ مورد نادرست طبقه بندی شده و خارج از قطر اصلی ماتریس اغتشاش می باشد. این ۱ مورد بجای اینکه در کلاس C2-S1 پیش بینی شود، به اشتباه در کلاس C3-S1 واقع شده است. بنابراین خطای ناشی از داده های بخش آموزش مدل برابر با ۱/۹ درصد خواهد بود. با توجه به اینکه در این رودخانه بیشترین پراکنش کلاس کیفیت مربوط به کلاس C2-S1 می باشد، بنابراین این کلاس بعنوان کلاس پیش فرض در نظر گرفته شده است. همچنین نتایج بدست آمده نشان می دهد مدل تصمیم درختی C5.0 برای پیش بینی کیفیت آب تنها از پارامترهای Ec و مجموع کاتیونها (SUM C) استفاده کرده است. به عبارتی دیگر برای تخمین کلاس کیفیت آب در رودخانه قنبرچای فقط اندازه گیری این دو پارامترکافی می باشد.

جدول (۸): خلاصه نتایج مدل تصمیم درختی C5.0 برای ایستگاه هیدرومتری قیرمیزی گؤل

شماره قانون	اگر ۱	اگر ۲	آنگاه (طبقه کیفیت آب)	تعدادنمونه صادق غیرصادق/	دقت لاپلاس(درصد) /LIFT
۱	EC≤243	-	C1-S1	27/0	96.6/5.9
۲	EC〉243	SUM C≤7.3	C2-S1	121/0	99.2/1.3
۳	SUM C〉7.3	-	C3-S1	16/0	94.4/9.7

جدول (۹): ماتریس اغتشاش برای نمونه های آموزش ایستگاه هیدرومتری قیرمیزی گؤل

کلاس	C1S1	C2S1	C3S1
C1S1	۲۷		
C2S1		۱۲۱	
C3S1			۳

جدول (۱۰): ماتریس اغتشاش برای نمونه های تست ایستگاه هیدرومتری قیرمیزی گؤل

کلاس	C1S1	C2S1	C3S1
C1S1	۹		
C2S1		۴۱	۱
C3S1			۳

رودخانه قلعه چای

در این رودخانه جمعا به تعداد ۱۰۳ مورد نمونه برداری از آب در فواصل زمانی مختلف در محل ایستگاه شیشوان انجام گرفته است. خلاصه نتایج مدل تصمیم درختی C5.0 برای رودخانه قلعه چای در جدول ۱۱ ارائه گردیده است.

نتایج ارزیابی نمونه های آموزشی در رودخانه قلعه چای نشان می دهد که در این مدل تصمیم درختی ۳ قانون جهت طبقه بندی استفاده شده است. همچنانکه از جدول ۱۱ ملاحظه می شود در مجموعه این قوانین مستخرج از نمونه های آموزشی تمامی نمونه ها از قوانین تبعیت کرده و این در حالیست که ۷۷ مورد درست طبقه بندی شده است. در این بخش از بین ۷۷ نمونه آموزشی ۳ نمونه در کلاس C1-S1، ۷۱ نمونه در کلاس C2-S1 و ۳ نمونه در کلاس C3-S1 توزیع شده است. ماتریس اغتشاش نمونه های آموزشی رودخانه قنبرچای در جدول ۱۲ ارائه گردیده است. در مورد ایستگاه شیشوان تمامی نمونه ها روی قطر اصلی واقع شده است و باعث شده که خطای این بخش صفر درصد باشد.

جدول (۱۱): خلاصه نتایج مدل تصمیم درختی C5.0 برای ایستگاه هیدرومتری شیشوان

شماره قانون	اگر		آنگاه (طبقه کیفیت آب)	تعدادنمونه صادق غیرصادق/	دقت لاپلاس(درصد) /LIFT
	۱	۲			
۱	SUM C≤2.35	-	C1-S1	3/0	80.0/20.5
۲	TDS≤458	SUM C〉2.35	C2-S1	71/0	98.6/1.1
۳	TDS〉458	-	C3-S1	3/0	80.0/20.5

جدول (۱۲): ماتریس اغتشاش برای نمونه های آموزش ایستگاه هیدرومتری شیشوان

کلاس	C1S1	C2S1	C3S1
C1S1	۳		
C2S1		۷۱	
C3S1			۳

پس از آموزش مدل و استخراج قوانین مربوطه برای بررسی صحت نتایج مدل با ۲۶ مورد نمونه مورد آزمون قرار گرفت. ماتریس اغتشاش مربوط به نمونه های آزمایشی در جدول ۱۳ ارائه گردیده است. از جدول ۱۳ ملاحظه می شود که از کل مجموع ۲۶ مورد نمونه آزمایشی فقط ۲ مورد نادرست طبقه بندی شده و خارج از

قطر اصلی ماتریس اغتشاش می باشد. این ۲ مورد بجای اینکه به ترتیب در کلاس C1-S1 و C2-S1 پیش بینی شوند، به اشتباه در کلاس‌های C2-S1 و C3-S1 واقع شده‌اند. بنابراین خطای ناشی از داده های بخش آموزش مدل برابر با ۷/۷ درصد خواهد بود. با توجه به اینکه در این رودخانه بیشترین پراکنش کلاس کیفیت مربوط به کلاس C2-S1 می باشد، بنابراین این کلاس بعنوان کلاس پیش فرض در نظر گرفته شده است. همچنین نتایج بدست آمده نشان می دهد مدل تصمیم درختی C5.0 برای پیش بینی کیفیت آب تنها از پارامترهای SUM C و TDS استفاده کرده است.

جدول (۱۳): ماتریس اغتشاش برای نمونه های تست ایستگاه هیدرومتری شیشوان

کلاس	C1S1	C2S1	C3S1
C1S1	۲	۱	
C2S1		۱۹	۱
C3S1			۳

رودخانه مغانجیق چای

در این رودخانه جمعا به تعداد ۲۲۸ مورد نمونه برداری از آب در فواصل زمانی مختلف در محل ایستگاه مغانجیق انجام گرفته است. خلاصه نتایج مدل تصمیم درختی C5.0 برای رودخانه مغانجیق در جدول ۱۴ ارائه گردیده است. در رودخانه مغانجیق چای ۳ قانون جهت طبقه بندی استفاده شده است. همچنانکه از جدول ۱۴ ملاحظه می شود در مجموعه این قوانین مستخرج از نمونه های آموزشی تمامی نمونه ها از قوانین تبعیت کرده است. در این بخش از بین ۱۱۷ نمونه آموزشی ۱۲ نمونه در کلاس C1-S1، ۱۵۵ نمونه در کلاس C2-S1 و ۴ نمونه در کلاس C3-S1 توزیع شده است. ماتریس اغتشاش نمونه های آموزشی رودخانه مغانجیق چای در جدول ۱۵ ارائه گردیده است. در مورد ایستگاه مغانجیق تمامی نمونه ها روی قطر اصلی واقع شده است و باعث شده که خطای این بخش صفر درصد باشد.

جدول (۱۴): خلاصه نتایج مدل تصمیم درختی C5.0 برای ایستگاه هیدرومتری مغانجیق

شماره قانون	اگر ۱	اگر ۲	آنگاه (طبقه کیفیت آب)	تعدادنمونه صادق/ غیرصادق	دقت لاپلاس(درصد) /LIFT
۱	TDS≤156.65	-	C1-S1	12/0	92.9/13.2
۲	TDS⟩156.65	TDS≤510	C2-S1	155/0	99.4/1.1
۳	TDS⟩510	-	C3-S1	4/0	83.3/35.6

جدول (۱۵): ماتریس اغتشاش برای نمونه های آموزش ایستگاه هیدرومتری مغانجیق

کلاس	C1S1	C2S1	C3S1
C1S1	۱۲		
C2S1		۱۵۵	
C3S1			۴

ماتریس اغتشاش مربوط به نمونه های آزمایشی در جدول ۱۶ ارائه گردیده است. از جدول ۱۶ ملاحظه می شود که از کل مجموع ۵۷ مورد نمونه آزمایشی فقط ۳ مورد نادرست طبقه‌بندی شده و خارج از قطر اصلی ماتریس اغتشاش می‌باشد. این ۳ مورد بجای اینکه در کلاس C3-S1 پیش‌بینی شوند به اشتباه در کلاس C2-S1 واقع شده‌اند. بنابراین خطای ناشی از داده‌های بخش آموزش مدل برابر با ۵/۳ درصد خواهد بود. در این

رودخانه کلاس C2-S1 بعنوان کلاس پیش فرض در نظر گرفته شده است. همچنین نتایج بدست آمده نشان می‌دهد مدل تصمیم درختی C5.0 برای پیش‌بینی کیفیت آب تنها از پارامتر TDS استفاده کرده است. خلاصه نتایج بدست آمده از روش تصمیم درختی برای رودخانه های منتخب دامنه جنوبی سهند در جدول ۱۷ زیر آورده شده است. همانگونه که در این جدول ملاحظه می شود، روش درختی توانسته کیفیت آب را در تمام ایستگاه های مورد بررسی با دقت بسیار بالایی پیش‌بینی کند.

جدول (۱۶): ماتریس اغتشاش برای نمونه های تست ایستگاه هیدرومتری مغانجیق

کلاس	C1S1	C2S1	C3S1
C1S1	۴		
C2S1		۴۸	
C3S1		۳	۲

جدول (۱۷): خلاصه نتایج مدل تصمیم درختی C5.0 برای رودخانه های دامنه جنوبی کوه سهند

ایستگاه	تعداد قانون	آموزش		آزمون		کلاس پیش فرض
		خطا (درصد)	دقت مدل (درصد)	خطا (درصد)	دقت مدل (درصد)	
چکان	۳	۰/۰	۱۰۰	۰/۰	۱۰۰	C2-S1
تازه کند	۳	۰/۶	۹۹/۴	۳/۳	۹۶/۷	C3-S1
قیرمیزی گؤل	۳	۰/۰	۱۰۰	۱/۹	۹۸/۱	C2-S1
شیشوان	۳	۰/۰	۱۰۰	۷/۷	۹۲/۳	C2-S1
مغانجیق	۳	۰/۰	۱۰۰	۵/۳	۹۴/۷	C2-S1

نتیجه‌گیری

در این مقاله براساس نمونه برداری های انجام یافته در ۶ رودخانه دامنه جنوبی کوه سهند در مجموع برای هر یک از رودخانه ها دبی و ۱۲ پارامتر هیدروشیمیایی مورد توجه قرار گرفت. ۱۲ صفت و ویژگی هیدروشیمیایی بعنوان ورودی مدل درخت تصمیم و کلاس کیفیت آب بعنوان خروجی و ویژگی هدف در نظر گرفته شد. نتایج بدست آمده نشان داد تصمیم گیری درختی عمدتا با استفاده از ۴ پارامتر Ec، SUM C، SAR و TDS قادر به طبقه‌بندی کلاس کیفیت آب با دقت بسیار بالایی می‌باشد. طبیعتا تعیین پارامترهای موثر در تعیین کلاس کیفیت آب های سطحی با استفاده از روش تصمیم‌گیری درختی، باعث خواهد شد که علاوه بر کاهش چشمگیر هزینه های نمونه برداری و آنالیز شیمیائی نمونه‌ها، زمان اختصاص یافته برای تعیین کلاس کیفیت آب نیز کاهش یابد و در عین حال نتایج از دقت قابل قبولی برخوردار باشند.

منابع

۱. حاجیان نژاد م. و ا.ر. رهسپار. ۱۳۸۹. بررسی تاثیر روان آب ها و پساب تصفیه‌خانه فاضلاب بر پارامترهای کیفی آب رودخانه زاینده رود. مجله تحقیقات نظام سلامت /سال ششم/ویژه نامه، ص ۸۲۱-۸۲۸.

۲. رحمانی ع. ر.، م.ت. صمدی و م. حیدری. ۱۳۸۷. ارزیابی کیفیت آب رودخانه های جاری در دشت همدان-بهار برای آبیاری بر مبنای دیاگرام ویلکوکس. فن آوری زیستی در کشاورزی. سال هشتم، شماره ۱، ص ۲۷-۳۶.

۳. سلاجقه ع.، س. رضوان زاده، ن. ا. خراسانی، م. حمیدی فر و س. سلاجقه. ۱۳۹۰. تغییرات کاربری اراضی و آثار آن بر کیفیت آب رودخانه (مطالعه موردی: حوضه آبخیزکرخه). محیط شناسی، سال سی و هفت، شماره ۵۸، ص ۸۱-۸۶.

۴. علیایی ا.، ح. بانژاد، م.ت. صمدی، ع.ر. رحمانی. و م.ح. ساقی. ۱۳۸۹. ارزیابی کارایی شبکه عصبی مصنوعی در پیش‌بینی شاخص‌های کیفی (BOD و DO) آب رودخانه دره مرادبیک همدان. مجله دانش آب و خاک، جلد ۲۰/۱، شماره ۳، ص ۱۹۹-۲۱۰.

۵. گلجان ف.، ع. ر. کرباسی، ن. حاجی زاده ذاکر و غ.ر. نبی بیدهندی. ۱۳۸۸. تعیین کلاسه کیفی آب رودخانه‌های شهرستان نور. فصل‌نامه تحقیقات علوم آب، سال اول، شماره اول، ص ۳۵-۴۸.

6. Mirabbasi, R., Mazloumzadeh, S.M., & Rahnama, M.B., (2008). Evaluation of irrigation water quality using fuzzy logic, Research Journal of Environmental Sciences, 2(5): 340-352.

7. Quinlan, J.R. (1993). C4.5 Programs for machine learning, Morgan, Kaufmann, San Mateo, California.

8. Quinlan, J.R. (2000). Data mining tools See5 and C5.0 [cited Feb 2012]. Available from http://www.rulequest.com/see5-info.html.

9. Santos, M.F., Cortez, P., Quintela, H., Neves, J., Vicente, H. & Arteiro, J. (2005). Ecological Mining - A Case Study on Dam Water Quality6. In A. Zanasi, C. Brebbia and N. Ebecken (Eds.), Data Mining VI - Data Mining, Text Mining and their Business Applications, WIT Transactions of Information and Communication Technologies Vol. 35, pp. 523-531, WIT Press, 2005, ISBN:1-84564-017-9, ISSN:1746-4463.

10. U.S. Salinity Laboratory Staff, (1954). Diagnosis and improvement of saline and alkali soils: U.S. Dept. Agric. Handbook No.60, 160 p.

11. Wilcox, L.V. (1955). Classification and use of irrigation waters: U.S. Dept. Agric. Circ. 969, 19p.

12. Yahya, S.M., Rahman, A.U., Abbasi, H.N. (2012). Assessment of seasonal and polluting effects on the quality of river water by using regression analysis: A Case Study of River Indus in Province of Sindh, Pakistan. International Journal of Environmental Protection. 2(2): 10-16.

Surface water quality prediction using decision tree method

Mohammad Taghi Sattari[1], Mehdi Abbasgoli Naebzad[2] and Rasoul Mirabbasi Najafabadi[3]*

Abstract

Consideration of water quality and implementation of appropriate actions for preventing of water resources pollution is a very important issue in Iran because of surface water deficit. Sustainable development of agriculture is impossible without considering of surface water quality. Water quality control is a noteworthy issue in irrigation scheduling program of agricultural land. Since surface water quality monitoring and assessment is very expensive and time consuming. Thus, finding a cheap, simple and relatively exact method which can predict the water quality class base on minimum hydro chemical parameters would be very useful. Decision tree as one of the data mining techniques classify data sets based on a tree structure and uses for prediction base on extracting the exiting patterns and roles among data sets. In this study, the decision tree method was used to classify water quality in some hydrometrics stations located at southern side of Sahand Mountain, including Chekan, Girmizigol, Shishovan, Tazekand and Moghanjig. The water quality classes were defined based on if-then rules. The results showed that the decision tree method is able to predict the water quality classes based on small number of hydro chemical parameters with high accuracy.

Keywords: surface water quality, data mining, decision tree, hydro chemical parameters, Sahand Mountain.

[1]. Assistant Professor, Department of Water Engineering, University of Tabriz, Tabriz, Iran.
[2] MSc student of Civil Engineering, Islamic Azad University of Maragheh, Maragheh, Iran.
[3] Assistant Professor, Department of Water Engineering, ShahrekordUniversity, Shahrekord. Iran.