بررسی مدل‌های آماری در مدل‌سازی فرآیند تصفیه فاضلاب با استفاده از روش داده کاوی

علیرضا رایگان شیرازی نژاد[1]، مرتضی زارع[2]، فهیمه زارع[3]، محمد مهدی بانشی[*1]، سهیلا رضایی[2]

۱. مرکز تحقیقات عوامل اجتماعی موثر بر سلامت، دانشگاه علوم پزشکی یاسوج، یاسوج، ایران

۲. گروه مهندسی بهداشت محیط، دانشکده بهداشت دانشگاه علوم پزشکی یاسوج، یاسوج، ایران

۳. گروه مهندسی فناوری اطلاعات، دانشگاه شهید بهشتی، تهران، ایران

تاریخ دریافت مقاله: ۱۳۹۳/۱۰/۲۵ ؛ تاریخ پذیرش: ۱۳۹٤/۲/۱

چکیده

زمینه و هدف: تصفیه فاضلاب شامل فرآیندهای فیزیکی، شیمیایی و بیولوژیکی بسیار پیچیده و وابسته به هم می باشد، با استفاده از روش های داده کاوی می توان با دقت زیاد و بوسیله مدل های بدون محاسبات پیچیده ریاضی فرآیندهای تصفیه فاضلاب را مدلسازی کرد.

مواد و روش‌ها: در این پژوهش از داده های مربوط به فرآیندهای تصفیه فاضلاب موجود در شرکت آب و فاضلاب استان کهگیلویه و بویر احمد استفاده گردید. در مجموع ۳۳۰٦ داده مربوط به COD، TSS، PH و کدورت جمع آوری گردید. در نهایت داده های گردآوری شده با استفاده از نرم افزار های تحلیل آماری SPSS-16 (آمارتوصیفی) و داده کاوی Ibm Spss Modeler 14.2 و از طریق ۹ الگوریتم مورد تجزیه و تحلیل قرار گرفتند.

نتایج: با توجه به نتایج بدست آمده بر روی الگوریتم های رگرسیون لجستیکی، شبکه عصبی، شبکه بیزی، تحلیل تفکیکی، درخت تصمیم C5، درخت C&R، CHAID، QUESTو SVM به ترتیب دارای درصد دقت درستی ۹۰/۱٦، ۹٤/۱۷ ، ۸۱/۳۷ ، ۷۰/٤۸ ، ۹۷/۸۹ ، ۹٦/۵٦ ، ۹٦/٤٦ ، ۹٦/۸٤ ، ۸۸/۹۲ بودند.

بحث ونتیجه‌گیری: در این مطالعه، الگوریتم C5 به عنوان بهترین و کارآمدترین الگوریتم در جهت مدلسازی فرآیندهای تصفیه فاضلاب با دقت ۹۷/۸۹۹ درصد انتخاب گردید و موثرترین متغیر ها در این مدل به ترتیب PH، COD، TSS و کدورت بودند.

کلمات کلیدی: مدلسازی فرآیند تصفیه فاضلاب، داده کاوی، دسته بندی، کهگیلویه و بویر احمد

مقدمه

تصفیه فاضلاب شامل فرآیندهای فیزیکی، شیمیایی و بیولوژیکی بسیار پیچیده و وابسته به هم می باشد. راهبری تصفیه خانه های فاضلاب به صورت تجربی یا اطلاعات برگرفته از مطالعات پایلوت می باشد. به همین دلیل، اجرای به موقع الزامات بهره برداری با مشکلات مختلفی روبرو است. تغییرات کمی و کیفی فاضلاب، ریزش های شدید ممکن است باعث در هم شکستن تاسیسات و سرریز شدن فاضلاب از طریق تخمین نادرست شود. با توجه به این موارد و همچین رفتارهای غیرخطی فرآیندهای تصفیه فاضلاب، تکنولوژی مناسب و مفیدی برای کنترل آنها نیازمند می باشد[1].

با گسترش تکنولوژی اطلاعات و ابزارهای اتوماتیک، حجم

* گروه مهندسی بهداشت محیط، دانشکده بهداشت دانشگاه علوم پزشکی یاسوج، یاسوج، ایران
ایمیل: **mmbaneshi@yahoo.com** - شماره تماس: ۰۹۱۷۳۰۸۳٤۲۹

بالایی از داده ها در تصفیه خانه های فاضلاب به صورت روزانه ثبت می گردد که می توان بر اساس این داده ها، تغییرات کمی و کیفی فاضلاب ورودی پیش بینی کرد و براساس آن، بهره بردار می تواند قبل از بروز مشکلات تصمیم های لازم را اتخاذ کند و بدین ترتیب، کنترل و بهره برداری مناسبی را اعمال نماید. بنابراین، مدلسازی و بهینه سازی فرآیندهای فاضلاب یک موضوع قابل اهمیت است، اگرچه داده های ارائه شده توسط تصفیه خانه های فاضلاب بسیار زیاد است، اما بطور کامل نمی تواند در جهت استخراج اطلاعات معنی دار به منظور بهبود عملکرد تصفیه خانه ها مورد استفاده قرار بگیرد[۲].

روش های مبتنی بر داده (داده کاوی)، بر اساس کشف دانش مفید از بین داده های اولیه عمل می کند. روش های اصلی داده کاوی دو دسته می باشند: توصیفی و پیش بینانه. وظایف توصیفی خواص عمومی داده ها را مشخص می کنند. هدف از توصیف، یافتن الگوهایی در مورد داده هاست که برای انسان قابل تفسیر باشد. وظایف پیش بینانه به منظور پیش بینی رفتارهای آینده آنها استفاده می شوند. منظور از پیش بینی به کارگیری چند متغیر یا فیلد در پایگاه داده برای پیش بینی مقادیر آینده یا ناشناخته دیگر متغیرهای مورد علاقه است[۳].

تا کنون روش های داده کاوی، موفقیت های زیادی را در زمینه کسب وکار، بازاریابی، تولیدات و علوم مهندسی و پزشکی کسب کرده است. با رویکرد مبتنی بر داده، فرآیند تصفیه می تواند با دقت زیاد به وسیله مدل های بدون محاسبات پیچیده ریاضی ارائه می شود. این مدل ها می توانند برای پیش بینی رفتار تصفیه خانه ها و تنظیماتی برای کنترل بهینه در جهت ذخیره ی انرژی و بهبود بهره وری از آن انجام داد. مثال هایی از کاربرد داده کاوی در تصفیه فاضلاب، مدلسازی اثر فرآیندهای بیولوژیکی با شبکه عصبی TDNN [۴]، پیش بینی مقدار پساب تولیدی با استفاده از پیش بینی کننده k-step[۵]، ترکیب مدلسازی ریاضی و شبکه عصبی مصنوعی و الگوریتم ژنتیک برای مدلسازی فرآیند تصفیه[۶]، مدلسازی فاضلاب صنعتی بوسیله شبکه عصبی مصنوعی[۷و۸] می باشد. یکی از عملکردهای داده کاوی، که در حیطه روش های پیش بینانه قرار می گیرد، دسته بندی می باشد. در این تحقیق، از الگوریتم های دسته بندی برای مدلسازی فرآیند استفاده شده و سعی شده است تا با این روش، مدل های آماری در مدلسازی فرآیند تصفیه خانه فاضلاب شهر یاسوج مورد بررسی قرار گیرد و کارآمدترین الگوریتم برای فرآیندهای فاضلاب در این تصفیه خانه مشخص گردد.

مواد وروش ها

تصفیه خانه فاضلاب شهر یاسوج در زمینی به مساحت ۳۰ هکتار برای جمعیت ۲۰۰ هزار نفر تا سال ۱۴۱۰، در دو مدول طراحی و از نوع لاگون هوادهی می باشد که در سال ۸۳ در مدار بهره برداری قرار گرفت. دبی فاضلاب ۵/۶ لیتر برثانیه پیش بینی شده است. مجموعه داده های مربوط به فرآیند تصفیه فاضلاب از شرکت آب و فاضلاب استان کهگیلویه وبویر احمد جمع آوری شد. در این مطالعه، از ۹ الگوریتم دسته بندی استفاده گردید (جدول۱). چهار متغیر (Chemical Oxygen Demand) COD، PH ، کدورت و (Total Suspended Solids) TSS در جهت تشخیص الگوریتم مورد بررسی قرار گرفتند. این متغیرها، با توجه به استاندارد خروجی فاضلاب، به دو طبقه در محدوده استاندارد (T) و خارج از استاندارد (F) قرار می گیرند. بنابراین بر اساس این استاندارد، متغیر هدف ساخته می شود. استانداردهای مربوط به این متغیرها را می توان در جدول شماره ۲ مشاهده نمود. در شروع کار، ۷۰ درصد داده ها به عنوان داده های آزمایشی و ۳۰ درصد به عنوان داده های آموزشی انتخاب شدند. در نهایت، مدل های مربوط به هر الگوریتم توسط چهار معیار مقدار درستی (Accuracy)، نمودار لیفت Lift، ماتریس تطابق (confusion *matrix*)، FN (False Negative) و FP (False Positive)

مورد ارزیابی قرار گرفتند. در این مطالعه از نرم افزار SPSS-16 برای تحلیل آمار توصیفی (فراوانی، درصد، میانگین و انحراف معیار) و همچنین از نرم افزار Ibm Spss Modeler 14.2 برای داده کاوی استفاده گردید.

جدول ۱: الگوریتم های دسته بندی مورد استفاده

نام الگوریتم	توصیف
الگوریتم رگرسیون منطقی یا لجستیک (Logistic regression algorithm)	رگرسیون منطقی یک روش آماری برای مدل سازی‌هایی که نتایج دودویی دارند، است. از رگرسیون لجستیک می‌توان به عنوان نوع دیگری از الگوریتم شبکه‌های عصبی نام برد
ماشین بردار پشتیبانی (Support vector machines - SVMs)	الگوریتمSVM، جز الگوریتم‌های تشخیص الگو دسته بندی می باشد. از الگوریتمSVM ، در هر جایی که نیاز به تشخیص الگو یا دسته بندی اشیا در کلاس‌های خاص باشد می توان استفاده کرد.
الگوریتم بیز (Naive Bayes Algorithm)	این الگوریتم بر پایه‌ی قضیه بیز برای مدل سازی پیش‌گویانه ارائه شده است. قضیه بیز از روشی برای دسته‌بندی پدیده‌ها بر پایه احتمال وقوع یا عدم وقوع یک پدیده استفاده می‌کند و احتمال رخ دادن یک پدیده محاسبه و دسته بندی می‌شود.
الگوریتم شبکه‌های عصبی (Neural Network Algorithm)	شبکه های عصبی از پرکاربردترین و عملی ترین روش‌های مدلسازی مسائل پیچیده و بزرگ که شامل صدها متغیر هستند، می‌باشد. شبکه های عصبی می‌توانند برای مسائل طبقه‌بندی (که خروجی یک کلاس است) یا مسائل رگرسیون (که خروجی یک مقدار عددی است) استفاده شوند.
الگوریتم درخت تصمیم (Decision Trees algorithm)	درخت تصمیم یکی از قوی‌ترین و پرکاربردترین الگوریتم‌های داده‌کاوی است که برای کاوش در داده‌ها و کشف دانش کاربرد دارد. این الگوریتم داده‌ها را به مجموعه‌های مشخصی تقسیم می‌کند. هر مجموعه شامل چندین زیر مجموعه از داده‌های کم و بیش همگن که دارای ویژگی‌های قابل پیش‌بینی هستند تقسیم می‌شود.
الگوریتم CHAID (Chi-squared Automatic Interaction Detector)	در این روش از مقدار P- Value آماره کای–دو مربوط به آزمون استقلال جداول توافقی استفاده می‌شود. از بین متغیرهای موجود، متغیری که دارای P- Value کوچکتری باشد در مرحله اول برای تقسیمات روی یک گره در نظر گرفته می شود.
الگوریتم QUEST (Quick Unbiased Efficient Statistical Trees)	درخت رده بندی حاصل از این الگوریتم دارای تقسیمات دوتایی بوده و ملاک تصمیم برای انتخاب متغیرها با استفاده از مقدار P- Value مربوط به آماره F آزمون ANOVA برای متغیرهای کمی و P- Value آماره کای– دو مربوط به جداول توافقی برای متغیرهای کیفی صورت می پذیرد.
الگوریتم C5	بهبود یافته الگوریتم C4.5 است. c4.5 الگوریتمی است که برای تولید یک درخت تصمیم استفاده می شود، درخت های تصمیم تولید شده توسط C4.5 می توانند برای دسته بندی استفاده شوند.
تجزیه و تحلیل مشخص خطی Linear Discriminant Analysis	آنالیز افتراقی خطی بسیار به تحلیل واریانس و تحلیل رگرسیونی نزدیک است؛ در هر سه این روش‌های آماری متغیر وابسته به صورت یک ترکیب خطی از متغیرهای دیگر مدل‌سازی می‌شود.

جدول ۲: استاندارد های متغیر های مورد استفاده[۹]

متغیر	محدوده غیرمجاز غلظت متغیر(کلاس F)	محدوده مجاز غلظت متغیر (کلاس T)

COD	>60(لحظه ای 100)	60(لحظه ای 100)
TSS	>40(لحظه ای 60)	40(لحظه ای 60)
PH	>6/5-8/5<	6/5-8/5
کدورت	50<	50

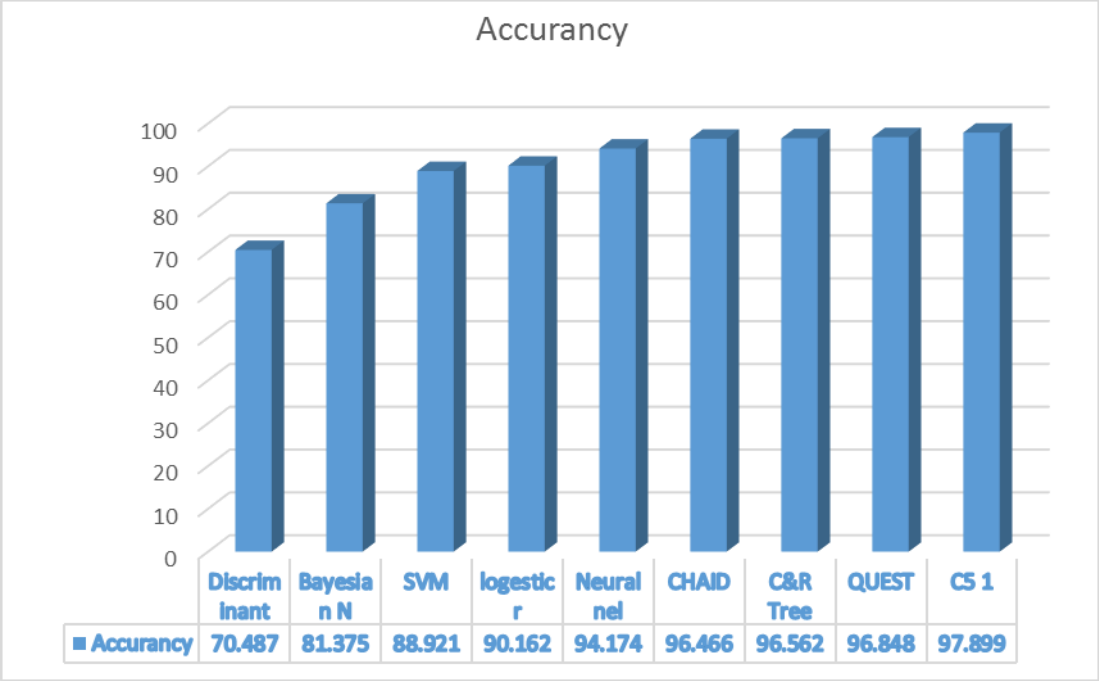
جدول ۳: آمار توصیفی مربوط به متغیرها

نام متغیر	کمترین	بیشترین	میانگین	انحراف معیار
COD	۰	۱٦٤/۰۱	۸٥/۷۰	٦/۷٦
TSS	۰	۱۳۲/۲٤	۱٥/٥۱	٦/۱٤
PH	۳/۷٦	۱۰	۷/۱۷	۳/۱٦
کدورت	۱٦/۹٥	۱۸۹/۷۳	۲۸	۸/۷٥

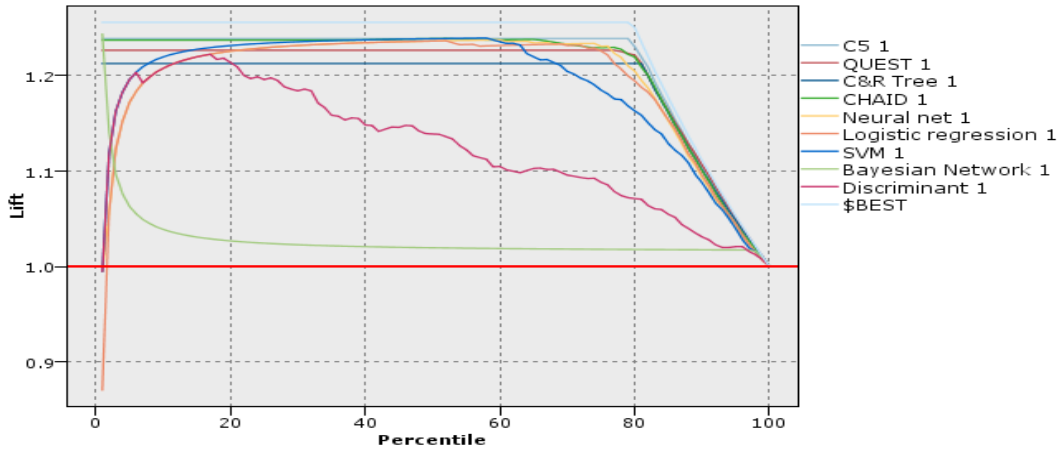
یافته‌ها و بحث

در مرحله اول، تعداد ٦۰۰۰ نمونه ثبت شده مربوط به ٤ متغیر COD، TSS، PH و کدورت مورد بررسی قرار گرفتند. متغیرهای استخراج شده توسط استاندارد کشوری پساب تصفیه خانه به دو دسته استاندارد و غیر استاندارد طبقه بندی گردید. با حذف رکوردهایی که مقادیر آنها به دلیل عدم اندازه گیری و یا خرابی دستگاه، در روز نمونه گیری ثبت نشده بود، حذف و در نهایت ۳۳۰٦ رکورد جمع آوری گردید. از این تعداد، ٦٦٥ (۲۰/۱درصد) داده خارج از استاندارد کشوری و ۲٦٤۰ (۷۹/۹ درصد) از داده ها در محدوده استاندارد قرار داشتند. جدول شماره ۳، شاخص های آمار توصیفی متغیرها را نشان می دهد.

مقایسه مقادیر پارامتر، میزان درستی مربوط به هر ۹ الگوریتم اعمال شده بروی مجموعه داده های پساب خروجی تصفیه خانه در نمودار ۱ نشان داده شده است.



نمودار ۱: مقادیر پارامتر میزان درستی مربوط به الگوریتم ها



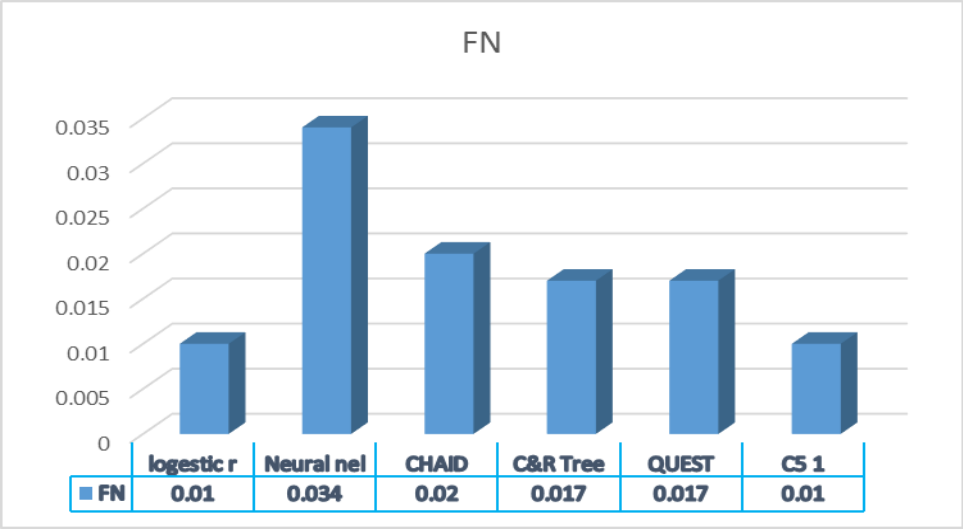
نمودار۲: نمودار لیفت مربوط به ارزیابی الگوریتم ها

با توجه به نمودار ۱، الگوریتم های C5 1، QUEST و C&R دارای بیشترین مقدار برای پارامتر درستی بوده، الگوریتم های SVM، شبکه بیزی و تحلیل تفکیکی دارای کمترین مقدار هستند. در نتیجه توانایی کمی در مدلسازی این سیستم دارند.

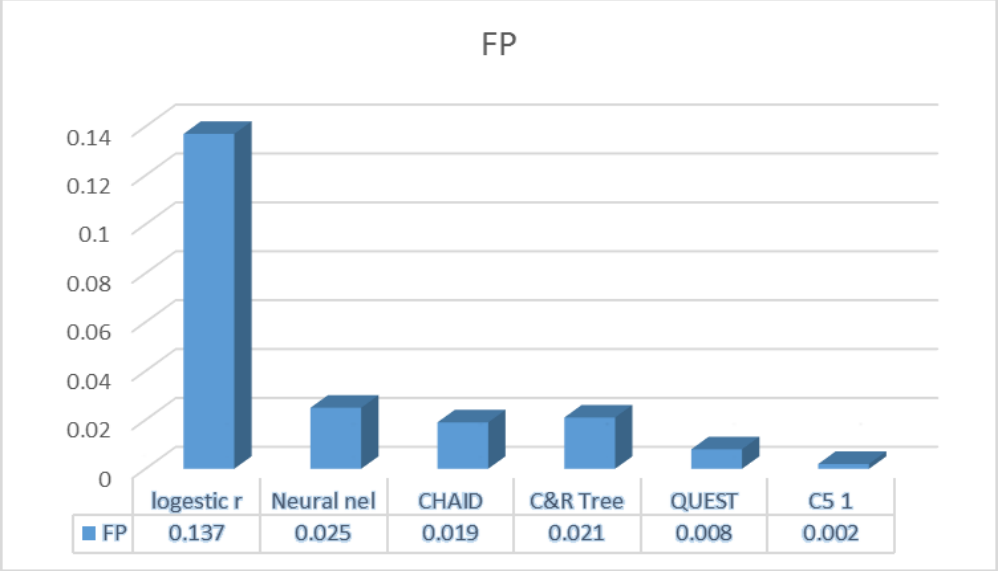
نمودار ۲ نشان می دهد که الگوریتم c5 بهترین حالت و بالاترین توانایی در مدلسازی این سیستم را دارد. با توجه به این نمودار، الگوریتم هایی که به خوبی با داده ها انطباق داشته باشند باید در سمت چپ، بیشترین مقدار و بعد از گذشت حالت ایستا با شیب تند در سمت چپ نزول کرده و به ۱ برسد.

همانطور که از نمودار ۳ می توان دریافت، الگوریتم های C5 و لجستیک به ترتیب دارای کمترین و یا به عبارتی بهترین

مقـادیر FN هسـتند. الگـوریتم هـای شـبکه عصـبی و درخـت CHAID به ترتیب دارای بیشترین مقدار FN بوده است.



نمودار ۳: نمودار مقادیر FN الگوریتم ها



نمودار٤: نمودار مقادیر FP الگوریتم ها

در رابطه با مقادیرFP الگوریتم هـا، مـی تـوان نمـودار ٤ را مورد بررسی قرار داد. همان طور که در شکل دیـده مـی شـود، کمترین مقادیر FP به الگوریتم C5 تعلق دارد و بیشترین میـزان آن، متعلق به الگوریتم لجستیک می باشد.

جدول ٤: جدول ماتریس تطابق مربوط به ارزیابی الگوریتم ها

الگوریتم	شرایط	غیر مجاز	مجاز
Logestic r	مجاز	۱۱	۸۳۱
	غیر مجاز	٦۱	۱٤٤
Neural nel	مجاز	۳٦	۸۰٦
	غیر مجاز	۱۷۸	۲۷
CHAID	مجاز	۲۱	۸۲۱
	غیر مجاز	۱۸۵	۲۰
C&R Tree	مجاز	۱۸	۸۲٤
	غیر مجاز	٤۸۳	۲۲
QUEST	مجاز	۱۸	۸۲٤
	غیر مجاز	۱۹٦	۹
C5	مجاز	۱۱	۸۳۱
	غیر مجاز	۲۰۲	۳

با توجه به جدول ٤، تطابق ماتریس الگوریتم C5 بهترین پیش بینی را انجام داده است و از ۲۰۵ رکوردی که خروجی آنها غیر مجاز بوده است، تنها ۳ رکورد آن را به اشتباه مجاز در نظر گرفته است و ۲۰۲ رکورد را به درستی پیش بینی کرده است. اما در مورد روش لجستیک همان طور که مشاهده می کنید از ۲۰۵ حالت غیر مجاز تنها ٦۱ رکورد را به درستی تشخیص داده است. در مطالعه ای که توسط سلطانی نژاد و همکارانش بر روی فرآیند تصفیه خانه شهر منرسای اسپانیا انجام شد، بهترین الگوریتم برای این تصفیه خانه را رگرسیون لجستیکی معرفی کردند.[۱۰]

نتیجه‌گیری

بررسی و مقایسه نتایج حاصل از الگوریتم های مختلف در تشخیص بهترین و کارآمدترین الگوریتم در مدلسازی فرآیندهای تصفیه فاضلاب نشان داد که نتایج نهایی هر الگوریتم بستگی زیادی به صفات یا متغیر های ورودی به الگوریتم و همچنین شرایط هر تصفیه خانه دارد. این نتیجه در مطالعات انجام شده قبلی بر روی روش های مبتنی بر داده نیز دست آمده است. بنابراین هرگز نمی توان یک الگوریتم خاص را به عنوان بهینه معرفی نمود وممکن است که در تصفیه خانه های مختلف متفاوت باشد. از این رو، با انتخاب متغیر های تاثیر گذار مناسب در هر تصفیه خانه و خصوصیات مرتبط با آنها می توان الگوریتمی را به عنوان بهترین و کارآمدترین الگوریتم انتخاب نمود. دراین پژوهش الگوریتم C5 به عنوان بهترین الگوریتم برای تصفیه خانه شهر یاسوج انتخاب‌گردید.

تشکر وقدردانی

نویسندگان این مقاله بر خود لازم می دانند تا از زحمات و کارکنان مسئولان شرکت آب و فاضلاب استان کهگیلویه و بویر احمد که نهایت همکاری را در اجرای این طرح داشته‌اند، تشکر و قدردانی نمایند.

 [DOI: 10.18869/acadpub.jehe.2.3.186]

منابع

1. Verma A, Wei X, Kusiak A. Predicting the total suspended solids in wastewater: a data-mining approach. Eng Appl Artif Intel 2013;26(4): 1366-72.
2. Li X, Zeng G, Huang G, Li J, Jiang R. Short-term prediction of the influent quantity time series of wastewater treatment plant based on a chaos neural network model. Frontiers Environ Sci Eng China. 2007;1(3): 334-8.
3. Han J, Kamber M, Pei J. Data mining: concepts and techniques: concepts and techniques: Elsevier; 2011.
4. Zhu J, Zurcher J, Rao M, Meng MQ. An on-line wastewater quality predication system based on a time-delay neural network. Eng Appl Artif Intel 1998;11(6): 747-58.
5. Tan P, Berger C, Dabke K, Mein R. Recursive identification and adaptive prediction of wastewater flows. Automatica 1991;27(5): 761-8.
6. Wan J, Huang M, Ma Y, Guo W, Wang Y, Zhang H, et al. Prediction of effluent quality of a paper mill wastewater treatment using an adaptive network-based fuzzy inference system. Appl Soft Comput 2011;11(3): 3238-46.
7. Oliveira-Esquerre K, Mori M, Bruns R. Simulation of an industrial wastewater treatment plant using artificial neural networks and principal components analysis. Brazilian J Chem Eng 2002;19(4): 365-70.
8. Gontarski C, Rodrigues P, Mori M, Prenem L. Simulation of an industrial wastewater treatment plant using artificial neural networks. Comput Chem Eng 2000;24(2): 1719-23.
9. Mohammad SA, Alireza R. Human Environmental Law, Regulation Criteria and Standard. 1, editor. Tehran, Iran: Hack; 2012: 277-278.
10. Soltaninejad A, Afsahi MM, Nakhaeizadeh G, Amiri MC. Evaluation of statistical and artificial intelligence methods in wastewater treatment process modeling. National Conference on Water and Wastewater Engineering (ncwwe) 2012.

Application of Statistical Model in Wastewater Treatment Process Modeling Using Data Analysis

Alireza Raygan Shirazinezhad[1], Morteza Zare[2], Fahime Zare[3], Mohammad Mehdi Baneshi*[1], Soheila Rezaei[2]

1. Social Determinants of Health Research Center, Yasuj University of Medical Sciences, Yasuj, Iran

2. Department of Environmental Health Engineering, School of Public Health Yasouj University of Medical Sciences, Yasouj, Iran

3. Department of Information Technology, Shahid Beheshti University, Tehran, Iran

E-mail: mmbaneshi@yahoo.com

Received: 15 Jan 2015 ; Accepted:21 Apr 2015

ABSTRACT

Background: Wastewater treatment includes very complex and interrelated physical, chemical and biological processes which using data analysis techniques can be rigorously modeled by a non-complex mathematical calculation models.

Materials and Methods: In this study, data on wastewater treatment processes from water and wastewater company of Kohgiluyeh and Boyer Ahmad were used. A total of 3306 data for COD, TSS, PH and turbidity were collected, then analyzed by SPSS-16 software (descriptive statistics) and data analysis IBM SPSS Modeler 14.2, through 9 algorithm.

Results: According to the results on logistic regression algorithms, neural networks, Bayesian networks, discriminant analysis, decision tree C5, tree C & R, CHAID, QUEST and SVM had accuracy precision of 90.16, 94.17, 81.37, 70.48, 97.89, 96.56, 96.46, 96.84 and 88.92, respectively.

Discussion and conclusion: The C5 algorithm as the best and most applicable algorithms for modeling of wastewater treatment processes were chosen carefully with accuracy of 97.899 and the most influential variables in this model were PH, COD, TSS and turbidity.

Keywords: Wastewater Treatment Process Modeling, Data Analyzing, Classification, Kohgiluyeh and Boyer Ahmad