نشریه علمی پژوهشی مهندسی فناوری اطلاعات مکانی

سال سوم، شماره چهارم، زمستان ۱۳۹۴

Vol.3, No.4, Winter 2016

۴۳-۶۳

ارائه‌ی روشی پویا برای پیش‌بینی مکانی-زمانی آلودگی هوای شهر تهران بر مبنای ماشین بردار پشتیبان

زینب قائمی[۱]*، مهدی فرنقی[۲]، عباس علیمحمدی[۳]

۱- دانشجوی کارشناسی ارشد سیستم‌های اطلاعات مکانی دانشگاه صنعتی خواجه نصیرالدین طوسی

۲- استادیار دانشکده مهندسی نقشه‌برداری دانشگاه صنعتی خواجه نصیرالدین طوسی

۳- دانشیار دانشکده مهندسی نقشه‌برداری دانشگاه صنعتی خواجه نصیرالدین طوسی

تاریخ دریافت مقاله ۱۳۹۴/۰۵/۲۱ تاریخ پذیرش مقاله: ۱۳۹۴/۱۱/۱۰

چکیده

با توجه به آثار سوء آلودگی هوا بر سلامت انسان‌ها و محیط، پیش‌بینی و مدلسازی این پدیده از جمله مسائل مهم در چند دهه‌ی گذشته بوده است. دینامیک غیرخطی و حجم بالای داده‌های آلودگی هوا، مشکلات پیش‌بینی این پدیده‌ی پیچیده را، بویژه در پردازش‌های پویا، دوچندان کرده است. هدف این پژوهش، ارائه‌ی الگوریتمی برخط است که بتواند با حل مشکلات روش‌های پیشین در پیش‌بینی برخط آلودگی هوا، سری زمانی آلودگی هوای شهر تهران را به صورت پویا پیش‌بینی کند. الگوریتم برخط ارائه شده بر مبنای ماشین بردار پشتیبان طراحی شده است. در الگوریتم ارائه شده، پیش‌بینی مبتنی بر داده‌های جریانی جمع‌آوری شده توسط سنجنده‌های آلودگی هوا، سنجنده‌های هواشناسی و همچنین داده‌های مکانی همچون ترافیک، ارتفاع متوسط منطقه و ویژگی‌های سطح زمین انجام می‌شود. نتایج حاصل شده بیانگر دقت مناسب الگوریتم برخط، جهت پیش‌بینی پویای آلودگی هوای شهر تهران می‌باشد. استفاده از داده‌های یک سال جهت انجام تست، دقت ۰.۷۱ و خطای جذر میانگین مربعات ۰.۵۴ و ضریب تعیین ۰.۸۱ را حاصل کرده است. افزون بر دقت مناسب، سرعت بالای پردازش‌ها در الگوریتم برخط، کارایی این الگوریتم را برای طراحی سیستمی آنلاین جهت پیش‌بینی آلودگی هوای شهر تهران برای چند ساعت آینده به اثبات می‌رساند.

واژه‌های کلیدی: پیش‌بینی پویای آلودگی هوای شهری، ماشین بردار پشتیبان برخط، داده‌های حجیم، سری زمانی، سیستم اطلاعات مکانی.

* نویسنده مکاتبه کننده: تهران - خیابان ولیعصر - تقاطع میرداماد - روبروی ساختمان اسکان، دانشکده ژئودزی و ژئوماتیک دانشگاه خواجه نصیرالدین طوسی

تلفن: ۴-۸۸۷۷۹۴۷۳

Email: zghaemi@mail.kntu.ac.ir

۱- مقدمه

امروزه توسعه‌ی شهرنشینی، گسترش صنایع و افزایش بی‌رویه‌ی استفاده از وسایل نقلیه منجر به گسترش آلودگی‌های زیست محیطی شده است. در این میان، افزایش آلودگی هوا، بدلیل آثار زیانباری که بر سلامت انسان‌ها و سایر جانداران دارد، در چند دهه‌ی اخیر از دغدغه‌های اصلی دانشمندان و محققان بوده است. این موضوع به صورت ویژه در پیش‌بینی وضعیت آلودگی هوا در شهرهای صنعتی و پرجمعیت همچون تهران از اهمیت بسزایی برخوردار است. پیش‌بینی آلودگی هوا می‌تواند اطلاعات مورد نیاز مدیران را جهت ارائه‌ی راهکاری برای کاهش انتشار مواد آلاینده فراهم کند. همچنین پیش‌بینی آلودگی هوا، به ویژه بصورت پویا، این امکان را فراهم می‌کند که بتوان آلودگی را در لحظه و بصورت آنی برای چند ساعت آینده پیش‌بینی و مناطق پرخطر از نظر آلودگی را شناسایی نمود. با ایجاد ایستگاه‌های سنجش آلودگی در سطح شهرها، داده‌های مربوط به غلظت آلاینده‌ها در بازه‌های ساعتی و روزانه، بطور پیوسته و در حجم بالا برداشت می‌شوند. این داده‌ها می‌توانند بستر مناسب برای پیش‌بینی آلودگی هوا را فراهم سازند.

پیش‌بینی آلودگی هوا از جمله مسائل پیچیده‌ای است که بصورت غیرخطی در حال تغییر می‌باشد [۱] و پارامترهای متعددی در آن تاثیرگذار هستند [۲, ۳]. این پیچیدگی محققان را بر آن داشته است تا روش‌های نوینی را برای پیش‌بینی آلودگی هوا ارائه دهند. پیش‌بینی سری‌های زمانی یکی از روش‌های متداول برای پیش‌بینی آلودگی هوا می‌باشد. مجموعه‌ای از مشاهدات وابسته که در طول زمان و به صورت پی در پی جمع‌آوری می‌شوند را سری زمانی می‌نامند. با استفاده از سری‌های زمانی، می‌توان مقدار متغیر را در آینده و بر اساس داده‌های کنونی و گذشته، تخمین زد [۴]. ورودی این سری زمانی داده‌های دوره‌ای گذشته و حال آلودگی هوا و خروجی آن میزان آلودگی برای چند ساعت آینده می‌باشد.

تا کنون مطالعات بسیاری در زمینه مدلسازی سری زمای آلودگی هوا انجام شده است. بطور کلی این روش‌ها را می‌توان در دو گروه روش‌های قطعی و روش‌های آماری طبقه‌بندی کرد [۱]. مدل‌های انتشار از جمله روش‌های قطعی هستند که در مناطق مختلفی برای مدلسازی و پایش آلودگی هوا توسعه داده شده‌اند [۵, ۶]. اما خروجی این مدل‌ها بسیار وابسته به داده‌های ورودی بوده و به منظور استفاده از آنها باید چگونگی پراکنش و انتشار مواد آلاینده در جو با دقت بالایی در دسترس باشد [۷, ۸]. بنابراین استفاده از این مدل‌ها، در شرایطی که داده‌های کافی و دقیق در دسترس نیستند، با مشکلات فراوانی روبرو خواهد بود. با توجه به این مهم که جمع‌آوری اطلاعات مورد نیاز مدل‌های انتشار کاری سخت و در.ابعاد بزرگ غیرممکن می‌باشد، محققان به استفاده از روش‌هایی برتر از جمله روش‌های آماری روی آورده‌اند [۹].

مدل‌های هوش مصنوعی از جمله روش‌هایی هستند که کارایی خود را مدلسازی مسائل پیچیده نشان داده‌اند. شبکه‌های عصبی، یکی از پرکاربردترین روش‌های هوش مصنوعی، به دلیل قابلیت بالا در انجام پردازش‌های پیچیده بطور وسیعی در پیش‌بینی سری‌های زمانی آلودگی هوا مورد استفاده قرار گرفته‌اند [۱۰, ۱۱]. در تحقیقی واحید[1] و همکاران از شبکه‌های عصبی برای حل مشکل توزیع نامناسب ایستگاه‌های پایش آلودگی هوا استفاده کرده‌اند [۱۲]. در مطالعه‌ای دیگر النگسینی[2] و همکاران جهت پیش‌بینی آلودگی هوای شهری، از داده‌های ترافیکی به همراه داده‌های هواشناسی به‌عنوان ورودی شبکه‌ی عصبی استفاده کرده‌اند [۱۱]. با وجود توانایی

[1] Wahid
[2] Elangasinghe

بالای شبکه‌های عصبی در مدلسازی مسائل غیرخطی، این روش‌ها همچنان با محدودیت‌هایی همچون بیش‌برازش[1]، بهینه محلی[2]، پردازش‌های زمانبر و تعمیم ضعیف[3] در مواجهه با حجم بالای داده‌های ورودی روبرو هستند [۱۳, ۱۴]. توانایی ماشین بردار پشتیبان، که در سال‌های اخیر به عنوان الگوریتمی نوین ارائه شده، در حل بسیاری از مسائل پیچیده به اثبات رسیده است. این روش، به دلیل قابلیت‌های محاسباتی و توانایی تعمیم بالا، توانسته بسیاری از مشکلات روش‌های کنونی را در پیش‌بینی سری‌های زمانی پیچیده برطرف کند [۱۵, ۱۶]. در تحقیقی که توسط ریکارد[4] انجام گرفته است [۱۷]، کارایی شبکه‌های عصبی و ماشین بردار پشتیبان در پیش‌بینی آلودگی هوا مقایسه شده که نتایج حاصله بیانگر عملکرد بهتر روش ماشین بردار پشتیبان نسبت به شبکه عصبی بوده است. جوز[5] و همکاران نیز کارایی دو روش ماشین بردار پشتیبان و شبکه عصبی در پیش‌بینی سری زمانی NO_2 و NO ناشی از ترافیک را مورد مقایسه قرار داده‌اند. برای کاهش ابعاد داده‌های ورودی از ترکیب آنالیز مولفه اصلی[6] با این روش‌ها استفاده شده است. نتایج حاصل شده عملکرد بهتر ماشین بردار پشتیبان نسبت به شبکه عصبی را اثبات نموده است [۱۸].

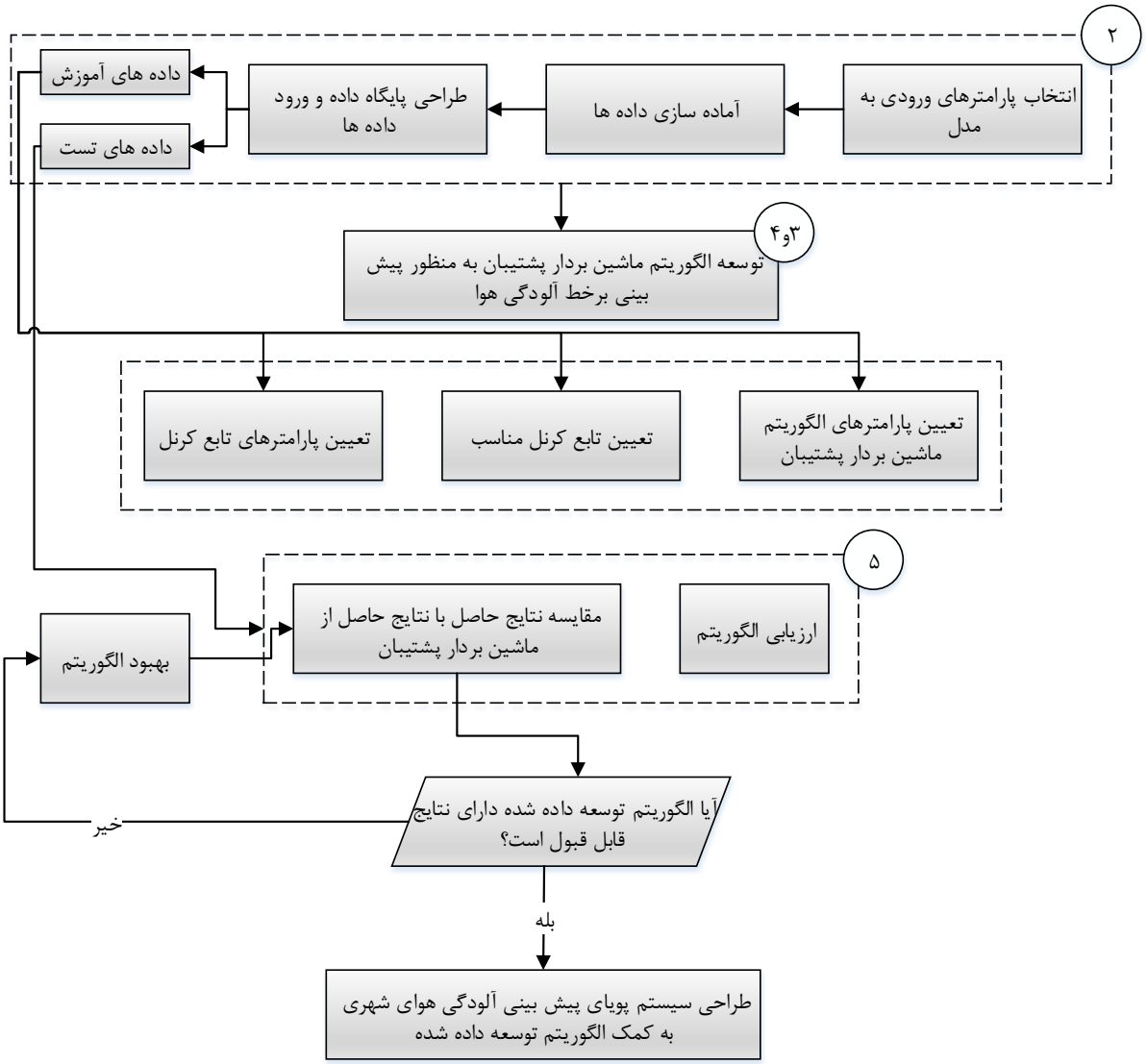
علاوه بر پیچیدگی رفتار سری زمانی آلودگی هوا، توسعه‌ی روش‌های جمع‌آوری داده همچون ایستگاه‌های پایش آلودگی هوا و سنسورهای تعبیه شده در تلفن‌های همراه داده‌های متنوعی از پارامترهای موثر در کیفیت هوا را در اختیار محققان قرار می‌دهد. این داده‌ها در حجم بالا و بصورت آنی و جریانی[7] در فرمت‌های مختلف و با دقت متفاوتی جمع‌آوری می‌شوند. این خصوصیات موجب می‌شود روش‌های سنتی پردازش و تحلیل داده‌ها، قادر به حل تمام ابعاد داده‌های کیفیت هوا نباشند. روش ماشین بردار پشتیبان نیز، با وجود توانایی بالا در حل مسائل پیچیده، با محدودیت‌هایی در زمینه‌ی پاسخگویی به حجم بالای داده‌های جریانی، مواجه است بطوریکه با افزایش حجم داده‌های ورودی، زمان پردازش‌ها نیز بطور قابل توجهی افزایش می‌یابد [۱۹]. این مشکلات موجب کاهش کارائی این روش در رویارویی با داده‌های حجیم و جریانی می‌گردد. به منظور پاسخگویی به مشکلات فوق، استفاده از الگوریتم‌های برخط که قادر به پردازش حجم بالایی از داده‌های جریانی و در مدت زمانی قابل قبول باشند، ضروری به‌نظر می‌رسد. در همین راستا در سال‌های اخیر، الگوریتم‌هایی بر مبنای ماشین بردار پشتیبان طراحی شده‌اند که می‌توانند جهت انجام پردازش‌های پویا و پیچیده بویژه پیش‌بینی برخط آلودگی هوا مورد استفاده قرار بگیرند.

هدف این پژوهش ارائه‌ی سیستمی جهت پیش‌بینی برخط آلودگی هوای شهر تهران برای ۲۴ ساعت آینده می‌باشد. جهت انجام پیش‌بینی، از یک الگوریتم برخط که بر مبنای ماشین بردار پشتیبان طراحی شده است، استفاده می‌شود. سری‌های زمانی مربوط به غلظت آلاینده‌ها و وضعیت هواشناسی به همراه داده‌های مکانی بطور پیوسته و جریانی به الگوریتم ارائه شده، معرفی می‌شوند. الگوریتم بر اساس داده‌های جریانی دریافت شده به صورت پیوسته آموزش می‌بیند. برای حل محدودیت‌های روش ماشین بردار پشتیبان در کار با داده‌های جریانی و حجیم، داده‌های اضافی که تاثیری در آموزش مدل ندارند، از پردازش‌ها حذف می‌گردند. همین‌طور از راهکاری در جریان آموزش استفاده می‌شود که با ورود داده‌ی آموزشی جدید، نیازی به آموزش دوباره‌ی مدل وجود نداشته باشد. حذف داده‌های اضافی، موجب کاهش حجم داده‌های وارد شده به مدل برای آموزش می‌شود.

[3] overfitting
[4] Local minimum
[5] Poor generalization
[6] Reikard
[7] Juhos
[8] Principal Component Analysis
[9] Streaming

مدل آموزش دیده در هر لحظه امکان پیش‌بینی آلودگی هوا در هر نقطه از شهر تهران را برای ۲۴ ساعت آینده فراهم می‌کند. جهت ارزیابی روش ارائه شده، نتایج حاصل از الگوریتم برخط با نتایج حاصل از ماشین بردار پشتیبان معمولی از نظر زمان پردازش و دقت مقایسه شده‌اند.

در ادامه ساختار مقاله به این صورت خواهد بود. در بخش ۲ داده‌های مورد نظر و منطقه‌ی مطالعاتی معرفی شده و نحوه‌ی آماده‌سازی داده‌ها تشریح می‌شود. مبانی نظری ماشین بردار پشتیبان و الگوریتم برخط استفاده شده در بخش ۳ توضیح داده شده است. در بخش ۴ نحوه‌ی طراحی و پیاده‌سازی سیستم پویای پیش‌بینی آلودگی هوا برای شهر تهران توضیح داده می‌شود. نتایج حاصل از پیاده‌سازی مدل ارائه شده در بخش ۵ مورد بحث قرار گرفته و در بخش ۶ نتیجه‌گیری و پیشنهادات مورد نظر برای کارهای آتی ارائه می‌گردد. کلیه‌ی این مراحل در شکل ۱ نمایش داده شده‌اند. جزئیات مربوط به هر قسمت در بخش ذکر شده در شکل تشریح شده است.



شکل ۱: مراحل انجام تحقیق

۲- روش‌ها

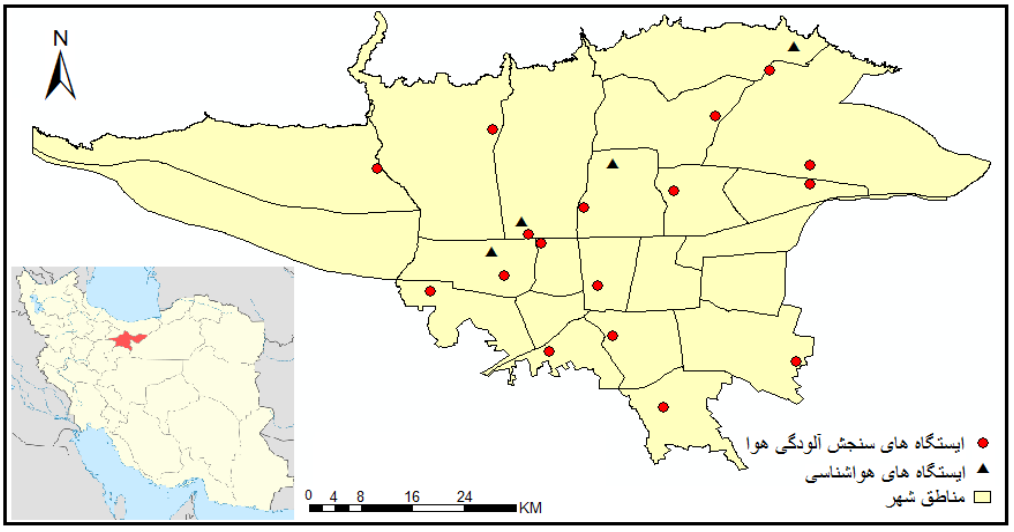
در این بخش منطقه‌ی مورد مطالعه و داده‌های استفاده شده معرفی می‌شوند. همچنین نحوه‌ی آماده‌سازی داده‌های استفاده شده تشریح می‌گردد.

۲-۱- منطقه مطالعاتی و داده‌های مورد استفاده

شهر تهران به‌عنوان پایتخت ایران، مهمترین کلانشهر و مرکز سیاسی و تجاری کشور محسوب می‌شود که توسط رشته‌کوه‌های البرز از سمت شمال و دشت کویر از سمت جنوب احاطه شده است. موقعیت جغرافیایی منطقه از یک سو و گسترش صنایع و افزایش استفاده از وسایل نقلیه از سوی دیگر موجب تشدید آلودگی هوا در این شهر شده است که این امر ضرورت پیش‌بینی آلودگی هوا برای شهر تهران را نمایان می‌کند.

جهت پیش‌بینی آلودگی هوا، کل شهر تهران به‌عنوان منطقه مطالعاتی انتخاب شده است. شکل ۲ موقعیت جغرافیایی شهر تهران را به‌همراه نحوه‌ی توزیع ایستگاه‌های سنجش آلودگی هوا و هواشناسی در سطح شهر نمایش می‌دهد.

برای انجام پیش‌بینی دقیق نیاز است پارامترهای تاثیرگذار در آلودگی هوا به درستی شناسایی شوند. داده‌های استفاده شده در این تحقیق شامل داده‌های ساعتی غلظت آلاینده‌ها، برداشت شده از ۲۱ ایستگاه سنجش آلودگی هوا بوده که شامل غلظت CO، O_3، PM_{10}، SO_2 و NO_2 [۱۵] می‌باشد. هم‌چنین داده‌های هواشناسی مربوط به ۴ ایستگاه هواشناسی شامل دما، فشار، رطوبت نسبی، سرعت باد و پوشش ابر [۱۳، ۲۰] نیز جهت بهبود دقت روش ارائه شده، مورد استفاده قرار گرفته‌اند. داده‌های غلظت آلاینده‌ها توسط شرکت کنترل ترافیک شهر تهران و داده‌های هواشناسی توسط مرکز تحقیقات هواشناسی تهران در دسترس قرار گرفته‌اند. داده‌های استفاده شده در بازه‌ی زمانی ۴ ساله از سال ۱۳۸۹ تا ۱۳۹۳ جمع‌آوری شده‌اند. همچنین، با توجه به نقش بسیار مهم مکان در توزیع آلودگی هوا در سطح شهر، در این تحقیق، پارامترهای فاصله از جاده، ارتفاع متوسط منطقه و توپوگرافی سطح زمین [۳]، به عنوان پارمترهای مکانی تاثیرگذار در نظر گرفته شده‌اند.



شکل ۲: موقعیت جغرافیایی شهر تهران و توزیع ایستگاه‌های هواشناسی و سنجش آلودگی

۲-۲- آماده‌سازی داده‌ها

در این بخش، نحوه‌ی آماده‌سازی داده‌های آلودگی هوا، هواشناسی و داده‌های مکانی مورد استفاده به تفصیل تشریح می‌شود.

۲-۲-۱- شاخص آلودگی هوا[11]

شاخص کیفیت هوا که از جمله شاخص‌های اصلی در بیان میزان آلودگی می‌باشد، به عنوان فاکتور اصلی در پیش‌بینی آلودگی هوا مورد استفاده قرار می‌گیرد. داده‌هایی که توسط ایستگاه‌های سنجش آلودگی برداشت می‌شوند غلظت آلاینده‌ها را در اختیار قرار می‌دهند و شاخص مربوط به هر یک از آلاینده‌ها را ارائه نمی‌کنند. غلظت آلاینده‌های مختلف با واحدهای متفاوتی ارائه می‌شود که به منظور استفاده از غلظت آلاینده‌ها در پیش‌بینی آلودگی هوا، لازم است شاخص مربوط به هر آلاینده، به کمک غلظت آن آلاینده محاسبه شود. بنابراین، اولین مرحله در آماده‌سازی داده‌ها، محاسبه شاخص آلودگی هوا برای تمامی آلاینده‌ها و ارائه شاخص آلودگی کلی بر اساس تلفیق شاخص‌های مربوط به هر یک از آلاینده‌ها می‌باشد. شاخص کیفیت آلودگی در ایران بر اساس استاندارد ارائه شده توسط سازمان حفاظت محیط زیست ایالات متحده[12] محاسبه می‌شود. شاخص کیفیت آلودگی برای هر یک از آلاینده‌ها با استفاده از معادله (۱) که توسط این سازمان ارائه گردیده، محاسبه می‌شود [۲۱]. پس از محاسبه شاخص آلودگی هوا، این شاخص بر اساس جدولی استاندارد [۲۱] به هفت کلاس طبقه‌بندی می‌شود. این هفت کلاس در جدول۱ نمایش داده شده‌اند. در نهایت این کلاس به عنوان ورودی مدل برای پیش‌بینی آلودگی هوا مورد استفاده قرار می‌گیرد.

رابطه (۱) $$I = \frac{I_{high}-I_{low}}{C_{high}-C_{low}}(C - C_{low}) + I_{low}$$

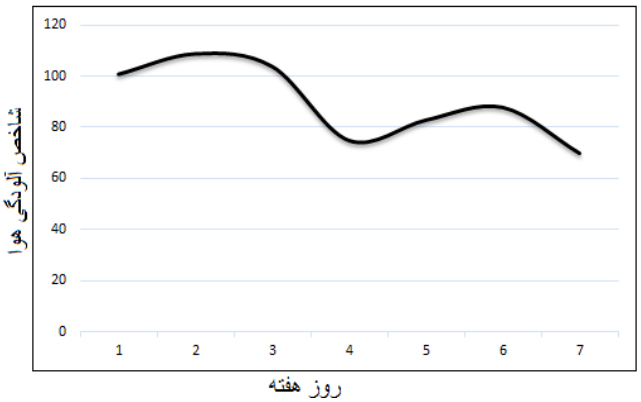
در معادله (۱) I شاخص کیفیت هوا، C غلظت آلاینده، C_{low}، C_{high}، I_{low} و I_{High} نقاط انفصال مربوط به آلاینده هستند (منظور از نقاط انفصال نقاطی است که محدوده‌ی بازه‌ی مورد نظر برای طبقه‌بندی آلاینده در یک کلاس را مشخص می‌کند) که از جداول ارائه شده توسط سازمان حفاظت محیط زیست استخراج می‌شوند [۲۱].

جدول۱: جدول طبقه‌بندی شاخص کیفیت هوا [۲۱]

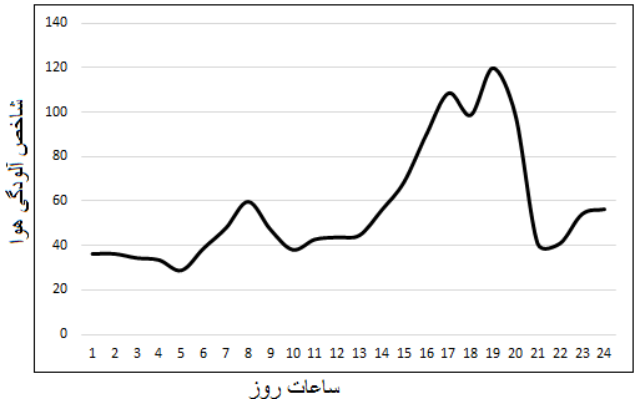
سطح اهمیت بهداشتی	شاخص کیفیت هوا
پاک	۰-۵۰
سالم	۵۱-۱۰۰
ناسالم برای گروه حساس	۱۰۱-۱۵۰
ناسالم	۱۵۱-۲۰۰
بسیار ناسالم	۲۰۱-۳۰۰
خطرناک	۳۰۱-۵۰۰
بسیار خطرناک	۴۰۱-۵۰۰

میزان شاخص آلودگی هوا در ساعات مختلف روز و روزهای مختلف هفته متفاوت می‌باشد. بر طبق بررسی‌های انجام شده میزان آلودگی هوا در سطح شهر تهران در اواخر هفته و روزهای تعطیل، به دلیل ترافیک کمتر، کاهش می‌یابد. شکل ۳ روند تغییر شاخص آلودگی هوا در طول هفته را نمایش می‌دهد. همانطور که شکل ۳ نشان می‌دهد میزان آلودگی هوا در اواخر هفته نسبت به روزهای ابتدایی کاهش می‌یابد. همچنین، ساعت اوج ترافیک در شهر تهران بین ساعات ۶ تا ۱۰ صبح و ۴ تا ۸ بعد از ظهر می‌باشد که ترافیک بیشتر موجب افزایش آلودگی هوا در این ساعات می‌شود. تغییرات آلودگی هوا در طول روز و ساعات اوج آلودگی در شکل ۴ نمایش داده شده است. بنابراین در این تحقیق، ساعت و ورز هفته به‌عنوان پارامترهای تاثیرگذار در پیش‌بینی آلودگی هوا در نظر گرفته شده‌اند.

[11] Air Quality Index (AQI)
[12] The United States Environmental Protection Agency (EPA)



شکل ۳: تغییرات میزان آلودگی هوا در طول هفته (این نمودار با میانگین‌گیری از داده‌های شش ماه آلودگی هوا ترسیم شده است)



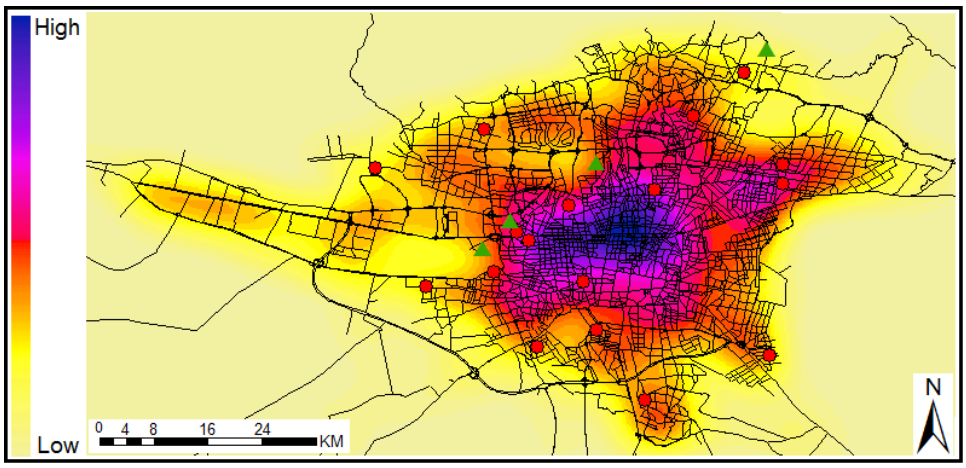
شکل ۴: تغییرات آلودگی هوا در طول روز (این نمودار با میانگین‌گیری از داده‌های یک ماه آلودگی هوا ترسیم شده است)

۲-۲-۲- داده‌های مکانی

فاصله از جاده: ترافیک به‌عنوان یکی از منابع اصلی و از جمله پارامترهای تاثیرگذار در آلودگی هوای شهری، بخصوص کلانشهر تهران، شناخته شده است [۲۲]. از آنجایی که رابطه‌ی مستقیمی میان آلودگی ناشی از ترافیک و فاصله از جاده وجود دارد و بدلیل در دسترس نبودن داده‌های ترافیکی شهر تهران، در این مطالعه اثر ترافیک به صورت تابعی از فاصله از جاده در نظر گرفته شده است. برای این منظور با استفاده از تحلیل مکانی KDE[13] در نرم افزار ArcGIS لایه‌ی رستری ایجاد شده که مقدار هر پیکسل بیانگر فاصله از جاده بوده و به عنوان اثر ترافیک به مدل ارائه می‌شود. در واقع در این تحلیل فرض شده است که هر چه فاصله‌ی یک نقطه از جاده بیشتر باشد، اثر ترافیک در آن نقطه کمتر خواهد بود. شکل ۵ رستر تولید شده را نمایش می‌دهد. همانطور که در شکل نیز نمایان است، در مناطق مرکزی، ترافیک بیشترین تاثیر را بر میزان آلودگی هوای شهر دارد.

ارتفاع متوسط منطقه: شهر تهران در ارتفاع متوسط ۱۱۹۰ متر بالاتر از سطح آبهای آزاد جهان قرار دارد. به علت گستردگی و وسعت زیاد این شهر، اختلاف ارتفاع میان پایین‌ترین و بالاترین نقاط شهر به ۷۰۰ متر می‌رسد. این تغییرات ارتفاع که موجب تغییرات آب و هوایی در مناطق مختلف شهر می‌شود، تاثیر بسزایی در آلودگی هوای شهر دارد [۳].

[13] Kernel Density Estimation



شکل ۵: اثر ترافیک در آلودگی هوا

همچنین تهران از جمله شهرهای ساخته شده بر روی مناطق تپه‌ای می‌باشد. بنابراین در سطح شهر نقاطی وجود دارد که ارتفاع آنها نسبت به مناطق اطراف بالاتر است. این در حالی است که ارتفاع مطلق آنها ممکن است پایین‌تر از سایر مناطق شهر باشد. بر روی مناطق تپه‌ای، اگرچه ارتفاع مطلق نقطه پایین‌تر از مناطق شمالی شهر است، ارتفاع محلی آن بالاتر از نقاط همسایه می‌باشد. در این مطالعه، برای مدل کردن اثر محلی ارتفاع در آلودگی هوا، از ارتفاع متوسط محلی استفاده شده است. برای این منظور، ارتفاع متوسط محلی هر نقطه در ناحیه‌ای دایره‌ای به شعاع ۲۵۰۰ متر، به کمک نرم افزار ArcGIS محاسبه شده است. سپس اختلاف این مقدار از ارتفاع مطلق نقطه بدست آمده و خروجی به عنوان متغیر تاثیرگذار در آلودگی هوا به مدل ارائه شده است.

شکل زمین: ویژگی‌های سطح زمین می‌تواند بر میزان غلظت آلاینده‌ها در مناطق مختلف تاثیرگذار باشد. از جمله این ویژگی‌ها می‌توان به تحدب و تقعر سطح زمین اشاره کرد. در مناطقی که سطح زمین حالت مقعر دارد، آلودگی هوا می‌تواند برای مدت طولانی‌تری باقی بماند و به همین ترتیب، در مناطق محدب، وزش باد آلودگی هوا را نسبت به سایر مناطق، سریعتر دور می‌کند. بنابراین، در این مطالعه، تحدب و تقعر سطح زمین به عنوان دیگر پارامتر تاثیرگذار در آلودگی هوا در نظر گرفته شده است. برای محاسبه این پارامتر از DEM Surface Tools 10[1] به‌عنوان یک افزونه برای نرم‌افزار ArcGIS استفاده شده است. میزان تقعر و تحدب با استفاده از معادله (۲) بدست آمده است:

رابطه (۲)[۲۳]

$$General\ Curvature = -2(r + t)$$

که در معادله (۲)، t و r مشتقات دوم نسبت به ارتفاع هستند:

رابطه (۳) $$r = \delta^2 z/\delta x^2$$

رابطه (۴) $$t = \delta^2 z/\delta y^2$$

۲-۲-۳- داده‌های هواشناسی

شهر تهران دارای ۵ ایستگاه هواشناسی است که داده‌های سری زمانی مربوط به وضعیت آب و هوا را در بازه‌های ۳ ساعته جمع آوری می‌کنند. این داده‌ها توسط اداره کل هواشناسی استان تهران جمع‌آوری شده و در دسترس قرار می‌گیرند. در این مطالعه، از میان داده‌های جمع آوری شده، دما، فشار، سرعت وزش باد، میزان پوشش ابر و رطوبت به‌عنوان پارامترهای تاثیرگذار در آلودگی هوا مورداستفاده قرار گرفته‌اند.

[14] http://www.jennessent.com/arcgis/surface_area.htm

۳- مبانی نظری تحقیق

در این بخش مبانی نظری ماشین بردار پشتیبان و همچنین ماشین بردار پشتیبان برخط به طور مختصر تشریح شده است.

۳-۱- ماشین بردار پشتیبان

سری‌های زمانی روشی مناسب برای تحلیل و پیش‌بینی رفتار متغیرها در طول زمان می‌باشند. کارایی روش ماشین بردار پشتیبان در پیش‌بینی سری‌های زمانی در تحقیقات مختلفی به اثبات رسیده است [۲۴, ۲۵]. ماشین بردار پشتیبان که بر اساس مبانی آماری بنا شده است، اولین بار توسط وپنیک[۱] [۲۶] مطرح گردید. اگرچه ماشین بردار پشتیبان[۲] یک طبقه‌بندی کننده ی[۳] دودویی است که برای کلاسه‌بندی[۴] و تشخیص الگوها[۵] طراحی شده است، می‌تواند برای پیش‌بینی سری‌های زمانی نیز مورد استفاده قرار گیرد [۱۹].

مجموعه‌ای از داده‌های آموزشی $\{x_1, x_2, \dots, x_i\}$ را به همراه کلاس مربوط به آنها $\{y_1, y_2 \dots y_i\}$ در نظر می‌گیریم، به طوریکه

$y_i \in \{-1, +1\}$ در صورتی‌که دو کلاس بصورت خطی قابل جداسازی باشند، ماشین بردار پشتیبان ابرصفحه‌ای[۶] را ایجاد می‌کند که دسته‌ها را به گونه‌ای از هم جدا کند تا فاصله‌ی میان نزدیکترین نمونه‌های دو کلاس، در راستای عمود بر مرز تصمیم‌گیری، بیشینه شود. در حالت خطی، طبقه‌بندی بهینه با استفاده از رابطه‌ی (۵) حاصل می‌شود:

$$f(x) = sign\{W.x_i + b\}$$ رابطه (۵)

به نحوی که شروط (۶) و (۷) برقرار باشند.

$$\sum_{i=1}^{l} \alpha_i y_i = 0$$ رابطه (۶)

$$\alpha_i \geq 0 \quad \forall_i$$ رابطه (۷)

در رابطه‌های (۵)، (۶) و (۷) α_i ضرایب لاگرانژ، W بردار نرمال عمود بر ابرصفحه و b فاصله از مبدا مختصات می‌باشد. W و b به ترتیب با استفاده از معادلات (۸) و (۹) محاسبه می‌گردند:

$$W = \sum_{i=1}^{l} \alpha_i y_i x_i$$ رابطه (۸)

$$b = y_i - \sum_{i=1}^{l} y_i \alpha_i x_i x_j$$ رابطه (۹)

تنها نمونه‌هایی که ضرایب لاگرانژ آنها مخالف صفر است ($\alpha_i \neq 0$) در تشکیل ابرصفحه مشارکت دارند. این داده‌ها، که نزدیکترین نمونه‌ها به ابرصفحه هستند، بردارهای پشتیبان[۷] نامیده می‌شوند. سایر نمونه‌ها با $\alpha_i = 0$، تاثیری در شکل‌گیری ابرصفحه نخواهند داشت [۲۷].

معادلات (۵)، (۸) و (۹) برای طبقه‌بندی در حالت خطی قابل استفاده است، در حالی‌که بسیاری از پدیده‌های طبیعی قابل مدلسازی در فضاهای خطی نمی‌باشند. در چنین شرایطی از تابع کرنل[۸] برای نگاشت داده‌ها به فضایی با ابعاد بیشتر استفاده می‌شود، بطوریکه در فضای جدید، داده‌ها بصورت خطی قابل جداسازی باشند. برای حالت غیرخطی رابطه‌ی (۵) به رابطه‌ی (۱۰) تبدیل شده و W و b مشابه با حالت خطی بر اساس رابطه‌های (۸) و (۹) محاسبه می‌گردند:

رابطه (۱۰)

$$f(x) = sign\{\sum_{i=1}^{l} \alpha_i y_i k(x_i, x_j) + b\}$$

به نحوی که شروط (۱۱) و (۱۲) برقرار باشند.

$$\sum_{i=1}^{l} \alpha_i y_i = 0$$ رابطه (۱۱)

$$0 \leq \alpha_i \leq C \forall_i$$ رابطه (۱۲)

در رابطه‌ی (۱۰) $k(x_i, x_j)$ تابع کرنل و C عددی ثابت است که توسط کاربر تعیین می‌شود. پارامتر C

---

[15] Vapnik

[۲] تمامی معادلات ریاضی این قسمت از[۲۷] برداشت شده است.

[17] Classifier
[18] Classification
[19] Pattern Recognition
[20] Hyperplane

[21] Support Vectors
[22] Kernel Function

مشخص می‌کند چه میزان خطا در طبقه‌بندی قابل صرف نظر کردن است.

۳-۲- ماشین بردار پشتیبان برخط

الگوریتم ماشین بردار پشتیبان، که در بخش قبل بطور مختصر تشریح شد، به‌عنوان الگوریتمی مناسب برای حل مسائل غیرخطی شناخته شده است. اما زمانی که با حجم بالای داده‌های جریانی روبرو هستیم، کارایی الگوریتم کاهش می‌یابد. عملکرد الگوریتم به این صورت است که ابتدا تمام نمونه‌ها جهت آموزش وارد مدل شده و مدل آموزش می‌بینند. اما پس از پایان مرحله‌ی آموزش، در صورت افزودن داده‌های آموزشی جدید، نیاز است آموزش مجددا با استفاده از تمامی داده‌های موجود انجام شود [۲۸, ۲۹]. نیاز به آموزش مجدد با استفاده از تمامی داده‌ها، امکان استفاده از الگوریتم ماشین بردار پشتیبان را در حل مسائلی مانند پیش‌بینی پویای آلودگی هوا که در آن داده‌های ایستگاه‌های پایش آلودگی هوا و ایستگاه‌های هواشناسی به‌صورت لحظه‌ای جمع‌آوری می‌شوند و حجم این داده‌ها حتی در طی چند روز هم بالا خواهد بود، غیرممکن به نظر می‌رسد. این محدودیت‌های الگوریتم ماشین بردار پشتیبان محققان را به سمت ارائه‌ی روش‌های پویا برای حل مشکلات ماشین بردار پشتیبان جهت پردازش داده‌های جریانی و حجیم سوق داده است. از جمله‌ی این روش‌ها می‌توان به الگوریتم برخط LaSVM اشاره کرد که اولین بار توسط بوردز[۱] ارائه گردید [۳۰]. اساس کار این الگوریتم به این صورت است که با ورود داده‌ی آموزشی جدید، ضرایب لاگرانژ (α_i) در دو مرحله با نام‌های "Process"و "Reprocess" به‌روز می‌شوند. با ورود نمونه‌ی آموزشی جدید، مرحله Process آغاز می‌شود. در این مرحله بررسی می‌شود که آیا نمونه‌ی جدید می‌تواند یک بردار پشتیبان باشد یا خیر. در صورتی که نمونه جدید یک بردار پشتیبان باشد، به مجموعه بردارهای پشتیبان اضافه می‌شود. مرحله‌ی Reprocess عملکردی معکوس دارد. در این مرحله، بردارهای پشتیبانی که ضرایب آنها می‌تواند بعد از به‌روز شدن ضرایب صفر شوند، از مجموعه‌ی بردارهای پشتیبان حذف می‌شوند [۳۱]. پس از انجام این دو مرحله، تمامی نمونه‌هایی که بردار پشتیبان نبوده و به عبارت دیگر دارای ضرایب صفر هستند، حذف می گردند. با اضافه و حذف شدن بردارهای پشتیبان، ابرصفحه‌ی جداکننده به طور پیوسته به‌روز می‌شود. در مرحله‌ی بعد داده‌ی آموزشی جدید به الگوریتم ارائه می‌شود و آموزش با استفاده از این داده‌ی جدید و بردارهای پشتیبان استخراج شده از مراحل قبل انجام می‌شود. بنابراین الگوریتم بسیار سریعتر از حالتی که از تمامی داده‌ها برای آموزش استفاده می‌شود، آموزش می‌بینند. این خصوصیت LaSVM آن را به الگوریتمی مناسب برای پردازش داده‌های جریانی حجیم تبدیل می‌کند. در این پژوهش نیز الگوریتم LaSVM برای پیش‌بینی پویای آلودگی هوا مورد استفاده قرار گرفته است.

۴- پیاده‌سازی الگوریتم پویای پیش‌بینی آلودگی هوا و ارزیابی نتایج

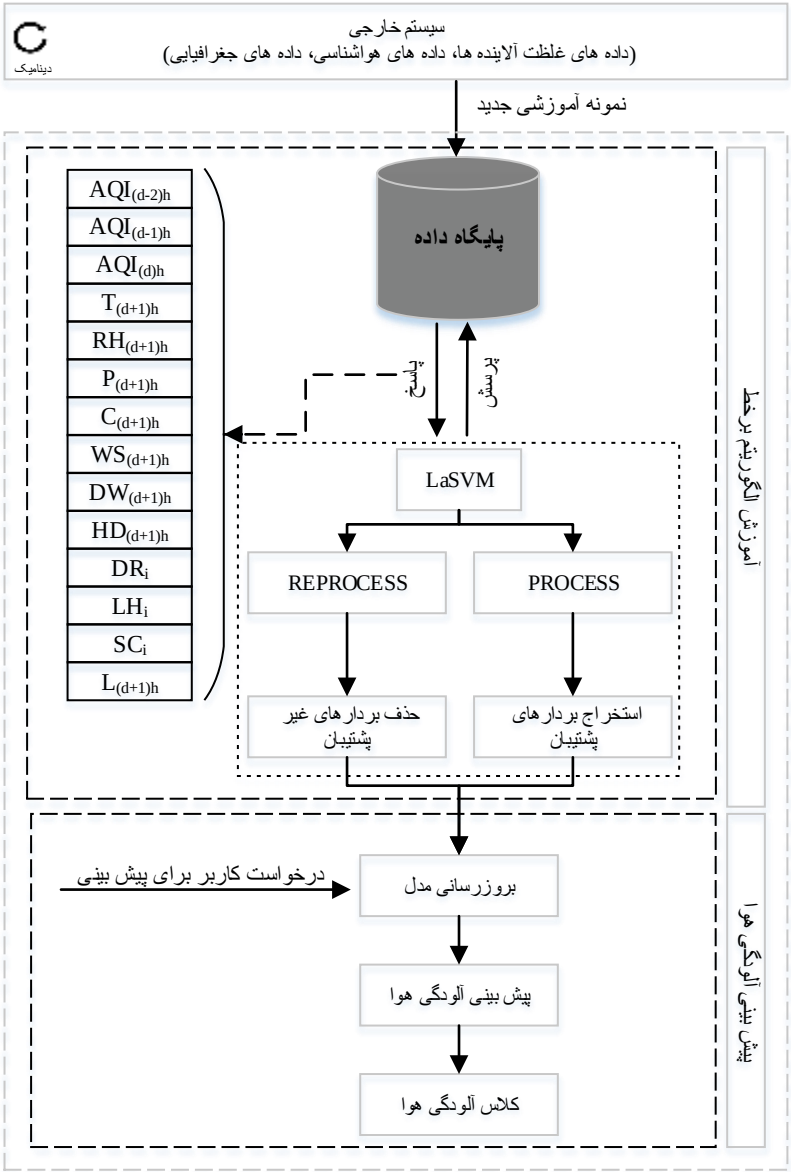
در این بخش نحوه‌ی پیاده‌سازی ماشین بردار پشتیبان به‌صورت پویا برای پیش‌بینی آلودگی هوای شهر تهران، مدل طراحی شده و نتایج حاصله تشریح می‌شوند.

۴-۱- مدل توسعه داده شده

جهت انجام پردازش‌های پویا از الگوریتم برخط LaSVM استفاده شده است. سری زمانی مربوط به داده‌های آلودگی هوا و هواشناسی به طور پیوسته توسط ایستگاه‌های سنجش برداشت شده و در پایگاه داده ذخیره می‌شود. در این مطالعه، PostgreSQL به عنوان پایگاه داده مناسب جهت ذخیره‌سازی داده‌های غلظت آلاینده‌ها،

[23] Bordes

داده‌های هواشناسی و داده‌های مکانی مورد استفاده قرار گرفته است. برای توسعه‌ی الگوریتم برخط از زبان برنامه‌نویسی جاوا (نسخه ۱.۸.۰) استفاده شده است. همچنین، برنامه در یک رایانه‌ی شخصی با مشخصات (Intel 1.8 GHz machine, 5 cores and 6 GB of RAM)، پیاده‌سازی و تست شده است. شمای کلی سیستم و نحوه عملکرد این سیستم در شکل ۶ نمایش داده شده است. در عمل، برنامه داده‌ها را از پایگاه داده بازیابی می‌کند، الگوریتم را آموزش می‌دهد و از الگوریتم آموزش داده شده برای پیش‌بینی سری زمانی آلودگی هوا در هر نقطه‌ی مورد نظراستفاده می‌کند.



شکل ۶: شمای کلی از سیستم پیش‌بینی آلودگی هوای شهری

برای پیش‌بینی سری زمانی برای چند ساعت آینده لازم است داده‌ها با ساختار مشخصی از پایگاه داده بازیابی و به عنوان ورودی برای آموزش الگوریتم مورد استفاده قرار بگیرند. رابطه (۱۳) ساختار سری زمانی استفاده شده برای ورود داده به الگوریتم را نمایش می‌دهد. در این ساختار اندیس‌های d و h به ترتیب روز و ساعت را مشخص می‌کنند. سه پارامتر ابتدائی، شاخص آلودگی هوا در سه روز قبل و در ساعت مورد نظر می‌باشند. T_{d+1}^h، RH_{d+1}^h، P_{d+1}^hو WS_{d+1}^h به‌ترتیب دما، فشار، رطوبت نسبی و سرعت باد مربوط به زمان پیش‌بینی و DW_{d+1}^h و HD_{d+1}^h روز و رابطه (۱۳)

$$L_{d+1}^h = f(AQI_d^h, AQI_{d-1}^h, AQI_{d-2}^h, T_{d+1}^h, RH_{d+1}^h, P_{d+1}^h, C_{d+1}^h, WS_{d+1}^h, DW_{d+1}^h, HD_{d+1}^h, DR_i, LH_i, SC_i)$$

ساعت پیش‌بینی می‌باشند. سه پارامتر انتهایی، دادههای مکانی استفاده شده در این مطالعه هستند که DR_i فاصله از جاده، LH_i ارتفاع متوسط منطقه و SC_i تحدب و تقعر سطح زمین را نمایش می‌دهند. i مختصات نقطه‌ی مورد نظر برای پیش‌بینی می‌باشد. نحوه‌ی آماده‌سازی این پارامترها در بخش ۲-۲ توضیح داده شده است. L_{d+1}^h نیز مشخص کننده‌ی کلاس آلودگی هوای مربوط به این سطر از داده می‌باشد. در واقع $d+1$ بیانگر زمان پیش‌بینی یعنی ۲۴ ساعت آینده می‌باشد.

هدف اصلی سری زمانی یافتن تابع $f(x)$ مناسب است که بتواند به درستی ارتباط میان مقادیر قابل پیش‌بینی برای آینده (L_{d+1}^h) و مقادیر پارامترها در گذشته را برقرار کند. الگوریتم برخط LaSVM استفاده شده در این مطالعه، این تابع را استخراج می‌کند. برای آموزش، داده‌های ۳ سال گذشته به صورت پویا و پیوسته به الگوریتم LaSVM وارد می‌شوند. با اضافه شدن هر داده‌ی آموزشی جدید، مجموعه‌ی بردارهای پشتیبان و درنتیجه صفحات جداکننده، به‌روز می‌شوند. علاوه بر داده‌های ۳ سال گذشته، هر داده‌ی جدیدی که توسط ایستگاه‌های سنجش آلودگی و هواشناسی بصورت آنی برداشت می‌شود نیز جهت آموزش الگوریتم مورد استفاده قرار می‌گیرد. خروجی آموزش، تابع $f(x)$ است که ارتباط میان ورودی‌های سری زمانی و مقدار پیش‌بینی را برقرار می‌کند. برای آموزش صحیح الگوریتم و بدست آوردن نتایج دقیق نیاز است پارامترهای الگوریتم، نوع تابع کرنل و پارامترهای مربوط به آن به درستی تعیین شوند. از آنجایی که عملکرد بهتر تابع گوسین نسبت به سایر توابع کرنل در بسیاری از مطالعات پیشین به اثبات رسیده است [۱۵, ۱۸, ۳۲]، در این مطالعه نیز تابع گوسین به عنوان تابع کرنل مناسب انتخاب شده است و مقادیر پارامترهای مورد نیاز به صورت تجربی انتخاب شده‌اند. مقدار پارامتر C معادل ۲ و مقدار پارامتر گاما معادل ۰.۰۰۱۹ در نظر گرفته شده است.

همانطور که در بخش ۳ نیز مطرح شد، ماشین بردار پشتیبان و در نتیجه الگوریتم LaSVM طبقه‌بندی کننده‌ی دودویی هستند. اما در بیشتر مسائل دنیای واقعی از جمله مسئله‌ی آلودگی هوا، نیاز به طبقه‌بندی چندگانه[۱] است. برای حل یک مسئله با طبقه‌بندی چندگانه با استفاده از طبقه‌بندی کننده‌های دودویی، می‌توان از ترکیب چند طبقه‌بندی کننده‌ی دودویی استفاده کرد [۳۳]. روش‌های یکی در برابر همه[۲] و یکی در برابر یکی[۳] از جمله روش‌هایی هستند که می‌توانند

[24] Multiple Classification
[25] One-against-all
[26] One-against-one

یک مسئله‌ی طبقه‌بندی چندگانه را به چندین مسئله‌ی طبقه‌بندی دوگانه بشکنند. در این مطالعه، برای ایجاد طبقه‌بندی کننده‌ی چندگانه با استفاده از طبقه‌بندی کننده‌ی دوگانه‌ی LaSVM، روش یکی در برابر همه انتخاب شده است. برای مسئله‌ی آلودگی هوا با ۷ کلاس شاخص آلودگی، ۷ LaSVM دودویی ایجاد شده است و با استفاده از مکانیزم یکی در برابر همه کلاس مناسب آلودگی هوا انتخاب می‌شود. توضیحات بیشتر در زمینه‌ی طبقه‌بندی چندگانه در [۲۶] در دسترس می‌باشد.

۵- نتایج و تحلیل

برای ارزیابی الگوریتم برخط استفاده شده در پیش‌بینی سری زمانی آلودگی هوای شهر تهران، نتایج حاصله با روش ماشین بردار پشتیبان معمولی مقایسه شده است. برای داشتن مقایسه‌ای صحیح، داده‌های آموزشی مشابهی به هر دو الگوریتم وارد شده‌اند. اما از آنجایی که نحوه‌ی عمکرد آنها با یکدیگر متفاوت است، داده‌ها با روش‌های متفاوتی به دو الگوریتم وارد می‌شوند. برای الگوریتم برخط LaSVM، داده‌ها بصورت پیوسته و جریانی به الگوریتم معرفی می‌شوند. اما از آنجایی که ماشین بردار پشتیبان قابلیت کار با داده‌های جریانی را ندارد، داده‌ها به دسته‌هایی با اندازه‌ی مساوی تقسیم شده و در هر مرحله از آموزش، یکی از دسته‌ها به داده‌های موجود اضافه شده و آموزش از سر گرفته می‌شود. نتایج حاصل از آموزش هر الگوریتم، در پایگاه داده ذخیره شده و با یکدیگر مقایسه شده‌اند.

از آنجایی که هدف اصلی این مطالعه پیش‌بینی آلودگی هوا به‌صورت پویا و برخط است، زمان پردازش یکی از پارامترهای اصلی برای ارزیابی کارایی الگوریتم ارائه شده است. علاوه بر پارامتر زمان، پارامترهای دقت[۱]، خطای مجذور میانگین مربعات[۲] و ضریب تعیین[۳] نیز برای ارزیابی دقت نتایج استفاده شده‌اند. روابط (۱۴)، (۱۵) و (۱۶) به‌ترتیب نحوه‌ی محاسبه‌ی پارامترهای دقت، خطای مجذور میانگین مربعات و ضریب تعیین را بیان می‌کنند:

رابطه (۱۴)

$$\text{دقت} = \frac{\text{تعداد نمونه‌های تست که به درستی پیش‌بینی شده‌اند}}{\text{تعداد کل نمونه‌های تست}}$$

رابطه (۱۵)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - Y_i^*\right|^2}$$

رابطه (۱۶)

$$R^2 = 1 - \sum_{i=1}^{n}(Y_i - Y_i^*)^2 / \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

در معادلات (۱۵) و (۱۶)، Y_i^*، مقدار پیش‌بینی شده و Y_i، مقدار مشاهداتی و در معادله‌ی (۱۶) $\bar{Y}$، میانگین مقادیر مشاهداتی می‌باشد. در ادامه مقایسه‌ی دو الگوریتم بر اساس زمان پردازش و دقت حاصله، تشریح شده است.

۵-۱- زمان پردازش

شکل ۷ نشان‌دهنده‌ی ارتباط میان زمان پردازش و تعداد داده‌های آموزشی برای هر دو الگوریتم ماشین بردار پشتیبان پویا و ماشین بردار پشتیبان معمولی می‌باشد. در ابتدای پردازش، هر دو الگوریتم زمان تقریبا یکسانی را برای آموزش صرف می‌کنند. اما با گذشت زمان و اضافه شدن داده‌های آموزشی، زمان آموزش ماشین بردار پشتیبان معمولی بصورت نمایی افزایش پیدا می‌کند. با اضافه کردن هزاران داده‌ی آموزشی، ساعت‌ها زمان نیاز است که الگوریتم با اضافه شدن یک داده‌ی جدید، آموزش ببیند. به طور خاص در این مطالعه، زمانی که تعداد داده‌ها به ۱۵۰۰۰ می‌رسد، الگوریتم برای آموزش، به بیش از ۱۶ ساعت زمان نیاز دارد. روند افزایش نمایی زمان آموزش برای ماشین بردار پشتیبان

[27] Accuracy

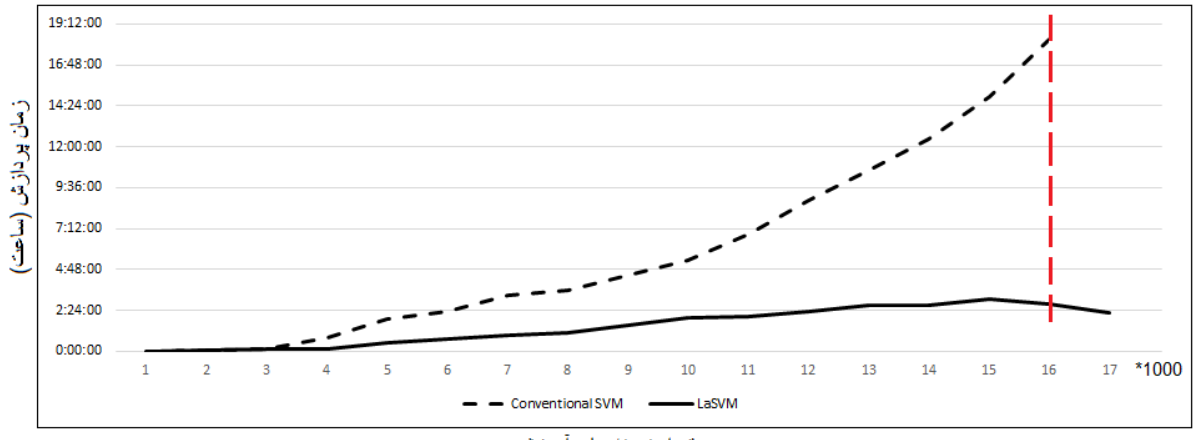
[28] Root Mean Square Error (RMSE)

[29] RSquared

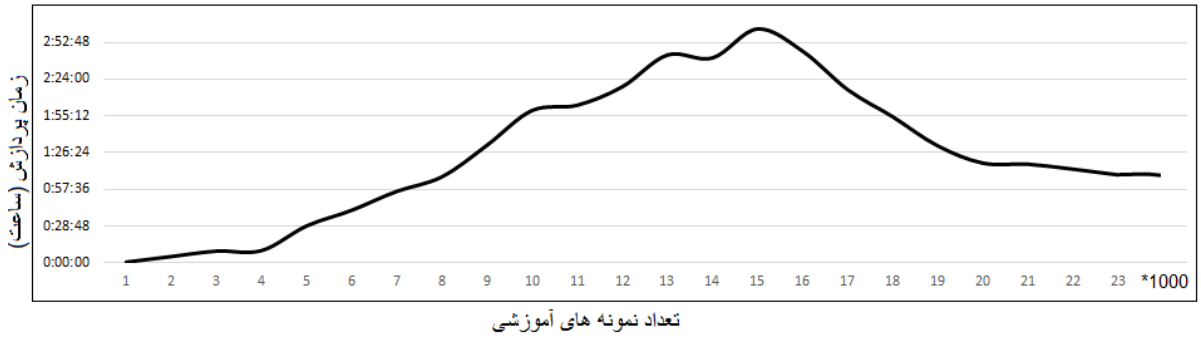
به این دلیل است که با ورود هر نمونه‌ی آموزشی جدید، الگوریتم نیاز دارد آموزش را مجددا با تمام داده‌های موجود انجام دهد. واضح است که این نحوه‌ی عملکرد ماشین بردار پشتیبان، استفاده از آن را برای پردازش‌های برخط غیرممکن می‌کند. اگرچه زمان آموزش الگوریتم برخط ارائه شده نیز در ابتدای آموزش افزایش پیدا می‌کند، اما این روند افزایش، در مقایسه با افزایش زمان آموزش ماشین بردار پشتیبان معمولی، بسیار کمتر است. سرعت کمتر در افزایش زمان پردازش الگوریتم برخط به این دلیل است که با اضافه شدن داده‌ی آموزشی جدید، الگوریتم نیازی ندارد برای آموزش مجدد از تمامی داده‌های موجود استفاده کند. بلکه تنها از بردارهای پشتیبان استخراج شده از مراحل قبل و نمونه‌ی آموزشی جدید استفاده می‌کند.

شکل ۸ روند افزایش زمان پردازش الگوریتم برخط را با جزئیات بیشتری نمایش می‌دهد. اگرچه زمان در ابتدای آموزش افزایش پیدا می‌کند، اما با افزودن تعداد نمونه‌های آموزشی بیشتر و یافتن تعداد بردارهای پشتیبان مناسب و در نتیجه ایجاد ابرصفحه‌ی جدا کننده‌ی بهینه، زمان پردازش به‌طور قابل توجهی کاهش پیدا می‌کند. شکل ۹ که نشان‌دهنده‌ی تعداد بردارهای پشتیان می‌باشد، موید این مطلب است. در حقیقت، کاهش زمان آموزش الگوریتم برخط، همزمان با افزایش تعداد بردارهای پشتیبان می‌باشد. با ثابت شدن تعداد بردارهای پشتیبان نیز، زمان پردازش‌ها تقریبا ثابت باقی ماند. از این لحظه به بعد، الگوریتم رفتار متعادلی را نشان می‌دهد. ثابت ماندن تعداد بردارهای پشتیبان می‌تواند به دو دلیل اتفاق افتد: تعداد بردارهای پشتیبان یافت شده در مرحله‌ی Process برابر با تعداد بردارهای پشتیبان حذف شده در مرحله‌ی Reprocess می‌باشد و یا ابرصفحه با دقت بالایی ایجاد شده است که ورود نمونه‌ی آموزشی جدید باعث نقض شروط مربوط به ایجاد ابرصفحه نمی‌شود. بدین معنا که افزودن نمونه‌ی آموزشی جدید، منجر به ایجاد بردار پشتیبان جدید در مرحله‌ی Process نمی‌شود.

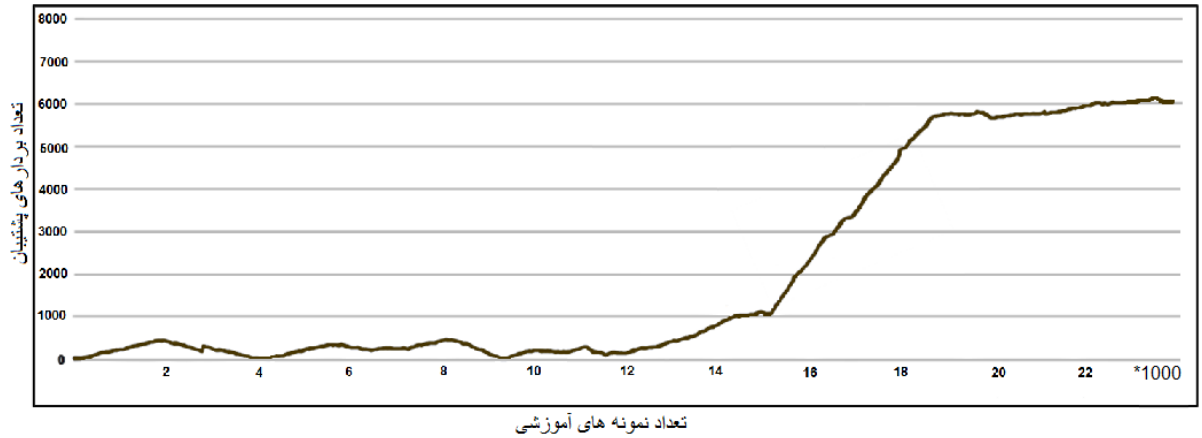
شکل ۱۰ بیانگر رابطه‌ی میان تعداد کل داده‌های ورودی به مدل و تعداد داده‌هایی است که توسط الگوریتم حفظ شده‌اند. همانطور که در این شکل نیز نمایش داده شده است، تعداد داده‌هایی که در مدل باقی می‌مانند، با ثابت شدن تعداد بردارهای پشتیبان، تقریبا ثابت باقی می‌ماند. حذف داده‌های اضافی که نقشی در آموزش مدل ندارند، علاوه بر افزایش سرعت پردازش‌ها، حجم حافظه‌ی مورد نیاز برای نگهداری داده‌ها را نیز کاهش می‌دهد.



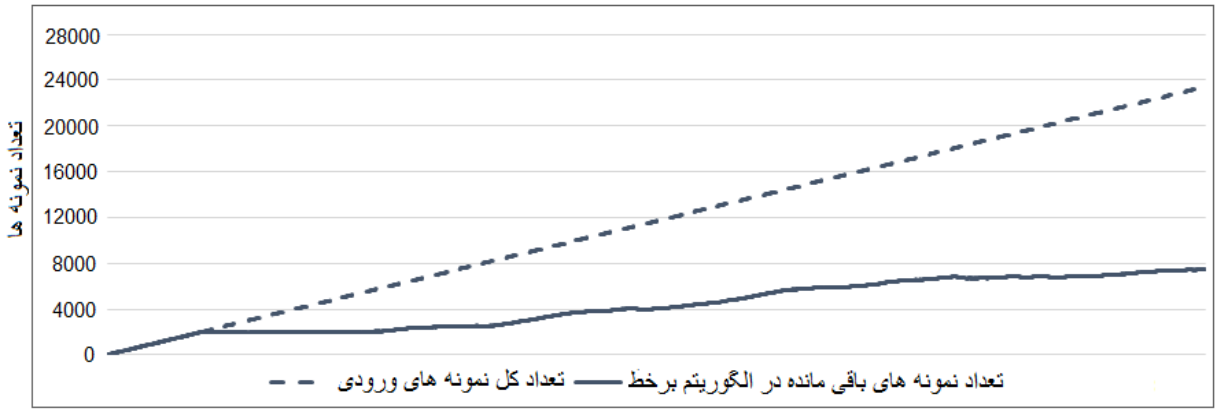
شکل ۷: مقایسه زمان آموزش برای دو الگوریتم ماشین بردار پشتیبان پویا و ماشین بردار پشتیبان معمولی



شکل ۸: تغییرات زمان آموزش الگوریتم برخط با توجه به تعداد داده‌ی ورودی



شکل ۹: تعداد بردارهای پشتیبان انتخاب شده توسط الگوریتم

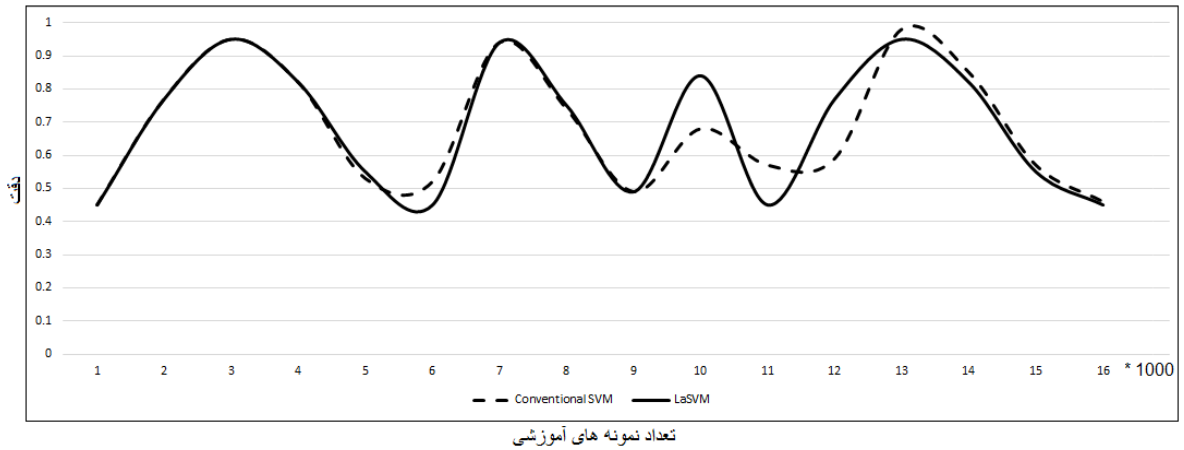


شکل ۱۰: تعداد نمونه‌های باقی مانده در مقایسه با تعداد کل داده‌های ورودی

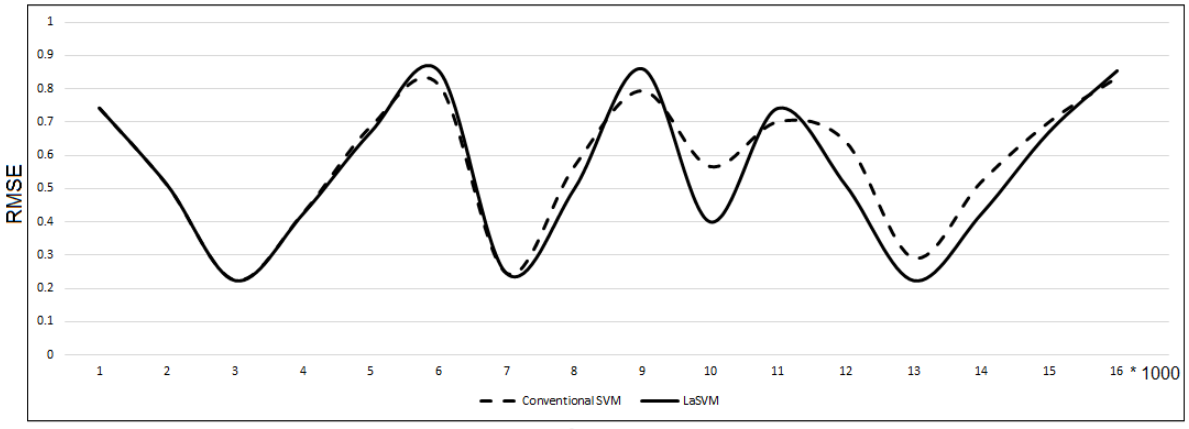
۵-۲- دقت

اشکال ۱۱، ۱۲ و ۱۳ به ترتیب اختلاف میان دقت، خطای مجذور میانگین مربعات و ضریب تعیین مربوط به هر دو الگوریتم برخط و ماشین بردار پشتیبان معمولی را تا زمانی که ماشین بردار پشتیبان به علت حجم بالای داده متوقف می‌شود، نمایش می‌دهند. در گام‌های اولیه‌ی آموزش، دقت حاصل شده از هر دو الگوریتم تقریبا برابر است. گذشت زمان و حذف نمونه‌های آموزشی بیشتر توسط الگوریتم برخط موجب ایجاد اختلاف در دقت‌های حاصل شده در گام‌های میانی آموزش می‌شود. اما دقت حاصل از الگوریتم برخط، پس از یافتن تعداد کافی

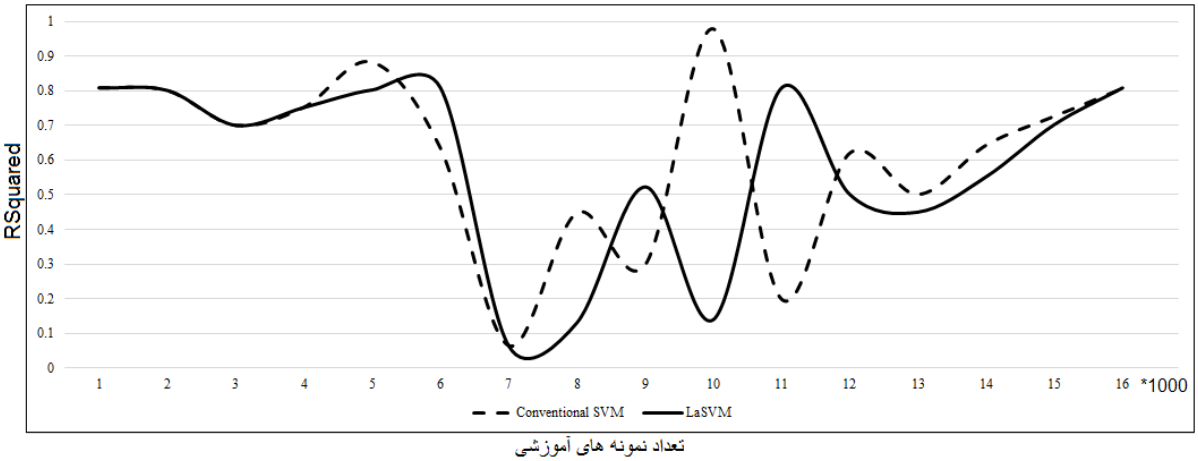
بردارهای پشتیبان برای ایجاد ابرصفحه‌ی جداکننده‌ی مناسب، به دقت ماشین بردار پشتیبان معمولی همگرا می‌شود. روندی مشابه برای تغییرات خطای مجذور میانگین مربعات و ضریب تعیین مشاهده می‌شود (اشکال ۱۲ و ۱۳) میانگین اختلاف میان دقت، خطای مجذور میانگین مربعات و ضریب تعیین مربوط به هر دو الگوریتم برخط و ماشین بردار پشتیبان معمولی تا این مرحله از آموزش به ترتیب ۰.۰۴۱، ۰.۰۴۶ و ۰.۱۷ می‌باشند. افزون بر مقایسه‌ی دو الگوریتم در این پژوهش دقت، خطای مجذور میانگین مربعات و ضریب تعیین برای الگوریتم برخط توسعه داده شده با استفاده از داده‌های یکسال محاسبه شده است که دقت ۰.۷۱، خطای مجذور میانگین مربعات ۰.۵۴ و ضریب تعیین ۰.۸۱ را حاصل کرده است.



شکل ۱۱: مقایسه‌ی دقت حاصل شده از الگوریتم برخط و دقت حاصله از ماشین بردار پشتیبان معمولی



شکل ۱۲: مقایسه‌ی خطای مجذور میانگین مربعات حاصل شده از الگوریتم برخط و خطای مجذور میانگین مربعات حاصله از ماشین بردار پشتیبان معمولی



شکل ۱۳: مقایسه‌ی ضریب تعیین حاصل شده از الگوریتم برخط و ضریب تعیین حاصله از ماشین بردار پشتیبان معمولی

علاوه‌بر مقایسه‌ی دو الگوریتم ماشین بردار پشتیبان برخط و ماشین بردار پشتیبان معمولی بصورت نمودارهای فوق، مقایسه‌ی عددی مربوط به زمان آموزش، دقت، RMSE و R Squared براساس تعداد نمونه‌های آموزشی در جدول ۲ آمده است.

۵-۳- پیش‌بینی آلودگی هوا

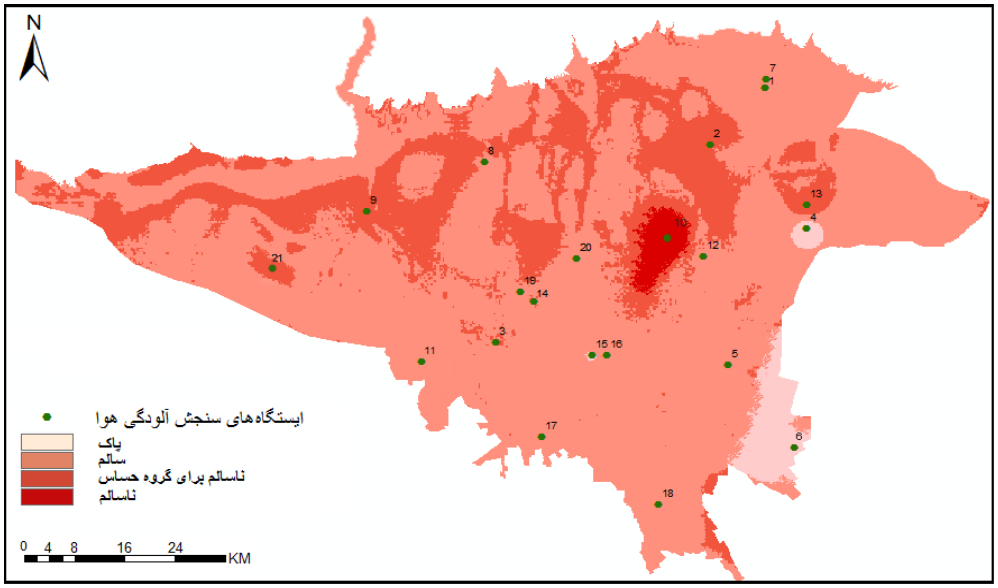
پس از ارزیابی الگوریتم برخط ارائه شده و طراحی سیستمی جهت پیش‌بینی پویای آلودگی هوا، می‌توان از این سیستم برای پیش‌بینی آلودگی هوای شهر تهران بر روی هر نقطه‌ی دلخواه استفاده نمود. به‌منظور نمایش کارآیی سیستم، پیش‌بینی وضعیت هوا در یک روز آلوده و مقایسه‌ی آن با وضعیت مشاهده شده توسط ایستگاه‌ها در آن روز در دستور کار قرار گرفت. نتایج پیش‌بینی انجام شده برای تاریخ ۱۳۹۱/۰۸/۱۸ ساعت ۹ صبح در شکل ۱۴ ارائه شده است. سیستم طراحی شده توانست آلودگی هوا را با دقت مناسبی برای تاریخ مذکور پیش‌بینی کند. پیش‌بینی به این صورت انجام شده است که الگوریتم بر روی هر پیکسل حرکت کرده و بر اساس پارامترهای مکانی، داده‌های هواشناسی و شاخص آلودگی هوا در سه روز گذشته، کلاس آلودگی هوا مربوط به هر پیکسل را پیش‌بینی می‌کند. همانطور که در این شکل مشخص است، آلودگی هوا در ایستگاه شماره ۱۰ بیشترین مقدار را داشته است. شکل ۱۵- الف نمای بزرگتری از این منطقه را نمایش می‌دهد. اشکال ۱۵- ب و ۱۵- ج نیز به ترتیب اثر ترافیک و ارتفاع متوسط منطقه را نمایش می‌دهند. همانطور که از شکل ۱۵- ب نیز مشخص است، بدلیل نزدیکی این ایستگاه به مرکز شهر، ترافیک موجب افزایش آلودگی هوا در این ناحیه شده است. همچنین، براساس شکل ۱۵- ج، ارتفاع متوسط این منطقه نسبت به مناطق همسایه کمتر بوده که موجب می‌شود آلودگی برای مدت بیشتری در این منطقه باقی مانده و شدت یابد. مقایسه‌ی کلاس آلودگی پیش‌بینی شده با مقادیرمشاهده شده بر روی ایستگاه‌های سنجش آلودگی هوا در شکل ۱۶ نمایش داده شده است. همانطور که در شکل نیز مشخص است، پیش‌بینی در ۱۸ ایستگاه بدرستی انجام شده و در ۳ ایستگاه، با یک کلاس اختلاف، پیش‌بینی درستی صورت نگرفته است.

یکی از مشکلات موجود در پیش بینی آلودگی هوا با استفاده از سیستم توسعه داده شده، وابستگی آن به

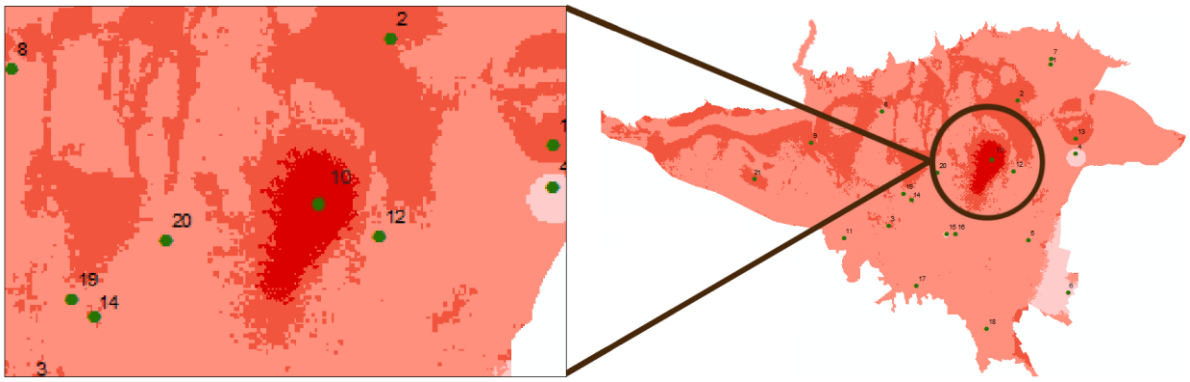
پارامترهای ورودی است. بطور مثال همانطور که در شکل ۱۴ نیز مشخص است، کلاس آلودگی هوا در ایستگاه شماره ۴ پاک پیش‌بینی شده است. این درحالی است که در ایستگاه شماره ۱۳ که در فاصله‌ی کمی از ایستگاه شماره ۴ قرار دارد، کلاس آلودگی هوا، ناسالم برای گروه‌های حساس پیش‌بینی شده است. این اختلاف به این دلیل است که در روزهای قبل از تاریخ تعیین شده برای پیش‌بینی آلودگی هوا، برخی از شاخص‌های آلودگی هوا توسط ایستگاه شماره ۴ اندازه‌گیری نشده‌اند. این مشکل موجب گردیده پیش‌بینی در این ایستگاه بدرستی انجام نگیرد.

جدول ۲: مقایسه‌ی عددی نتایج حاصل از ارزیابی دو الگوریتم ماشین بردار پشتیبان برخط و ماشین بردار پشتیبان معمولی

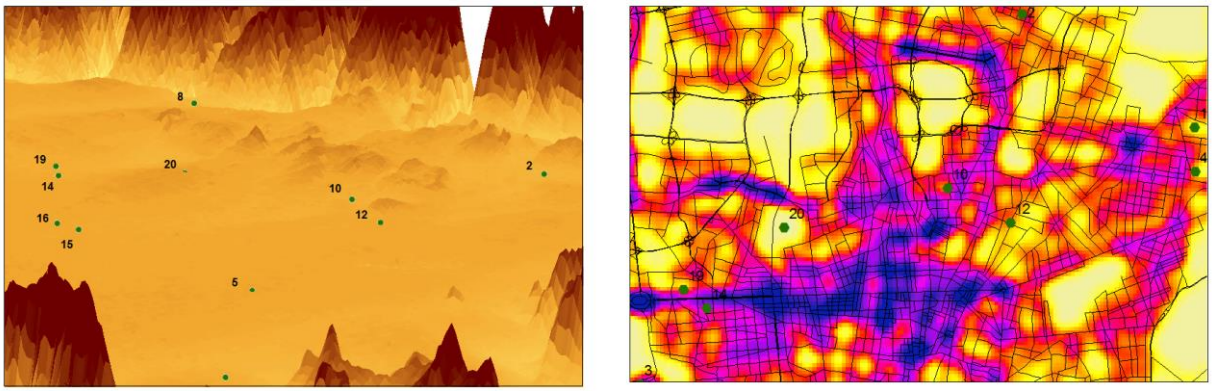
تعداد نمونه‌ی آموزشی	روش	زمان آموزش	دقت	RMSE	R Squared
2000	LaSVM	0:04:56	0.77	0.509902	0.801109000
	SVM	0:03:18	0.77	0.509902	0.801109350
4000	LaSVM	0:09:37	0.82	0.424264	0.752195000
	SVM	0:48:33	0.82	0.424264	0.752194514
6000	LaSVM	0:41:14	0.45	0.854400	0.808722000
	SVM	2:21:36	0.52	0.812404	0.635282458
8000	LaSVM	1:07:25	0.75	0.500000	0.130710000
	SVM	3:34:55	0.74	0.565685	0.447308910
10000	LaSVM	1:59:30	0.84	0.400000	0.139601000
	SVM	5:20:05	0.68	0.565685	0.979202279
12000	LaSVM	2:18:06	0.77	0.509902	0.501109000
	SVM	8:49:31	0.59	0.640312	0.619995111
14000	LaSVM	2:40:41	0.82	0.424264	0.552195000
	SVM	12:26:59	0.85	0.519615	0.643394613
16000	LaSVM	2:45:58	0.45	0.854400	0.808722000
	SVM	18:17:59	0.46	0.832870	0.807215064



شکل ۱۴: پیش‌بینی آلودگی هوا بر روی هر پیکسل

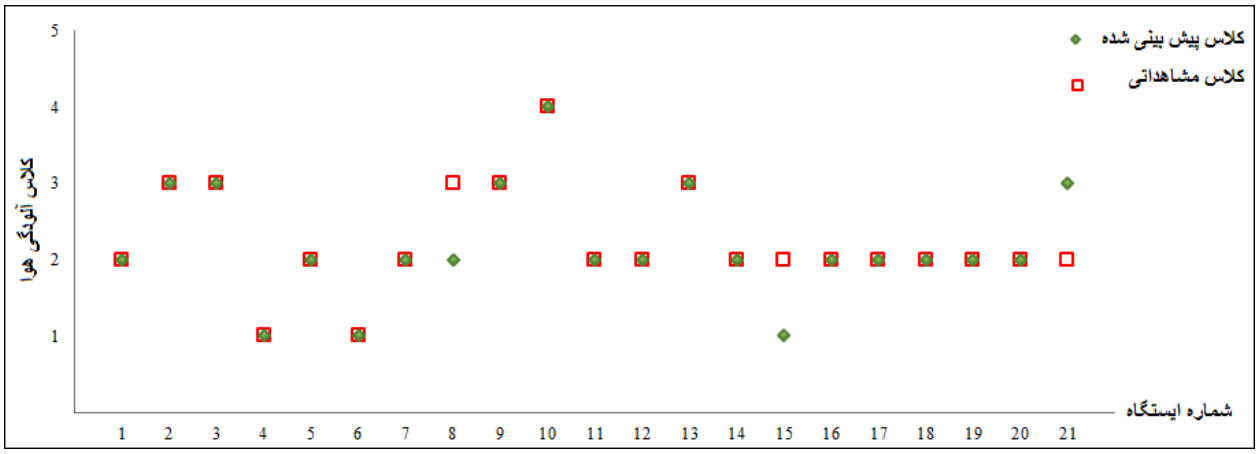


(الف)



(ب) (ج)

شکل ۱۵: بررسی اثر پارامترها در میزان آلودگی هوا: (ب): اثر ترافیک, (ج): اثر ارتفاع متوسط منطقه



شکل ۱۶: مقایسه‌ی مقادیر پیش‌بینی شده بر روی ایستگاه‌ها با مقادیر واقعی

۶- نتیجه‌گیری

هدف این پژوهش ارائه‌ی راهکاری جهت پیش‌بینی آلودگی هوای شهر تهران به‌صورت پویا می‌باشد. این پیش‌بینی برای ۲۴ ساعت آینده و بر مبنای داده‌های جریانی غلظت آلاینده‌ها، داده‌های هواشناسی و داده‌های مکانی انجام می‌شود. الگوریتم LaSVM که توسعه یافته‌ی ماشین بردار پشتیبان می‌باشد، به‌عنوان الگوریتم پویا و کارا، جهت انجام پردازش‌های برخط

مورد استفاده قرار گرفته است. به‌منظور ارزیابی الگوریتم برخط، نتایج حاصل از این الگوریتم با نتایج پیش‌بینی حاصل از ماشین بردار پشتیبان معمولی مورد مقایسه قرار گرفت. مقایسه‌ی نتایج بیانگر افزایش قابل توجه سرعت پردازش‌ها در الگوریتم برخط نسبت به ماشین بردار پشتیبان معمولی می‌باشد. علاوه بر کاهش زمان پردازش در الگوریتم برخط، دقت این الگوریتم نیز بسیار نزدیک به دقت حاصل از ماشین بردار پشتیبان بوده و با گذشت زمان دقت الگوریتم ارائه شده به دقت ماشین بردار پشتیبان همگرا می‌شود. بهبود قابل توجه سرعت آموزش الگوریم ارائه شده در پردازش‌های برخط و دقت مناسب بدست آمده، کارایی این الگوریتم را در پیش‌بینی پویای آلودگی هوا و کار با داده‌های حجیم به اثبات می‌رساند. نکته‌ی دیگر این تحقیق، استفاده از داده‌های مکانی جهت بهبود نتایج می‌باشد. استفاده از داده‌های جغرافیایی امکان پیش‌بینی مکانی آلودگی هوا برای هر نقطه دلخواه را فراهم می‌کند. بر این اساس، سیستمی جهت پیش‌بینی مکانی-زمانی آلودگی هوای شهر تهران برای چند ساعت آینده طراحی شده است. این سیستم می‌تواند داده‌هایی را که به صورت لحظه‌ای از ایستگاه‌های سنجش آلودگی هوا و هواشناسی برداشت می‌شوند به همراه موقعیت نقطه مورد نظر دریافت کرده و آلودگی هوا را در آن نقطه پیش‌بینی کند. با توجه به خطر آلودگی هوا بر سلامت انسان و میزان خساراتی که به طبیعت و جامعه وارد می‌کند، استفاده از این سیستم پیش‌بینی آلودگی هوا امری ضروری می‌نماید.

استفاده از پردازش‌های موازی در پیش‌بینی پویای آلودگی هوای شهر تهران به‌عنوان گام بعدی این پژوهش در نظر گرفته شده است. تقسیم پردازش‌ها بر روی چند پردازشگر می‌تواند تاثیر بسزایی در بهینه شدن زمان پردازش‌ها داشته باشد. همچنین، بدلیل تاثیر قابل توجه پارامترهای الگوریتم در نتایج، استفاده از روش‌هایی همچون Grid Search برای انتخاب این پارامترها در دستور کار قرار دارد. استفاده از آنالیز حساسیت جهت بررسی میزان تاثیر متغیرهای ورودی بر خروجی الگوریتم نیز به عنوان کار آینده در نظر گرفته شده است.

مراجع

[1] Wang, P., et al., A novel hybrid forecasting model for PM10 and SO2 daily concentrations. Science of The Total Environment, 2015. 505(0): p. 1202-1212.

[2] Hasenfratz, D., et al., Participatory air pollution monitoring using smartphones. Mobile Sensing,2012

[3] [Zheng, Y., F. Liu, and H.-P. Hsieh. U-Air: When urban air quality inference meets big data. in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013. ACM.

[4] Sapankevych, N. and R. Sankar, Time Series Prediction Using Support Vector Machines: A Survey. IEEE Computational Intelligence Magazine, 2009. 4(2): p. 24-38.

[5] Finardi, S., et al., A deterministic air quality forecasting system for Torino urban area, Italy. Environmental Modelling & Software, 2008. 23(3): p. 344-355.

[6] Ranzato, L., et al., A comparison of methods for the assessment of odor impacts on air quality: Field inspection (VDI 3940) and the air dispersion model CALPUFF. Atmospheric Environment, 2012. 61: p. 570-579.

[7] Chaloulakou, A., M. Saisana, and N. Spyrellis, Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Science of the Total Environment, 2003. 313(1): p. 1-13.

[8] Kumar, A. and P. Goyal, Forecasting of

daily air quality index in Delhi. Science of the total environment, 2011. 409(24): p. 5517-5523.

[9] Chen, Y., et al., Ensemble and enhanced PM 10 concentration forecast model based on stepwise regression and wavelet analysis. Atmospheric Environment, 2013. 74: p. 34.359-7

[10] Dong, M., et al., PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining. Expert Systems with Applications, 2009. 36(5): p. 9046-9055.

[11] Elangasinghe, M.A., et al., Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. Atmospheric Environment, 2014. 94(0): p. 106-116.

[12] [12] Wahid, H., et al., Neural network-based meta-modelling approach for estimating spatial distribution of air pollutant levels. Applied Soft Computing, 2013. 13(10): p. 4087-4096.

[13] [13] Niska, H., et al., Evolving the neural network model for forecasting air pollution time series. Engineering Applications of Artificial Intelligence, 2004. 17(2): p. 159-167.

[14] [14] Singh, K.P., S. Gupta, and P. Rai, Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment, 2013. 80: p. 426-437.

[15] [15] García Nieto, P.J., et al., A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study. Applied Mathematics and Computation, 2013. 219(17): p. 8923-8937.

[16] Ip, W., et al. Forecasting daily ambient air pollution based on least squares support vector machines. in Information and Automation (ICIA), 2010 IEEE International Conference on. 2010. IEEE.

[17] Reikard, G., Forecasting volcanic air pollution in Hawaii: Tests of time series models. Atmospheric Environment, 2012. 60: p. 593-600.

[18] Juhos, I., L. Makra, and B. Tóth, Forecasting of traffic origin NO and NO2 concentrations by Support Vector Machines and neural networks using Principal Component Analysis. Simulation Modelling Practice and Theory, 2008. 16(9): p. 1488-1502.

[19] Wang, W., C. Men, and W. Lu, Online prediction model based on support vector machine. Neurocomputing, 2008. 71(4): p. 550-558.

[20] Kurt, A. and A.B. Oktay, Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Systems with Applications, 2010. 37(12): p. 7986-7992.

[21] Mintz, D., Technical Assistance Document for the Reporting of Daily Air Quality-the Air Quality Index (AQI). 2012: US Environmental Protection Agency, Office of Air Quality Planning and Standards.

[22] Halek, F., A. Kavouci, and H. Montehaie, Role of motor-vehicles and trend of air borne particulate in the Great Tehran area, Iran. International journal of environmental health research, 2004. 14(4): p. 307-313.

[23] Jenness, J., DEM surface tools v. 2.1. 254. Jenness Enterprises, Flagstaff, Arizona, USA.[Cited 5 Jan 2012.] Available from URL: http://www. jennessent. com/arcgis/surface_area. htm, 2010.

[24] Müller, K.-R., et al., Predicting time series with support vector machines, in Artificial Neural Networks—ICANN'97. 1, 997Springer. p. 999-1004.

[25] Thissen, U., et al., Using support vector machines for time series prediction. Chemometrics and intelligent laboratory systems, 2003. 69(1): p. 35-49.

[26] Vapnik, V.N., Statistical learning theory. Vol. 2. 1998: Wiley New York.

[27] Burges, C.J., A tutorial on support vector machines for pattern recognition. Data

mining and knowledge discovery, 1998. 2(2): p. 121-167.

[28] Ertekin, S., et al. Learning on the border: active learning in imbalanced data classification. in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007. ACM.

[29] Laskov, P., et al., Incremental support vector learning: Analysis, implementation and applications. The Journal of Machine Learning Research, 2006:7 p. 1909-1936.

[30] Bordes, A., et al., Fast kernel classifiers with online and active learning. The Journal of Machine Learning Research, 2005. 6: p. 1579-1619.

[31] Bottou, L., Large-scale kernel machines. 2007: MIT Press.

[32] Yeganeh, B., et al., Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. Atmospheric Environment, 2012. 55: p. 357-365.

[33] Hastie, T. and R. Tibshirani, Classification by pairwise coupling. The annals of statistics, 1998. 26(2:(p. 451-471.

K. N. TOOSI UNIVERSITY OF TECHNOLOGY
FACULTY OF GEODESY AND GEOMATICS ENGINEERING

Journal of Geospatial Information Technology
Vol.3, No.4, Winter 2016

An Online Approach for Spatio-Temporal Prediction of Air Pollution in Tehran using Support Vector Machine

Zeinab Ghaemi [1]*, Mahdi Farnaghi [2], Abas Alimohammadi [3]

1-- MSc. student of geographic information systems, Faculty of Geodesy and Geomatics, K.N.Toosi University of Technology
2- Associate Professor, Faculty of Geodesy and Geomatics, K.N.Toosi University of Technology
3- Assistant Professor, Faculty of Geodesy and Geomatics, K.N.Toosi University of Technology

Abstract

Due to its critical impact on human health and the environment, monitoring and prediction of air pollution have become an important issue during the past decades. Non-linear behavior of air pollution in one hand and high volume of required data on the other hand, intensifies the complexity of air pollution prediction especially in online applications. In order to overcome the deficiencies of traditional methods, this study proposes an online algorithm based on Support Vector Machine (SVM) to predict the time series of air pollution in the city of Tehran, Iran. Prediction is performed on the basis of time series data of pollutant concentrations, weather condition, and geographical parameters such as traffic, surface curvature and local altitude. Evaluation of the outputs shows that prediction errors are within an acceptable range and the online algorithm has an outstanding speed in comparison with the conventional SVM. The overall accuracy of 0.71, RMSE of 0.54 and R^2 of 0.81 prove the efficiency of the proposed algorithm to develop a system to dynamically predict Tehran's air pollution in advance.

Key words: Online air pollution prediction, Support Vector Machine, Time series, Geographic Information System, Big data.

Correspondence Address: Faculty of Geodesy and Geomatics, K.N.Toosi University of Technology, Tehran, Iran.
Tel: +98 21 88779473-4.
Email: zghaemi@mail.kntu.ac.ir