
بسم الله الرحمن الرحيم

دانشگاه صنعتی شاهرود

دانشکده علوم ریاضی

گروه ریاضی کاربردی

کاربرد روش‌های مختلف یادگیری ماشین در استخراج پارامترهای کیفی آب از داده‌های تشعشع طیفی

دانشجو: **مسعود افشاری**

اساتید راهنما:

دکتر داود شاهسونی

دکتر حمید طاهری شهرآیینی

پایان‌نامه جهت اخذ درجه کارشناسی ارشد آمار ریاضی

بهمن ماه ۹۱

تقدیم به

روح پدر بزرگوارم،

مادر مهربانم

و

همه ی عزیزانی که دوستشان دارم

تشکر و قدردانی

حمد و سپاس خدایی را که لطف و کرمش شامل من شد و این توانایی را به من عطا نمود تا گامی کوچک در اقیانوس بی کران علم و معرفت بردارم.

اکنون که به یاری خداوند متعال، این دوره ی تحصیلی را به پایان رسانده ام، ابتدا بر خود می دانم که از خانواده ی صمیمی و مهربانم، کمال تشکر را داشته باشم. همچنین از آقایان دکتر داود شاهسونی و دکتر حمید طاهری شهرآیینی، که در طول این دو سال، همچون شمعی، روشنایی بخش مسیرم بودند، سپاسگزاری می نمایم. همچنین بر خود لازم می دانم که از آقای دکتر مجید عظیم محسنی، که در طول دوران تحصیلات دانشگاهی اینجانب، راهنمایی های خود را از من دریغ ننمودند، کمال تشکر و قدردانی را نمایم. و در پایان از همه ی اساتید بزرگوار، دوستان گرامی و کسانی که در گردآوری این مجموعه، حامی و پشتیبان من بوده اند، کمال تشکر را دارم. امید است که این پایان نامه، گامی هرچند کوچک، در رشد و تعالی عرصه ی علمی کشور عزیزم برداشته باشد.

مسعود افشاری

بهمن ۱۳۹۱

تعهد نامه

اینجانب **مسعود افشاری** دانشجوی دوره کارشناسی ارشد رشته **آمار ریاضی** دانشکده **ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه کاربرد روش‌های مختلف یادگیری ماشین در استخراج پارامترهای کیفی آب از داده‌های تشعشع طیفی تحت راهنمایی **دکتر داود شاهسونی** و **دکتر حمید طاهری شهرآیینی** متعهد می‌شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است .
- کلیه‌ی حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آمدن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه‌ی مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه‌ی مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ ۱۳۹۱/۱۱/۲۹

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه‌ی حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده‌است) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

پهنه‌های آبی، همواره به‌عنوان یکی از عوامل موثر در محیط زیست انسان و سایر موجودات زنده، شناخته می‌شود. از این‌رو، یکی از مهم‌ترین موضوعات پژوهش‌های زیست‌محیطی، بررسی کیفیت پهنه‌های آبی می‌باشد. در این راستا، آگاهی از **پارامترهای کیفی آب**، امری لازم و اجتناب‌ناپذیر است. با توجه به وجود برخی مشکلات در اندازه‌گیری این پارامترها در مناطق گوناگون، و از طرفی به دلیل وجود اثرات واکنش‌های پارامترهای مختلف، امروزه، محققین با به‌کارگیری روش‌های مختلف **یادگیری ماشین** و روش‌های پیشرفته‌ی آماری، اقدام به برآورد پارامترهای مورد نظر می‌کنند. بدین منظور، با توجه به پیشرفت علوم فضایی، استفاده از **داده‌های تشعشع طیفی**، در دستور کار پژوهش‌گران قرار گرفته است. با در نظر گرفتن خطای اندازه‌گیری این داده‌ها، و همچنین تاثیر اتمسفر بر روی داده‌های تشعشع طیفی، همواره **نوفه** به‌عنوان یکی از اجزای جداناپذیر این نوع داده‌ها مطرح است. بنابراین چنانچه روشی قابلیت مدل‌سازی در شرایط نوفه‌ای را دارا باشد، مطالعه‌ی آن روش به‌منظور تبدیل داده‌های تشعشع طیفی به داده‌های کیفی آب، سودمند خواهد بود. در این پایان‌نامه، ضمن معرفی دو روش **جنگل‌های تصادفی** (RF) و **ماشین بردار پشتیبان** (SVM)، عملکرد این دو روش در برآورد پارامترهای کیفی آب در داده-های تشعشع طیفی مورد ارزیابی قرار گرفته است. با توجه به نتایج حاصل می‌توان گفت در به‌کارگیری روش‌های RF و SVM به‌منظور برآورد غلظت کلروفیل-a پایگاه داده NOMAD و برآورد غلظت رنگدانه پایگاه داده SeaBAM، استفاده از متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ منجر به نتایج بهتری نسبت به استفاده از متغیرهای $R_{rs}(\lambda)$ به‌عنوان متغیرهای توضیحی می‌گردد. در برآورد غلظت کلروفیل-a پایگاه داده NOMAD، روش SVM به‌ازای کلیه‌ی مقادیر نوفه، منجر به کمترین مقدار خطای MPAE در بین سه روش RF، SVM و روش یادگیری فعال (ALM) می‌شود. در برآورد غلظت رنگدانه‌ی پایگاه داده

SeaBAM، مقدار خطای RMSE حاصل از روش‌های RF و SVM، نسبت به روش‌های ALM، شبکه‌های عصبی مصنوعی (ANN) و برخی الگوریتم‌های تجربی، تا حد قابل قبولی کاهش می‌یابد. در پایگاه داده MOMO، به‌طور کلی استفاده از دو روش RF و SVM در برآورد کیفیت آب منجر به بهبود نتایج روش ANN می‌گردد. به‌طور کلی با در نظر گرفتن نتایج حاصل و همچنین هزینه‌ی محاسبات، می‌توان گفت که در این تحقیق عملکرد روش RF تا حدی بهتر از روش SVM می‌باشد.

کلمات کلیدی: یادگیری ماشین، جنگل‌های تصادفی، ماشین بردار پشتیبان، پارامترهای کیفی آب، نوفه، داده‌های تشعشع طیفی.

فهرست مطالب

فهرست اَشکال

فهرست جداول

فصل اول

مقدمه

۱-۱ مقدمه

انسان سال‌ها است که در جهت کشف و کنترل پدیده‌های طبیعی و علمی مختلف تلاش می‌کند تا با شناسایی رفتار این پدیده‌ها، بتواند بهترین تصمیم را در بهترین زمان ممکن اتخاذ کند؛ که البته در این زمینه، پیشرفت‌های چشمگیری داشته است. از جمله‌ی این پیشرفت‌ها می‌توان به پیش‌بینی زمان و مکان وقوع بلایای طبیعی، پیش‌بینی قیمت طلا، کشف عوامل موثر در بهینه کردن محصولات کشاورزی و کشف ذرات موثر در اکوسیستم‌های مختلف اشاره کرد. رشد این دستاوردها در زمینه‌های مختلف، متفاوت بوده است. به عنوان مثال، امروزه پیش‌بینی وضعیت آب و هوای یک منطقه برای محققین، کار چندان دشواری نیست، در حالی‌که پژوهش‌گران قادر به پیش‌بینی دقیق زمان و مکان وقوع زلزله نیستند. رفتار بسیاری از این پدیده‌های طبیعی و علمی را می‌توان پس از شناسایی، به‌صورت قوانینی مدوّن درآورد و توسط مدل‌های آماری[1] شبیه‌سازی نمود. یک مدل آماری، با بررسی رفتار مشاهدات در یک پایگاه داده[2]، الگویی را تشکیل می‌دهد که با استفاده از آن می‌توان رفتار مشاهدات جدید را پیش‌بینی نمود. در واقع عملکرد این مدل‌ها همانند فرآیندی است که در آن تعدادی از متغیرها، در نقش ورودی بوده و یک یا چند متغیر نیز نقش خروجی را ایفا می‌کنند.

از گذشته تا به امروز همواره پهنه‌های آبی (اقیانوس‌ها، دریاها، دریاچه‌ها و غیره)، از نواحی مهم و استراتژیک در محیط زیست انسان و سایر موجودات زنده بوده است. از این‌رو، یکی از مهمترین زمینه‌های تحقیق در محیط‌زیست، پایش کیفیت پهنه‌های آب مانند دریاها و تالاب‌ها می‌باشد. آگاهی از پارامترهای کیفی آب، یک امر اجتناب‌ناپذیر در پایداری محیط زیست است، به‌طوری‌که مطالعات بسیاری برای کشف و شناخت بهتر پارامترهای کیفی آب صورت گرفته است. در این زمینه یکی از معضلات محققین، چگونگی

[1] Statistical Models
[2] Data Base

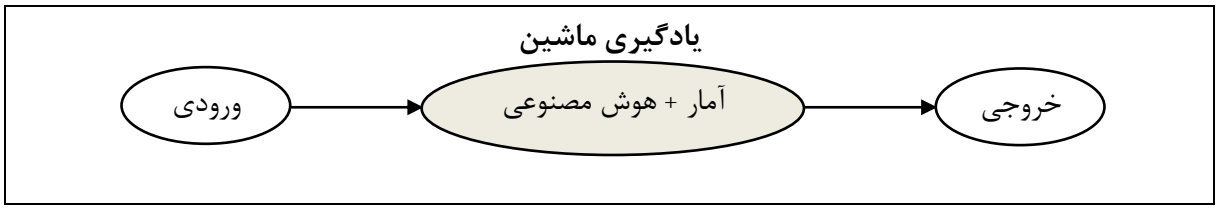
اندازه‌گیری این پارامترها در مناطق مختلف است. به‌همین دلیل امروزه پژوهش‌گران با استفاده از تصاویر ماهواره‌ای و *داده‌های تشعشع طیفی*[1]، مناطق آبی مختلف را مورد مطالعه و بررسی قرار می‌دهند و با اجرای روش‌های مختلف *یادگیری ماشین*[2] و روش‌های پیشرفته‌ی آماری بر روی این داده‌ها، پارامترهای کیفی آب را استخراج و برآورد می‌کنند. تاکنون روش‌های آماری متعددی توسط محققین معرفی گردیده است، اما در زمینه‌ی برآورد پارامترهای کیفی آب، تنها تعداد محدودی از این روش‌ها مورد استفاده قرار گرفته است. با توجه به آن‌که نوفه[3] یکی از اجزای جداناپذیر داده‌های ماهواره‌ای است، همواره برای پژوهش‌گران، روش‌هایی قابل‌توجه است که حساسیت کمتری نسبت به نوفه داشته باشند. در این پایان-نامه ضمن معرفی دو روش ***جنگل‌های تصادفی***[4] و ***ماشین بردار پشتیبان***[5] به عنوان دو روش پیشرفته‌ی آماری و یادگیری ماشین، عملکرد این دو روش در برآورد پارامترهای کیفی آب در داده‌های تشعشع طیفی مورد ارزیابی قرار می‌گیرد.

۱-۲ یادگیری ماشین

در علوم مختلف، تعاریف متعددی برای یادگیری ارایه شده است که یکی از جامع‌ترین تعاریف به این صورت است: *یادگیری عبارتست از به‌دست آوردن دانش و فهم و یا بهبود عملکرد از طریق مطالعه، آموزش یا تجربه*. امروزه بسیاری از ماشین‌های ساخت بشر، قابلیت یادگیری دارند و به علمی که تمرکز آن بر روی ایجاد قابلیت یادگیری در ماشین‌ها است، *یادگیری ماشین* می‌گویند. در واقع یادگیری ماشین به کشف و تنظیم الگوهایی می‌پردازد که بر اساس آن، کامپیوترها (در حالت کلی‌تر ماشین‌ها) توانایی تعلّم و یادگیری پیدا می‌کنند. با ظهور یادگیری ماشین، انسان همواره تلاش کرده تا بتواند با استفاده از قابلیت

[1] Spectral Radiance Data
[2] Machine Learning
[3] Noise
[4] Random Forests
[5] Support Vector Machine

یادگیری ماشین‌ها، پدیده‌های پیرامون خود را کنترل و حتی پیش‌بینی کند. یکی از معضلات عمده‌ی به‌کارگیری روش‌های آماری در بررسی رفتار پدیده‌ها، پیچیدگی ساختار این روش‌ها است، به‌ویژه زمانی که تعداد مشاهدات یا تعداد متغیرها افزایش یابد. علم آمار برای حل این مشکل از سایر علوم کمک گرفته است، به‌طوری‌که با استفاده از ماشین‌ها، دیگر نیازی نیست که محاسبات پیچیده توسط انسان انجام شود. به‌همین دلیل است که بسیاری از دانشمندان، یادگیری ماشین را تقاطع دو علم آمار و هوش مصنوعی[۱] می‌دانند. عملکرد یک ماشین به ترتیبی است که ابتدا رفتار پدیده‌های مختلف در قالب یک پایگاه داده، وارد ماشین می‌گردد و سپس الگوها و روابط بین داده‌ها توسط یادگیری ماشین کشف و شناسایی می‌گردد، که شناسایی این الگوها به‌وسیله‌ی روش‌های آماری صورت می‌گیرد. در نهایت، نتایج در قالب خروجی ماشین ظاهر می‌شود (شکل (۱-۱)).

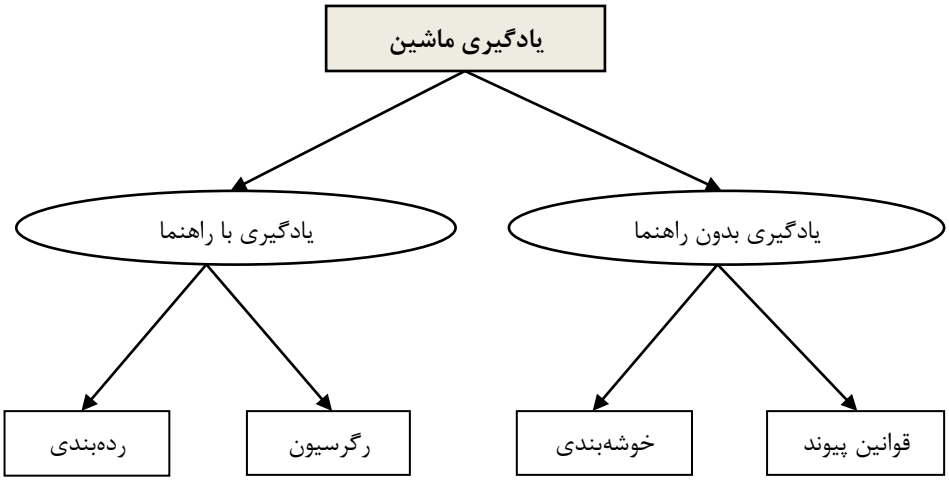


شکل (۱-۱)– ساختار کلی یادگیری ماشین

به‌طور کلی، یادگیری ماشین را می‌توان به دو دسته‌ی یادگیری با راهنما[۲] و یادگیری بدون راهنما[۳] تقسیم کرد. در یادگیری با راهنما، داده‌ها همواره شامل متغیر پاسخ[۴] هستند و هدف، پیش‌بینی متغیر پاسخ با استفاده از متغیرهای توضیحی[۵] می‌باشد. در حالی که در یادگیری بدون راهنما تلاش می‌شود تا ارتباط و الگوهای بین متغیرها بیان گردد و در این نوع یادگیری، متغیر پاسخی وجود ندارد. در یادگیری با راهنما، اگر متغیر پاسخ گسسته باشد، مساله‌ی رده‌بندی[۶] مطرح می‌شود و اگر متغیر پاسخ پیوسته

1 Artificial Intelligence
2 Supervised Learning
3 Unsupervised Learning
4 Response Variable
5 Explanatory Variable
6 Classification

باشد، مساله‌ی رگرسیونی مطرح خواهد شد. از جمله روش‌های مختلف یادگیری ماشین می‌توان به روش-های رگرسیون لجستیک[1]، شبکه‌های عصبی مصنوعی[2] (ANN)، تحلیل سبد بازار[3]، ماشین بردار پشتیبان، جنگل‌های تصادفی، تحلیل خوشه‌ای[4] و غیره اشاره نمود. در حالت کلی، می‌توان روش‌های مختلف یادگیری ماشین را به‌صورت شکل (۱-۲) دسته‌بندی کرد.



شکل (۱-۲)– دسته‌بندی روش‌های یادگیری ماشین

امروزه یادگیری ماشین در زمینه‌های متعددی به کار گرفته می‌شود که در این میان می‌توان به کاربرد آن در شناسایی چهره[5]، کشف تقلب در کارت‌های اعتباری[6]، پیش‌بینی قیمت ارز، طلا و سهام، تخمین پوشش گیاهی مناطق گوناگون، برآورد تأثیر عوامل مختلف در بروز سرطان و عوامل پیشگیری از مبتلا شدن به آن، بررسی دلایل افزایش یا کاهش جمعیت، و حتی برآورد پارامترهای موثر زیست‌محیطی اشاره کرد (وایتن[7] و همکاران، ۲۰۱۱؛ هلجی[8]، ۲۰۱۰).

[1] Logistic Regression
[2] Artificial Neural Networks
[3] Market Basket Analysis
[4] Cluster Analysis
[5] Face Recognition
[6] Credit Card Fraud Detection
[7] Witten I.H.
[8] Helgee E.A.

۱-۳ کاربرد یادگیری ماشین در سنجش از دور

سنجش از دور[1]، علمی است که به جمع‌آوری داده (اطلاعات) درباره‌ی یک شی، منطقه یا پدیده از طریق امواج الکترومغناطیس[2] می‌پردازد. سنجش از دور گرچه در بدو پیدایش به عنوان یک موضوع صرفاً علمی و تحقیقاتی شناخته می‌شد، اما امروزه کاملاً در اختیار شاخه‌های مختلف علوم قرار گرفته و کاربردهای گوناگونی پیدا کرده است. این فن‌آوری در دنیای پیشرفته‌ی امروزی به عنوان یکی از مهمترین منابع جمع‌آوری داده قلمداد می‌شود. وجود انواع زمینه‌های کاربردی و علمی برای سنجش از دور، آن را به‌عنوان یک ابزار قوی و کارآمد برای تولید داده در زمینه‌های مختلف، از جمله علوم مهندسی مطرح نموده است. امروزه اکثر داده‌های سنجش از دور در قالب تصاویر ماهواره‌ای ثبت می‌گردد. در اکثر موارد، این تصاویر به مقادیر عددی تبدیل می‌شوند تا توسط مدل‌ها و روش‌های مختلف قابل پردازش باشند. یکی از کاربردهای مهم این نوع داده‌ها، استفاده از اطلاعات آن در حل مشکلات و مسایل زیست‌محیطی و به خصوص اکوسیستم‌های دریایی است. برای توضیحات بیشتر در مورد سنجش از دور و داده‌های تشعشع طیفی، به پیوست **الف** مراجعه کنید (ونگ[3]، ۲۰۱۰).

۱-۴ تاریخچه‌ی تحقیق

در چهارم اکتبر ۱۹۵۷، نخستین ماهواره‌ی فضایی با نام اسپونیک-۱[4] توسط اتحاد جماهیر شوروی به فضا پرتاب شد که یکی از بزرگ‌ترین دستاوردهای علمی بشر در قرن بیستم محسوب می‌شود. پس از آن، کشورهای مختلفی اقدام به پرتاب ماهواره‌های فضایی نمودند و از اطلاعات و تصاویر ارسالی این ماهواره‌ها به‌منظور پیشبرد اهداف نظامی، زیست محیطی، تجاری و غیره بهره برده‌اند.

---

1 Remote Sensing
2 Electromagnetic
3 Weng Q.
4 Sputnik-1

با پیشرفت علوم فضایی و با توجه به اهمیت پهنه‌های آب در محیط زیست، در سال ۱۹۸۷ اولین ماهواره‌ی مطالعات دریایی و استخراج پارامترهای کیفی NASA[1] به فضا ارسال گردید تا با استفاده از تصاویر ثبت‌شده، و تجزیه و تحلیل داده‌های مربوط به آن، کیفیت پهنه‌های آب مورد بررسی قرار گیرد. حال با توجه به اثرات اتمسفر بر روی این تصاویر، می‌توان گفت که نوفه یکی از اجزای جداناپذیر داده‌های ماهواره‌ای است. لذا توسعه و رشد روش‌های برآوردیابی که دارای حساسیت کمتری نسبت به نوفه باشند، مورد نیاز است. در این راستا چنانچه روشی قابلیت مدل‌سازی در شرایط نوفه‌ای را داشته باشد، مطالعه‌ی کیفیت آب برای تبدیل داده‌های تشعشع طیفی به داده‌های کیفی آب سودمند خواهد بود. در این پژوهش‌ها محققین به‌دنبال یافتن بهترین روش برای برآورد پارامترهای کیفی آب مانند مواد معلق، کلروفیل[2] و مواد آلی محلول زرد رنگ[3] هستند که نقشی کلیدی در محیط زیست دارند.

تاکنون تحقیقات زیادی در رابطه با برآورد پارامترهای کیفی آب با استفاده از داده‌های تشعشع طیفی انجام شده است. در هر یک از این پژوهش‌ها، محققین با استفاده از توانمندی‌ها و خواص روش‌های مختلف مدل‌سازی، به برآورد پارامترهای موردنظر پرداخته‌اند. در این پژوهش‌ها معمولاً مجموعه‌ای از پایگاه داده‌های استاندارد جهت ارزیابی عملکرد این روش‌ها و الگوریتم‌ها، توسط محققین، به‌کار گرفته شده است. در همین زمینه، اُریلی[4] و همکاران (۱۹۹۸) از پایگاه داده SeaBAM[5] به‌منظور ارزیابی عملکرد دو روش نیمه‌تحلیلی و پانزده الگوریتم تجربی در برآورد غلظت کلروفیل استفاده کردند. این الگوریتم‌های تجربی بین سال‌های ۱۹۹۰ تا ۱۹۹۸ توسعه یافته‌اند. علاوه بر این اریلی و همکاران (۱۹۹۸)، برخی از الگوریتم‌های تجربی را بر روی پایگاه داده SeaBAM بهینه کردند و دریافتند که در بین این الگوریتم‌ها،

1 National Aeronautics and Space Administration
2 Chlorophyll
3 Colored Dissolved Organic Matters
4 O'Reilly J.
5 Sea WifFS Bio-Optical Algorithm Mini-Workshop

الگوریتم OC4[1] مناسب‌ترین روش در برآورد غلظت کلروفیل پایگاه داده SeaBAM است. همچنین میشل[2] و کارا[3] (۱۹۹۸)، به‌منظور برآورد غلظت رنگدانه در این پایگاه داده، روش CCO[4] را به‌کار گرفتند. پس از آن ژانگ[5] و همکاران (۲۰۰۳)، با استفاده از روش ANN، به برآورد غلظت کلروفیل در پایگاه داده SeaBAM پرداختند. نتایج به‌دست‌آمده، ثابت می‌کرد روش ANN از دقت بیشتری نسبت به الگوریتم OC4 برخوردار است. همچنین روش ANN حساسیت کمتری نسبت به افزایش نوفه نشان می‌دهد. ژان[6] و همکاران (۲۰۰۳)، روش ماشین بردار پشتیبان را به‌منظور برآورد غلظت کلروفیل در پایگاه داده SeaBAM به‌کار بردند که به عملکردی مشابه روش ANN دست یافتند. همچنین دارکی[7] و همکاران (۲۰۰۵)، به‌منظور برآورد غلظت کلروفیل دریای بالتیک، الگوریتم‌های OC2[8] و OC4 را به‌کار بردند. شرودر[9] (۲۰۰۵)، نیز در رساله‌ی دکترای خود، عملکرد روش ANN را در برآورد برخی از پارامترهای کیفی آب در یک پایگاه داده‌ی کامپیوتری مورد ارزیابی قرار داد. سو[10] و همکاران (۲۰۰۶) با به‌کارگیری روش شبکه‌های عصبی مصنوعی دو قسمتی[11] (BANN) بر روی پایگاه داده SeaBAM، و مقایسه‌ی نتایج حاصل با نتایج مدل ANN و الگوریتم‌های OC2 و OC4، نتایج اریلی و همکاران (۱۹۹۸) و ژانگ و همکاران (۲۰۰۳) را بهبود بخشیدند. در یکی از آخرین پژوهش‌های انجام شده، طاهری شهرآیینی و همکاران (۲۰۰۹) با به‌کارگیری روش یادگیری فعال[12] (ALM) بر روی مجموعه‌ای از پایگاه داده‌های استاندارد، به برآورد برخی از پارامترهای کیفی آب پرداختند.

[1] Ocean Colour 4
[2] Mitchel B.G.
[3] Kuhra M.
[4] CalCOFI Two-band Cubic
[5] Zhang T.
[6] Zhan H.
[7] Darecki M.
[8] Ocean Colour 2
[9] Schroeder T.
[10] Su F.C.
[11] Bi-partite Artificial Neural Networks
[12] Active Learning Method

با توجه به این‌که ذرات معلق، کلروفیل و رنگدانه‌های موجود در پهنه‌های آبی، نقش بسزایی در اکوسیستم‌های دریایی ایفا می‌کنند، پژوهش‌گران در تلاشند تا بهترین روش‌ها و الگوریتم‌ها را به‌منظور برآورد این ذرات به‌کار گیرند. همان‌طور که گفته شد تاکنون روش‌های مختلفی به‌کار گرفته شده است. ولی با توجه به اهمیت نقش این ذرات در محیط زیست، محققین همواره به‌دنبال بهبود این نتایج هستند. لذا اگر روشی ارایه گردد که بتواند با خطای کمتر و دقت بیشتری به برآورد این پارامترها بپردازد، مورد استقبال قرار خواهد گرفت. حال با توجه به این‌که روش ماشین بردار پشتیبان قابلیت پیش‌بینی در داده-هایی با ساختارهای پیچیده را دارا می‌باشد و از آنجایی که به‌ندرت در زمینه‌ی برآورد پارامترهای کیفی آب به‌کار گرفته شده است، استفاده از آن در حل این مساله می‌تواند مفید باشد. از طرفی، تاکنون از روش جنگل‌های تصادفی در این زمینه استفاده نشده است و با توجه به کارایی این روش در پایگاه‌های بزرگ داده و به‌خصوص داده‌های نوفه‌ای، به نظر می‌رسد استفاده از این دو روش منجر به نتایج ارزشمندی در برآورد پارامترهای کیفی آب در داده‌های تشعشع طیفی گردد (طاهری شهرآیینی و همکاران، ۲۰۰۹؛ کانیزارو[1] و کاردر[2]، ۲۰۰۶؛ ژانگ، ۲۰۰۳).

۱-۵ اهداف و ضرورت انجام پایان‌نامه

استخراج و برآورد پارامترهای کیفی آب از طریق داده‌های تشعشع طیفی (یا تشعشعات طیفی) توسط سنجنده[3]های مختلف، از گذشته مدنظر محققین و مهندسین بوده است. اما به خاطر پیچیدگی‌ها و اثرات واکنش‌های پارامترهای مختلف، استفاده از روش‌های پیشرفته‌ی آماری و یادگیری ماشین برای حل این مساله اجتناب‌ناپذیر است. از این‌رو، محققین همواره به‌دنبال به‌کارگیری و توسعه‌ی الگوریتم‌ها و روش‌های مناسبی، جهت برآورد پارامترهای کیفی آب از داده‌های تشعشع طیفی بوده‌اند. تاکنون از روش جنگل‌های

[1] Cannizzaro J.F.
[2] Carder K.L.
[3] Sensor

تصادفی در این زمینه استفاده نشده است. همچنین به‌ندرت روش ماشین بردار پشتیبان در زمینه‌ی برآورد پارامترهای کیفی آب به‌کار گرفته شده است. این مقدمه انگیزه‌ای شد تا در این پایان‌نامه، رویکرد کاربردی آمار مورد توجه قرار گیرد تا از آن طریق بتوان گوشه‌هایی از جایگاه نقش آمار را در سایر علوم به‌نمایش گذاشت. در این راستا، ابتدا دو روش نوین یادگیری ماشین (روش جنگل‌های تصادفی و روش ماشین بردار پشتیبان) مورد مطالعه و بررسی قرار گرفته است. سپس با در نظر گرفتن خواص هر روش، پارامترهای کیفی آب در چندین پایگاه داده، با استفاده از این روش‌ها برآورد شده و در نهایت نتایج این دو روش با نتایج روش‌های پیشین مورد مقایسه و ارزیابی قرار گرفته است.

۱-۶ ساختار پایان‌نامه

بخش‌های مختلف این پایان‌نامه مطابق ذیل تهیه و تنظیم گردیده است.

در فصل دوم، ابتدا پایگاه داده‌هایی که در این تحقیق استفاده شده‌اند معرفی گردیده است. سپس به معرفی دو روش جنگل‌های تصادفی و ماشین بردار پشتیبان، به‌عنوان دو روش نوین یادگیری ماشین، که اساس کار این پایان‌نامه را تشکیل می‌دهند، پرداخته شده است. در فصل سوم، ابتدا الگوریتم کلی تحقیق ذکر می‌شود. سپس به تشریح هر یک از مراحل این الگوریتم پرداخته شده است. در فصل چهارم، ابتدا به ارایه و تحلیل نتایج تجربی پرداخته شده و سپس به مقایسه‌ی نتایج حاصل از این پایان‌نامه، با نتایج تحقیقات گذشته پرداخته شده است. فصل پنجم، به بیان نتیجه‌گیری نهایی این پژوهش و ارایه‌ی پیشنهادات می‌پردازد. ضمن این‌که در پیوست **الف**، به‌طور کامل سنجش از دور و داده‌های تشعشع طیفی، معرفی شده است.

فصل دوم

مواد و روش‌ها

۲-۱ پایگاه داده‌ها

در این زیربخش، به معرفی پایگاه داده‌هایی که در این پژوهش و اکثر پژوهش‌های مشابه، مورد استفاده قرار گرفته‌اند، پرداخته می‌شود. در این میان برخی از این داده‌ها، به‌صورت میدانی جمع‌آوری شده‌اند. این در حالی است که برخی از این پایگاه داده‌ها توسط آزمایشات کامپیوتری تولید شده‌اند. برای این منظور، ابتدا تعریفی از آزمایش کامپیوتری ارایه می‌گردد.

آزمایش کامپیوتری

مطالعه‌ی بسیاری از پدیده‌های علمی در شرایط آزمایشگاهی بنا بر دلایلی همچون داشتن هزینه‌های گزاف انجام آزمایش، طولانی بودن زمان اجرا و عملی نبودن انجام آزمایش به‌خصوص در پدیده‌های زیست‌محیطی، مقدور نیست. از این‌رو، محققین به شبیه‌سازی این پدیده‌ها توسط مدل‌های ریاضی روی آورده‌اند. مدل‌های ریاضی غالباً شامل یک سری معادلات دیفرانسیل معمولی یا معادلاتی با مشتق‌های جزیی از نوع خطی و غیرخطی هستند که متغیرهای موجود در آن پدیده را به نوعی با یکدیگر مرتبط می‌سازند. به برنامه یا کد کامپیوتری که قادر به حل عددی این دستگاه معادلات پیچیده باشد، مدل کامپیوتری و اجرای این مدل با ورودی‌های متفاوت را آزمایش کامپیوتری گویند (جانفدا، ۱۳۹۰).

۲-۱-۱ پایگاه داده NOMAD

پایگاه داده NOMAD[1] شامل داده‌های تشعشع مربوط به ۲۰ طول موج مختلف می‌باشد. اما به‌دلیل این-که در بسیاری از طول موج‌ها، داده‌ی گمشده وجود داشته است، طاهری شهرآیینی و همکاران (۲۰۰۹) پایگاه داده‌ای را از NOMAD استخراج کردند که شامل ۶ متغیر توضیحی است که هر یک معرف میزان انعکاس نور خروجی از آب در طول موج خاصی می‌باشد (طول موج‌های ۴۱۱، ۴۴۳، ۴۸۹، ۵۱۰، ۵۵۵ و

[1] NASA bio-Optical Marine Algorithm Dataset

۶۶۵ نانومتر). در این پایگاه داده، متغیر پاسخ، غلظت کلروفیل-a است. این پایگاه داده شامل ۲۰۹۶ مشاهده است که به طور تصادفی به دو دسته‌ی ۱۰۴۸ تایی به عنوان داده‌های مدل‌ساز[۱] و داده‌های آزمون[۲] تقسیم شده‌اند. جدول (۲-۱) دامنه‌ی تغییرات متغیرهای این پایگاه داده را نشان می‌دهد. پایگاه داده NOMAD از طریق وب‌سایت SeaBASS[۳] برای عموم قابل دسترس می‌باشد. برای کسب اطلاعات بیشتر در مورد این پایگاه داده به وردل[۴] و بیلی[۵] (۲۰۰۵) مراجعه شود.

جدول (۲-۱)- دامنه‌ی تغییرات متغیرهای پایگاه داده NOMAD

Variable	Rrs(411)	Rrs(443)	Rrs(489)	Rrs(510)	Rrs(555)	Rrs(665)	Chl-a
Min	۰/۰۰۰۱۵	۰/۰۰۰۱۹	۰/۰۰۰۳۷	۰/۰۰۰۳۰	۰/۰۰۰۲۲	۰/۰۰۰۰۱۲۳	۰/۰۱۲
Max	۰/۰۳۰۶	۰/۰۳۶۷۶	۰/۰۶۳۸۱	۰/۰۷۷۷۴	۰/۱۱۱۴۸	۰/۰۲۷۶۳	۶۵/۱۴

۲-۱-۲ پایگاه داده SeaBAM

پایگاه داده SeaBAM شامل ۹۱۹ داده‌ی تشعشع طیفی است که در آن متغیر پاسخ، مقدار غلظت رنگدانه می‌باشد. این پایگاه داده شامل ۵ متغیر توضیحی است که هر یک معرف میزان انعکاس نور خروجی از آب در طول موج خاصی می‌باشد (طول موج‌های ۴۱۲، ۴۴۳، ۴۹۰، ۵۱۰ و ۵۵۵ نانومتر). در این پژوهش، دو دسته داده‌ی ۴۵۰ تایی به عنوان داده‌های مدل‌ساز و آزمون از این پایگاه به تصادف انتخاب شده‌است، تا به کمک آن بتوان عملکرد روش‌های RF و SVM را مورد بررسی قرار داد. پایگاه داده SeaBAM نیز از طریق وب‌سایت SeaBASS برای عموم قابل دسترس می‌باشد. جدول (۲-۲) دامنه‌ی تغییرات متغیرهای این پایگاه داده را نشان می‌دهد. برای کسب اطلاعات بیشتر در مورد این پایگاه داده به اُریلی و همکاران (۱۹۹۸)، ماریتورنا[۶] و همکاران (۲۰۰۲) و اُریلی و ماریتورنا (۲۰۰۲) مراجعه کنید.

[1] Training Data
[2] Test Data
[3] Sea WiFS Bio-optical Archive and Storage Systems (http:// seabass.gsfc.nasa.gov)
[4] Werdell P.J.
[5] Bailey S.W.
[6] Maritorena S.

جدول (۲-۲)- دامنه‌ی تغییرات متغیرهای پایگاه داده SeaBAM

Variable	Rrs(412)	Rrs(443)	Rrs(490)	Rrs(510)	Rrs(555)	Pigment
Min	۰/۰۰۰۸۰	۰/۰۰۰۹۱	۰/۰۰۱۳۰۴	۰/۰۰۱۱۹	۰/۰۰۰۷۱	۰/۰۲۶
Max	۰/۰۲۱۴۳	۰/۰۱۸۸۷	۰/۰۲۰۲۵	۰/۰۱۹۹۳	۰/۰۱۷۲۴	۳۳/۴۷۳

۲-۱-۳ پایگاه داده Synthetic

پایگاه داده Synthetic برخلاف دو پایگاه داده NOMAD و SeaBAM که با انجام عملیات میدانی جمع-آوری شده‌اند، توسط آزمایشات کامپیوتری حاصل از اجرای یک مدل تعیینی[1] ساخته شده‌است. این پایگاه داده شامل ۳۰۰ مشاهده است که فقط به‌عنوان داده‌های مدل‌ساز پایگاه داده SeaBAM به‌کار می-رود. این پایگاه داده شامل ۵ متغیر توضیحی است که هر یک معرف میزان انعکاس نور خروجی از آب در طول موج خاصی می‌باشد (طول موج‌های ۴۱۲، ۴۴۳، ۴۹۰، ۵۱۰ و ۵۵۵ نانومتر). همچنین متغیر پاسخ، غلظت رنگدانه می‌باشد. جدول (۲-۳) دامنه‌ی تغییرات متغیرهای این پایگاه داده را نشان می‌دهد. برای کسب اطلاعات بیشتر در مورد این پایگاه داده به ژانگ (۲۰۰۳) یا ژانگ و همکاران (۲۰۰۳) مراجعه شود.

جدول (۲-۳)- دامنه‌ی تغییرات متغیرهای پایگاه داده Synthetic

Variable	Rrs(412)	Rrs(443)	Rrs(490)	Rrs(510)	Rrs(555)	Pigment
Min	۰/۰۰۱۳۱	۰/۰۰۱۱۶	۰/۰۰۱۶۷	۰/۰۰۱۸۷	۰/۰۰۱۳۸	۰/۰۲۵
Max	۰/۰۱۳۳۱	۰/۰۱۱۵۷	۰/۰۰۷۵۳	۰/۰۰۲۹۵	۰/۰۰۲۸۳	۳۵

۲-۱-۴ پایگاه داده MOMO

پایگاه داده MOMO[2] به‌وسیله‌ی آزمایشات کامپیوتری ساخته شده است. فل[3] و فیشر[4] (۲۰۰۱) این پایگاه داده را با استفاده از ابرکامپیوترها و تحت شرایط آزمایشگاهی در دانشگاه برلین تولید کرده‌اند. این پایگاه داده، شامل داده‌های مدل‌ساز و آزمون می‌باشد که هر کدام شامل ۱۰۰۰۰۰ مشاهده است. در این پایگاه

[1] Deterministic Model
[2] Matrix Operator MethOd
[3] Fell F.
[4] Fischer J.

داده، سه متغیر پاسخ غلظت ذرات معلق[1] ، غلظت کلروفیل و غلظت مواد آلی محلول زرد رنگ وجود دارد که به‌طور مجزا بایستی هر یک برآورد گردد. پایگاه داده MOMO شامل ۱۸ متغیر توضیحی استجدول (۲-۴) دامنه‌ی تغییرات متغیرهای این پایگاه داده را نشان می‌دهد.

جدول (۲-۴)- دامنه‌ی تغییرات متغیرهای پایگاه داده MOMO

Variable	min	max
L1	۰/۰۱۷	۰/۰۸۱۴
L2	۰/۰۱۳۹	۰/۰۷۹۲
L3	۰/۰۱	۰/۰۸۲۷
L4	۰/۰۰۸۵۵	۰/۰۸۴۵
L5	۰/۰۰۵۶۹	۰/۰۹۰۵
L6	۰/۰۰۳۹	۰/۰۸۵
L7	۰/۰۰۳۳۴	۰/۰۸۰۳
L9	۰/۰۰۲۷۷	۰/۰۷۷۶
L10	۰/۰۰۲۳۲	۰/۰۶۸
L12	۰/۰۰۲۰۷	۰/۰۶۷۸
L13	۰/۰۰۱۴۸	۰/۰۶۴۳
L14	۰/۰۰۱۳۶	۰/۰۶۳۵
wind	۱/۵	۷/۲۳
pressure	۹۸۰	۱۰۴۰
cos	۰/۲۴۷	۱
obzx	-۰/۶۶۱	۰/۶۶۱
obzy	-۰/۶۶۱	۰/۶۶۱
obzz	۰/۷۵	۱
logchl	-۱/۳	۱/۷
logyel	-۲/۳	۰
logtsm	-۱/۳	۱/۷

که در این پایگاه داده، متغیرهای توضیحی عبارتند از: میزان انعکاس در بالای اتمسفر در طول موج‌های ۴۱۲، ۴۴۳، ۴۹۰، ۵۱۰، ۵۶۰، ۶۲۰، ۶۶۵، ۷۰۹، ۷۵۴، ۷۷۹، ۸۶۵ و ۸۸۵ نانومتر و سرعت باد بر حسب متر بر ثانیه، فشار سطحی بر حسب هکتوپاسکال، کسینوس زاویه‌ی زنیت[2] خورشیدی، حاصلضرب سینوس زاویه‌ی زنیت مشاهده‌گر (ماهواره) و کسینوس اختلاف زاویه‌ی آزیموت[3] بین خورشید و سنجنده،

[1] Total Suspended Matter
[2] Zenith
[3] Azimuth

کسینوس زاویه‌ی زنیت مشاهده‌گر، حاصلضرب سینوس زاویه‌ی زنیت سنجنده و سینوس اختلاف زاویه‌ی آزیموت بین خورشید و مشاهده‌گر.

۲-۲ روش جنگل‌های تصادفی

یکی از روش‌های نوین و پیشرفته‌ی آماری، روش جنگل‌های تصادفی است که در سال ۲۰۰۱ توسط *لیو بریمن*[1]، آماردان دانشگاه برکلی امریکا ارایه شد و از آن پس توسط پژوهش‌گران، به‌طور گسترده‌ای در مدل‌سازی و داده‌کاوی[2] به‌کار گرفته شد. اگرچه اکثر محققان، روش جنگل‌های تصادفی را با نام لیو بریمن می‌شناسند، اما آدله کاتلر[3] نیز در جهت توسعه‌ی این روش، تلاش‌های فراوانی نموده است. نام این روش از روشی به نام جنگل‌های تصمیم تصادفی[4] گرفته شد که توسط تین کام هو[5] (۱۹۹۸) ارایه شده بود.

روش جنگل‌های تصادفی که جزو تکنیک‌های یادگیری ماشین است، قابلیت استفاده برای پیش‌بینی داده‌های کمی و رده‌بندی داده‌های کیفی را داراست. بررسی‌های انجام‌شده حاکی از آن است که از این روش عموماً برای رده‌بندی و کمتر به منظور پیش‌بینی داده‌های کمی استفاده شده‌است. در چند سال اخیر، به‌کارگیری روش جنگل‌های تصادفی در پیش‌بینی داده‌های کمی و رده‌بندی داده‌های کیفی توسط محققین رواج بیشتری پیدا کرده و نتایج این روش نسبت به نتایج سایر روش‌ها، مورد ارزیابی قرار گرفته است. در جدول (۲-۵)، برخی از پژوهش‌هایی که در آن‌ها نتایج روش جنگل‌های تصادفی مورد ارزیابی قرار گرفته‌است دیده می‌شود. در اکثر این تحقیقات، روش جنگل‌های تصادفی، از عملکرد بهتری نسبت به سایر روش‌ها برخوردار بوده است (وریکاس[6] و همکاران، ۲۰۱۰).

[1] Leo Breiman (1928-2005)
[2] Data Mining
[3] Adele Cutler
[4] Random Decision Forests
[5] Tin Kam Ho
[6] Verikas A.

جدول(۲-۵)- برخی از پژوهش‌هایی که در آن‌ها روش جنگل‌های تصادفی به‌کار رفته است

سال	محققین	موضوع تحقیق	نوع متغیر پاسخ	روش‌هایی که جنگل‌های تصادفی نسبت به آن‌ها عملکرد بهتری داشته
۲۰۰۶	یان[1]	عیب‌یابی موتور هواپیما	کیفی	[2]CART, SVM
۲۰۰۹	ویترو[3] و همکاران	کشف تقلب در کارت‌های اعتباری	کیفی	CART, SVM, LR, [4]NB
۲۰۰۹	اسلابینک[5] و همکاران	شناسایی گونه‌های باکتری	کیفی	CART, SVM
۲۰۰۹	ژی[6] و همکاران	پیش‌بینی زنجیره‌ی مشتری	کیفی	[7]DT, SVM
۲۰۰۵	هانکوک[8] و همکاران	پیش‌بینی قدرت نگه‌دارندگی رنگ داروها	کمی	CART, [9]PLS
۲۰۰۷	بوکینژ[10] و همکاران	پیش‌بینی وفاداری مشتری	کمی	[11]MLR, [12]ARDNN
۲۰۰۹	کوژف[13] و همکاران	پیش‌بینی شرایط پوشش گیاهی ایالت ویکتوریای استرالیا	کمی	CART, [14]MTRT
۲۰۱۲	گودرزی و شاهسونی	زمان بازداری نسبی پُلی برم دار کردن فنیل‌اترها	کمی	-

بنیان کار روش جنگل‌های تصادفی، که از این پس به اختصار آن را RF نام می‌بریم، شبیه روش CART است. به‌همین دلیل لازم است که برای معرفی روش RF ابتدا روش CART معرفی گردد. روش CART قابلیت استفاده برای پیش‌بینی داده‌های کمی و رده‌بندی داده‌های کیفی را دارا می‌باشد و از

[1] Yan W.
[2] Classification And Regression Tree
[3] Whitrow C.
[4] Naïve Bayes
[5] Slabbinck B.
[6] Xie Y.
[7] Decision Tree
[8] Hancock T.
[9] Partial Least Squares
[10] Buckinx W.
[11] Multiple Linear Regression
[12] Automatic Relevance Determination Neural Network
[13] Kocev D.
[14] Multi-Target Regression Trees

آنجایی که تمرکز این پایان‌نامه بر روی پیش‌بینی داده‌های کمی است، لذا تنها به بیان بخش رگرسیونی روش CART با نام درخت رگرسیونی پرداخته می‌شود.

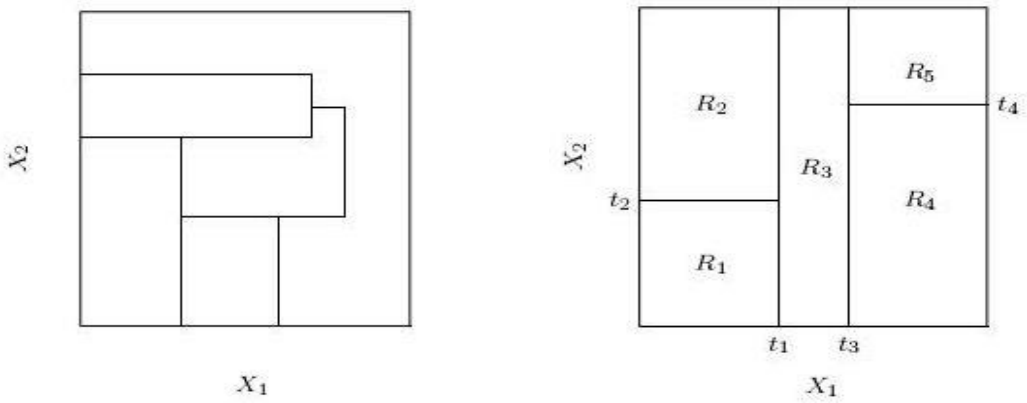
۲-۲-۱ روش درخت رگرسیونی

برای بیان اساس این روش، ابتدا با یک مثال دو متغیره و قابل درک شروع نموده و پس از آن به بیان این روش در حالت چند متغیره پرداخته می‌شود.

۲-۲-۱-۱ روش درخت رگرسیونی در حالت دو متغیره

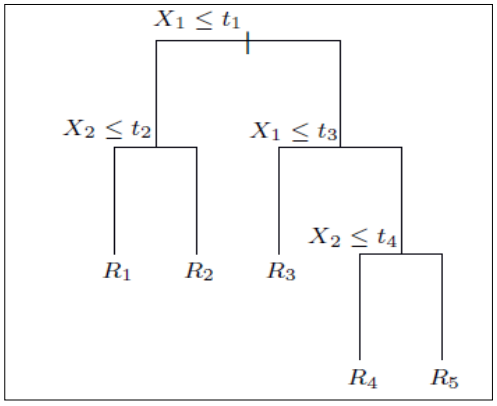
فرض کنید $\mathbf{S}=\{(x_{1i}, x_{2i}) ; i=1,..,n\}$ مجموعه مشاهدات دو متغیر توضیحی X_1 و X_2 و همچنین $\{y_1,\ldots,y_n\}$ مجموعه مقادیر متغیر پاسخ Y باشند. در این روش، فضای متغیرهای توضیحی به گونه‌ای تفکیک می‌شود که در هر ناحیه بتوان رویه‌ی پاسخ را توسط یک مدل ساده مانند یک صفحه تقریب نمود. در واقع هدف آن است که مقادیر متغیرهای توضیحی در هر ناحیه، تا حد ممکن همگون باشند به‌طوری‌که بتوان آن را توسط ساده‌ترین مدل برآورد نمود.

به‌طور کلی، افراز فضای متغیرهای توضیحی را می‌توان به هر شکلی انجام داد. اما در روش درخت رگرسیونی، تنها افرازی قابل قبول است که تمامی نواحی ساخته شده به شکل مربع یا مستطیل باشند. در حالت سه و بیش از سه متغیر، این قطعات به صورت مکعب مستطیل یا آبَرمکعب خواهند بود. در شکل (۲-۱) نمونه‌ای از افراز قابل قبول و غیر قابل قبول در روش درخت رگرسیونی نشان داده شده است.



شکل (۲-۱)- نمونه‌ای از افراز صحیح (قاب راست) و غیر صحیح (قاب چپ) در فضای دو متغیره

در شکل(۲-۲) نوع دیگری از نمایش این افراز نشان داده شده است. دلیل نامگذاری درخت رگرسیونی را می‌توان به نمایش درختی افراز فضای متغیرهای ورودی، نسبت داد که در آن هر مشاهده از نقطه‌ی بالای نمودار درختی وارد شده و در یکی از نواحی افراز قرار می‌گیرد و سپس مقدار $\hat{Y}$ مربوط به همان ناحیه را اختیار می‌کند.



شکل (۲-۲)- نمودار درختی افراز فضای انجام شده در شکل (۲-۱)- قاب راست

اکنون با فرض این‌که فضای متغیرهای X_1 و X_2 به پنج ناحیه‌ی R_1، R_2، ... و R_5 تفکیک شده باشد، به طور قطع هر ناحیه‌ی R_m ; m=1, ... , 5 شامل زیر مجموعه‌ای از داده‌های مشاهده شده به نام A_m است، به‌طوری که $\cup_{m=1}^{5} A_m = S$.

که در آن، S، فضای کل متغیرهای توضیحی می‌باشد. بنا بـه هـدف دنبـال شـده در بنیـان روش درخـت رگرسیونی، می‌توان رویه‌ی پاسخ را در ناحیه‌ی R_m توسط یک مدل ساده یعنی $\hat{Y}_{R_m} = c_m$ تقریب نمـود. بنابراین روش درخت رگرسیونی به‌صورت زیر بیان می‌شود.

$$\hat{y} = \sum_{m=1}^{5} c_m I_{A_m}(x_1, x_2) , \quad (۱-۲)$$

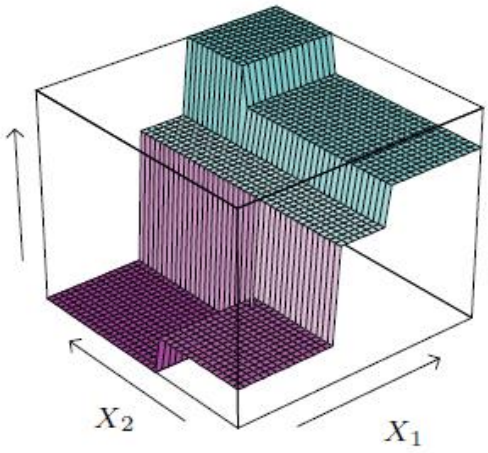
که در آن c_mها مقادیری ثابت می‌باشند و $I_{A_m}(\boldsymbol{x}) = \begin{cases} 1 , & \boldsymbol{x} \in A_m \\ 0 , & \boldsymbol{x} \notin A_m \end{cases}$.

با توجه به روش کمترین توان‌های دوم خطا، به آسانی ثابت می‌شود که c_m، میـانگین مقـادیر متغیـر پاسخ واقع در ناحیه‌ی R_m است. به عبارت دیگر

$$\hat{y}_{R_m} = c_m = \bar{y}_{A_m} = \frac{1}{n_m} \sum_{y_j \in A_m} y_j , \quad (۲-۲)$$

که در آن n_m تعداد مشاهدات مجموعه‌ی A_m است.

شکل (۲-۳) برآورد رویه‌ی پاسخ را به تصویر کشیده است. همان‌طور که دیده می‌شود رویـه‌ی پاسـخ پیوسته نبوده و ناپیوستگی آن در مرز نواحی رخ داده است (هستی[1] و همکاران، ۲۰۰۹).



شکل (۲-۳)- نمایش رویه‌ی پاسخ حاصل از روش درخت رگرسیونی در فضای دو متغیره

[1] Hastie T.

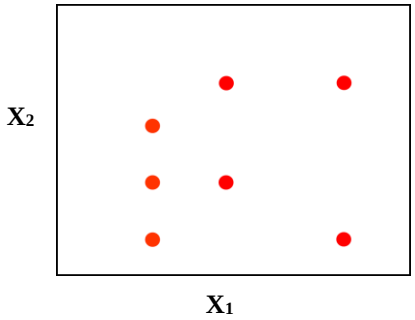
تعریف گره

به مقداری از هر متغیر که افراز در آن نقطه انجام می‌گیرد، گره[1] می‌گویند. در واقع گره محلی است که زیرشاخه‌ها به یکدیگر متصل می‌شوند. در شکل (۲-۲)، مقادیر t_1 , t_2 , t_3 , t_4 ، همان گره‌ها هستند.

۲-۲-۱-۲ الگوریتم تشکیل درخت رگرسیونی

تشکیل درخت رگرسیونی یا افراز فضای نمونه، طی مراحلی سلسله‌مراتبی انجام می‌شود که هر مرحله، بر اساس آن است که تفکیک فضا در راستای کدام متغیر و در چه مقداری از آن متغیر بایستی انجام گیرد. بدین منظور جزییات الگوریتم را با یک مثال شرح می‌دهیم.

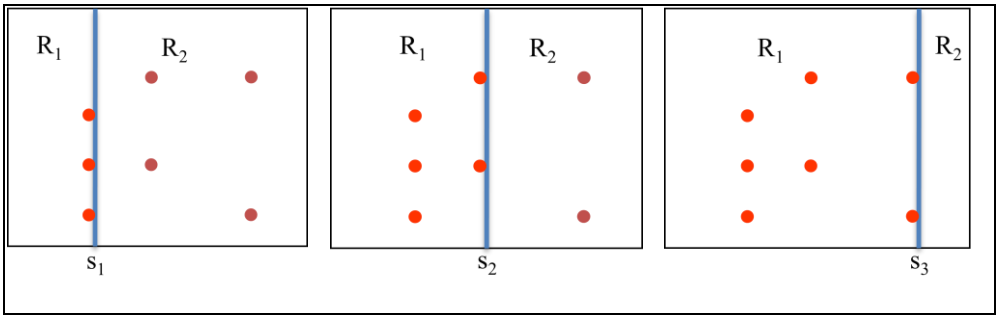
فرض کنید تعداد ۷ مشاهده برای دو متغیر X_1 و X_2 در اختیار باشد که نمودار پراکنش آن‌ها در شکل (۲-۴) آمده است.



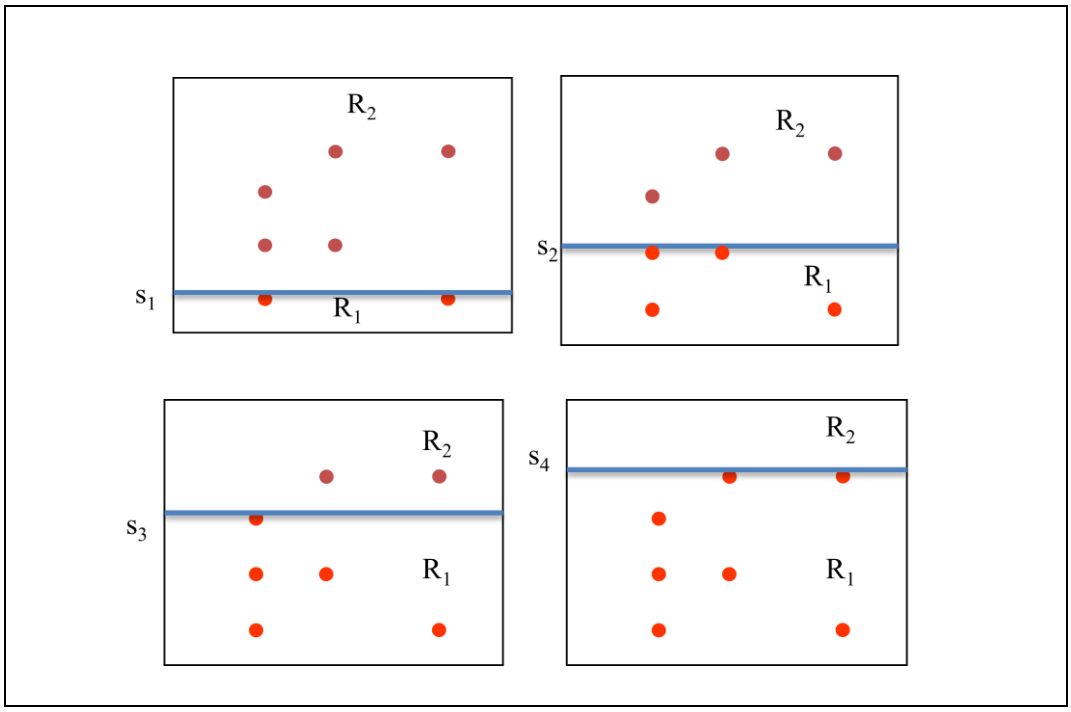
شکل (۲-۴)- نمودار پراکنش داده های دو متغیره

شکل‌های (۲-۵) و (۲-۶) کلیه‌ی افرازهای ممکن و موثر را به‌ترتیب در راستای محورهای X_1 و X_2 نشان می‌دهند.

[1] node



شکل (۲-۵)- سه افراز ممکن در راستای متغیر X_1



شکل (۲-۶)- چهار افراز ممکن در راستای متغیر X_2

برای انتخاب بهترین افراز مرحله‌ی اول، به هر یک از هفت حالت فوق، یک اندازه نسبت داده می‌شود. این اندازه به صورت رابطه‌ی (۲-۳) تعریف می‌گردد.

$$M = \sum_{R_1}(Y_i - \bar{Y})^2 + \sum_{R_2}(Y_i - \bar{Y})^2 , \qquad (۲-۳)$$

به‌عبارت دیگر، M، برابر با مجموع توان‌های دوم خطا در نواحی R_1 و R_2 برای هر یک از نمودارهای شکل‌های (۲-۵) و (۲-۶) است. افرازی که دارای کمترین مقدار M باشد، به عنوان بهترین افراز در

مرحله‌ی اول انتخاب می‌شود و بدین ترتیب بهترین جهت و بهترین نقطه برای تفکیک فضا در گام اول به‌دست خواهد آمد. همچنین می‌توان نوشت

$$M = n_1 Var_1 + n_2 Var_2$$

که در آن Var_1 و Var_2 به ترتیب واریانس مشاهدات واقع در نواحی R_1 و R_2 و همچنین n_1 و n_2 تعداد مشاهدات هریک از دو ناحیه‌ی افراز شده می‌باشد. بنابراین می‌توان M را مجموع موزون واریانس‌های مشاهدات در دو ناحیه‌ی R_1 و R_2 دانست.

اکنون فضا به دو ناحیه تفکیک شده است. در مرحله‌ی دوم و مراحل بعدی، فرآیند فوق در هر یک از نواحی تولید شده تکرار می‌شود تا در نهایت، فضای متغیرهای توضیحی به صورتی شبیه به قاب راست شکل (۲-۱) افراز شود.

۲-۲-۱-۳ روش درخت رگرسیونی در حالت چند متغیره

به طور کلی، بیان ریاضی الگوریتم افراز فضای متغیرهای توضیحی در روش درخت رگرسیونی به‌صورت زیر است. فرض کنید R_1 و R_2 دو ناحیه‌ی تولیدشده در هر مرحله از افراز، بر اثر تفکیک فضای متغیرها در راستای محور متغیر X_j و در نقطه‌ی s باشد، که به‌صورت زیر تعریف می‌گردند

$$R_1(j,s) = \{x \mid x_j \le s\} \quad ; \quad R_2(j,s) = \{x \mid x_j > s\} \ . \qquad (۴-۲)$$

یافتن بهترین افراز، معادل با انتخاب بهینه‌ی متغیر j و نقطه‌ی s بر اساس کمینه شدن مقدار M است. اگر برآورد رویه‌ی پاسخ در دو ناحیه‌ی R_1 و R_2 به ترتیب با c_1 و c_2 نشان داده‌شود، باید این مقادیر به گونه‌ای انتخاب گردد که مجموع توان‌های دوم خطا، $M(c_1,c_2)$ ، در کل دو ناحیه کمینه شود.

$$M_{j,s}(c_1, c_2) = \sum_{R_1(j,s)}(Y_i - c_1)^2 + \sum_{R_2(j,s)}(Y_i - c_2)^2, \qquad (۵-۲)$$

بنا به روش کمترین توان‌های دوم خطا، بهترین c_1 و c_2 به‌ترتیب عبارتند از $\bar{y}_{R_1}$ و $\bar{y}_{R_2}$ یعنی

$$M_{j,s}(\bar{y}_{R_1}, \bar{y}_{R_2}) = min_{c_1,c_2} M_{j,s}(c_1, c_2).$$

برای یافتن بهترین افراز در این مرحله از مراحل بازگشتی، بایستی $M_{j,s}$ را به‌ازای همه‌ی متغیرهای X_j و همه‌ی مقادیر ممکن s یافته و (j,s)-ای را انتخاب کرد که دارای کمترین مقدار M باشد. به‌عبارت دیگر اگر راستای X_L و مقدار K بهترین افراز را القا کند، داریم

$$(L, K) = \arg\min_{j,s} M_{j,s}(\bar{y}_{R_1}, \bar{y}_{R_2}) = \arg\min_{j,s} [\sum_{R_1(j,s)} (Y_i - \bar{y}_{R_1})^2 + \sum_{R_2(j,s)} (Y_i - \bar{y}_{R_2})^2].$$

۲-۲-۱-۴ اندازه‌ی درخت و هرس کردن

در این بخش به یک مطلب مهم در ساختار درخت رگرسیونی پرداخته می‌شود. تفکیک فضای متغیرها تا چه مرحله‌ای می‌تواند ادامه یابد. به‌عبارت دیگر، هر درخت تا کجا می‌تواند رشد کند. اگر درخت به اندازه‌ی کافی رشد نکند ممکن است مدل مناسبی ارایه نشود و اگر هم بیش از حد رشد کند، احتمال این‌که بیش‌برآوردی[1] رخ دهد، بسیار زیاد است. برای انتخاب اندازه‌ی درخت، نیاز به یک پارامتر در ساختار مدل است به‌طوری که این پارامتر مقدار بهینه‌ی اندازه درخت را به‌دست آورد.

در هر مرحله از رشد درخت، مدل حاصل از روش درخت رگرسیونی دقیق‌تر شده و مجموع توان‌های دوم خطا کاهش می‌یابد. یک راه برای توقف رشد درخت می‌تواند بر اساس میزان کاهش مجموع توان‌های دوم خطا باشد، بدین ترتیب که اگر در یک مرحله از رشد درخت، مجموع توان‌های دوم خطا دچار کاهش چندانی نشود می‌توان از رشد درخت در آن مرحله صرف نظر کرد. این میزان کاهش در مجموع توان‌های دوم خطا می‌تواند در اختیار کاربر باشد.

[1] Over Stimation

فرض کنید درختی را که به طور کامل رشد کرده با T_0 نشـان داده و T یکـی از زیـردرخـت‌هـای T_0 باشد، یعنی $T \subset T_0$. همچنین فرض کنید $Q_m(T)$ نشان‌دهنده‌ی میانگین مجموع توان‌هـای دوم خطـا در ناحیه‌ی R_m از درخت T باشد.

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \epsilon R_m} (Y_i - \bar{y}_m)^2 , \qquad (۶-۲)$$

که در آن، N_m تعداد مشاهدات در ناحیه‌ی R_m می‌باشد. اکنون با استفاده از رابطه (۲-۶)، معیار هزینه‌ی پیچیدگی مدل، برای تشکیل درخت T به صورت زیر تعریف می‌گردد که شامل یک پارامتر برای کنترل اندازه‌ی درخت است.

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|, \qquad \alpha > 0, \qquad (۷-۲)$$

که در آن، $|T|$ بیانگر تعداد گره‌های درخت T می‌باشد و α پارامتری است که ارتباط بین اندازه‌ی درخت و نیکویی برازش[1] مدل را کنترل می‌کند. در واقع $C_\alpha(T)$ شامل دو بخش است که بخش اول آن مبین مجموع موزون توان‌های دوم خطا در کلیه‌ی افرازهای مربوط به درخت T است و بخش دوم آن که شامل پارامتر α است، نقش کنترل‌کننده‌ی اندازه‌ی درخت را دارد. بهترین هرس درخت کامل T_0، زیردرختی است مانند T_α که به ازای هر α، مقدار هزینه‌ی پیچیدگی یعنی $C_\alpha(T)$ را کمینه کند. افزایش مقدار α موجب کاهش اندازه‌ی درخت T_α و کاهش آن منجر به افزایش اندازه‌ی درخت T_α می‌شود. همچنین قابل ذکر است که به ازای هر مقدار α، تنها یک درخت یکتا T_α وجود دارد که منجر به کمینه کردن $C_\alpha(T)$ می‌شود (هستی و همکاران، ۲۰۰۹).

در پایان این بخش باید یادآور شد که روش درخت رگرسیونی با بیش‌برآوردی همراه اسـت کـه بـرای جلوگیری از این مشکل، درخت نیاز به هرس دارد.

[1] Goodness of Fit Test

۲-۲-۲ جنگل‌های تصادفی

اساس روش RF وابسته به ماهیت روش درخت رگرسیونی است و از آن‌جایی که، در زیربخش۲-۲-۱، ساختار درخت رگرسیونی به طور کامل معرفی گردید، اینک می‌توان به معرفی روش RF پرداخت.

در روش RF مجموعه‌ای از درخت‌های رگرسیونی تشکیل می‌شوند و هر درخت مدلی را تولید می‌کند که مدل نهایی، برآیند یا ترکیبی از همه‌ی این مدل‌ها است. به عبارت دیگر هر یک از این درخت‌ها سهمی در مدل نهایی دارد. یکی از تفاوت‌های اساسی RF با درخت رگرسیونی آن است که در درخت رگرسیونی برای افراز فضای متغیرها در هر مرحله، از کلیه‌ی متغیرها استفاده می‌گردد در حالی که در RF فقط از زیر مجموعه‌ای از متغیرها استفاده می‌شود. دیگر تفاوت اساسی این دو روش، داده‌هایی است که در ساخت مدل شرکت دارند، به این معنا که در درخت رگرسیونی، همه‌ی داده‌ها در ساخت مدل شرکت دارند اما در روش RF تنها از بخشی از داده‌ها در ساخت مدل استفاده می‌شود. به‌طور کلی می‌توان گفت روش RF، ترکیبی از چندین درخت رگرسیونی است که در ساخت آن چندین نمونه‌ی باجایگذاری از داده‌ها شرکت دارند و در هر درخت برای ساخت هر گره، تنها یک زیر مجموعه‌ی تصادفی از متغیرهای توضیحی شرکت می‌کند.

۲-۲-۲-۱ الگوریتم روش جنگل‌های تصادفی

فرض کنید $(\boldsymbol{X}_i, Y_i)\ ;\ i = 1, \dots, N$ مجموعه‌ی داده‌های مدل‌ساز باشد که در آن، $\boldsymbol{X}_i = (X_{i1}, \dots, X_{iM})$ برداری از M متغیر توضیحی و Y_i متغیر پاسخ متناظر آن است. اگر تعداد کل درخت‌های مدل با $ntree$ نشان داده شود، مراحل پنج‌گانه‌ی زیر بیانگر الگوریتم ساخت درخت i-ام $(i = 1, \dots, ntree)$ است.

۱) یک نمونه‌ی Nتـایی بـه روش بـاجایگـذاری از مجموعـه‌ی داده‌هـای مـدل‌سـاز گرفتـه مـی‌شـود. زیرمجموعه‌ای از داده‌های اصلی که در این نمونه حضور ندارند را OOB[1] نامیده که برای هر درخت نقش داده‌های آزمون را ایفا می‌کند.

۲) به‌طور تصادفی از بین M متغیر توضیحی، m متغیر انتخاب می‌شود ($m \ll M$). برای تقسیم فضای متغیرهای توضیحی به دو قسمت و بر اساس اصول درخت رگرسیونی، فقـط از ایـن m متغیـر و نمونـه‌ی Nتایی انتخاب شده در گام ۱ استفاده می‌شود تا بهترین متغیر و بهترین نقطه‌ی افراز در اولـین مرحلـه از مراحل بازگشتی به‌دست آید. محدودیت ما آن است که تعداد مشاهدات موجود در هـر یـک از دو ناحیـه بایستی بیشتر از n_r باشد و n_r پارامتری است که در اختیار کاربر بوده و بیانگر حـداقل تعـداد مشـاهدات موجود در هر ناحیه است. معمولاً در مدل رگرسیونی، $m = \frac{M}{3}$ پیشنهاد مـی‌شـود و در مـدل رده‌بنـدی، $m = \sqrt{M}$ در نظر گرفته می‌شود.

۳) برای هر یک از دو ناحیه‌ی تولید شده در گام ۲، **مجدداً** به‌طور تصادفی از بین M متغیر توضیحی، m متغیر انتخاب می‌گردد و با استفاده از همـان نمونـه‌ی Nتـایی انتخـاب شـده در گـام ۱، روش درخـت رگرسیونی اعمال می‌گردد. این عمل منجر به افراز هر یک از نواحی موجود به دو قسمت می‌شود. لازم بـه ذکر است که افراز نواحی در صورتی انجام می‌شود که تعداد مشاهدات موجود درکـل آن ناحیـه بیشـتر از $2n_r$ باشد. یعنی اگر هر یک از نواحی دارای تعداد مشاهداتی کمتر از $2n_r$ باشد نباید افـراز دیگـری روی هیچ‌یک از آن نواحی صورت گیرد.

۴) گام ۳ برای تمام نواحی افراز شده تا زمانی تکرار می‌شود که تعداد مشاهدات در تمامی این نواحی کمتر از $2n_r$ باشد.

[1] Out of Bag

۵) حال درخت رگرسیونی i-ام طوری تشکیل شده است که فضای متغیرهای توضیحی به r_i ناحیه‌ی $R_{i\,1}, R_{i\,2}, \dots, R_{i\,r_i}$ تقسیم گردیده است و تعداد مشاهدات هر ناحیه کمتر از $2n_r$ می‌باشد. مدل درخت رگرسیونی به‌دست آمده، به‌صورت زیر است.

$$\hat{f}_i(\boldsymbol{x}) = \sum_{j=1}^{r_i} \hat{c}_j I_{R_{i\,j}}(\boldsymbol{X}), \qquad (۸-۲)$$

که در آن، $\hat{c}_j = \bar{y}_j$ و $I_{R_{i\,j}}(\boldsymbol{X}) = \begin{cases} 1, & \boldsymbol{X} \in R_{i\,j} \\ 0, & \boldsymbol{X} \notin R_{i\,j} \end{cases}$.

از آن‌جایی که در روش RF تعداد *ntree* درخت رگرسیونی وجود دارد، می‌توان گفت که تعداد *ntree* مدل به‌صورت معادله‌ی (۲-۸) خواهیم داشت. اگر برای مقدار مشاهده‌شده‌ی x، مدل خروجی درخت i-ام را به‌صورت $\hat{f}_i(x)$ نشان دهیم، آن‌گاه برآورد متغیر پاسخ در این نقطه، با میانگین‌گیری از مقادیر $\hat{f}_i(x); i = 1, \dots, ntree$ یعنی

$$\hat{y}(x) = \frac{1}{ntree} \sum_{i=1}^{ntree} \hat{f}_i(x), \qquad (۹-۲)$$

به‌دست می‌آید.

لازم به ذکر است که در روش RF، تعداد درخت‌ها (پارامتر *ntree*) ، تعداد متغیرهای به تصادف انتخاب‌شده (پارامتر m) و تعداد حداقل مشاهدات در هر افراز (پارامتر n_r)، پارامترهایی هستند که قابل تغییر بوده و توسط کاربر تعیین می‌گردند (بریمن، ۲۰۰۱؛ استورلی[1] و همکاران، ۲۰۰۹).

۲-۲-۲-۲ تعیین اهمیت متغیرها در روش جنگل‌های تصادفی

در ساختار روش RF ، امکانی وجود دارد که می‌توان میزان اهمیت متغیرها[2] را در مدل، تعیین نموده و متغیرهایی که دارای نقش بیشتری در هر درخت و در مدل نهایی هستند شناسایی شود. همان‌طور که در الگوریتم روش RF اشاره شد، برای تشکیل هر درخت، یک نمونه‌ی باجایگذاری از داده‌های اصلی مورد

[1] Storlie C.B.
[2] Variable Importance

استفاده قرار می‌گیرد. همچنین داده‌هایی را که در این نمونه حضور ندارند را *OOB* نامیدیم که به نوعی نقش داده‌های آزمایشی را برای ارزیابی آن درخت ایفا می‌کنند. فرض کنید خطای پیش‌بینی Y برای داده‌های *OOB* در درخت i-ام، با نماد $EOOB_i$ نشان داده شود. برای تعیین اهمیت متغیر j-ام، مقادیر این متغیر را به‌طور تصادفی، nPerm مرتبه (nPerm، پارامتری است که در اختیار کاربر می‌باشد) جابجا کرده و مجدداً خطا، به ازای مجموعه‌ی جدید، محاسبه می‌شود که مقدار این خطا با نماد $\widehat{EOOB_i^j}$ نشان داده می‌شود. میزان اهمیت متغیر j-ام در مدل درخت i-ام را با $VI_i(X^j)$ نشان داده و به‌صورت زیر تعریف می‌گردد.

$$VI_i(X^j) = \widehat{EOOB_i^j} - EOOB_i\,, \qquad (۱۰-۲)$$

سرانجام، میزان اهمیت متغیر j-ام در مدل نهایی RF به صورت زیر است.

$$VI(X^j) = \frac{1}{ntree}\sum_{i=1}^{ntree}\left(VI_i(X^j)\right), \qquad (۱۱-۲)$$

قابل ذکر است که اندازه‌ی اهمیت یک متغیر، به تنهایی قابل تفسیر نبوده و فقط برای رتبه‌بندی متغیرها بر اساس اهمیت آن‌ها در مدل به‌کار می‌رود (بریمن، ۲۰۰۱؛ جنر[1] و همکاران، ۲۰۱۰).

۲-۲-۲-۳ مزیت‌های روش جنگل‌های تصادفی

الف) به دلیل شرکت کردن نمونه‌ای از داده‌ها در تشکیل مدل، این روش از لحاظ هزینه محاسبات مقرون به‌صرفه است. ضمن این‌که از دقت قابل قبولی نیز برخوردار می‌باشد.

ب) این روش، میزان اهمیت متغیرها را در مدل نهایی مشخص می‌کند. این ویژگی در بسیاری از پژوهش‌ها، حایز اهمیت است. شناسایی متغیرهای مهم می‌تواند کمک‌های فراوانی به محققین کند.

[1] Genuer R.

پ) روش RF، در مجموعه داده‌هایی که متغیرهای توضیحی زیادی وجود دارد، از کارایی بالایی برخوردار است. دلیل این ویژگی انتخاب تصادفی m متغیر از M متغیر کل در هر مرحله از رشد مدل می-باشد

ت) در این روش، با وجود آن‌که هر درخت به طور کامل رشد می‌کند و هرس نمی‌گردد، با این حال مدل نهایی دچار بیش برآوردی نمی‌شود.

ث) این روش، قابلیت استفاده در رده‌بندی داده‌های کیفی و پیش‌بینی داده‌های کمی را دارا می‌باشد.

ج) در صورت در اختیار داشتن تعداد مشاهدات کم، با توجه به نقش داده‌های *OOB*، می‌توان از آن به عنوان داده‌های آزمون استفاده کرد (اشتینبرگ[1] و همکاران، ۲۰۰۴).

۳-۲ روش ماشین بردار پشتیبان

یکی دیگر از روش‌های پیشرفته‌ی یادگیری ماشین، روش ماشین بردار پشتیبان است. این روش در سال ۱۹۹۲ برای اولین بار توسط *ولادیمیر وپنیک*[2] محقق روسی دانشگاه لندن ارایه شد. روش ماشین بردار پشتیبان که جزو روش‌های هسته[3] در یادگیری ماشین محسوب می‌شود، از آن دسته روش‌هایی است که هم برای رده‌بندی داده‌های کیفی و هم برای پیش‌بینی داده‌های کمی قابل استفاده است. شهرت این روش در ابتدا به دلیل عملکرد موفقیت‌آمیزش در تشخیص حروف و اعداد دست‌نویس در مقایسه با روش شبکه‌های عصبی مصنوعی بوده است.

[1] Steinberg D.
[2] Veladimir Vapnik
[3] Kernel

بررسی‌ها حاکی از آن است که در اکثر تحقیقات، از روش ماشین بردار پشتیبان به منظور رده‌بندی داده‌های کیفی استفاده شده است و به‌ندرت جهت پیش‌بینی داده‌های کمی به‌کار رفته است. از این روش تاکنون در پژوهش‌هایی مانند تشخیص چهره (لی[1] و همکاران، ۲۰۱۱)، تشخیص دست‌خط (کومار[2] و همکاران، ۲۰۱۲)، تشخیص نوعی فلجی در گوسفندان (کونچوا[3] و همکاران، ۲۰۰۷)، و کشف تقلب در کارت‌های اعتباری (ویترو و همکاران، ۲۰۰۹) استفاده شده است. در هر یک از این پژوهش‌ها، نتایج روش ماشین بردار پشتیبان با نتایج روش‌های مختلفی مانند شبکه‌های عصبی مصنوعی، درخت رگرسیونی و درخت رده‌بندی مورد مقایسه و ارزیابی قرار گرفته است که در بسیاری از موارد، استفاده از روش ماشین بردار پشتیبان منجر به نتایج بهتری شده است. ژان و همکاران (۲۰۰۳)، روش ماشین بردار پشتیبان را به منظور برآورد غلظت کلروفیل در پایگاه داده SeaBAM به‌کار بردند که نتایج تحقیق نشان داد روش ماشین بردار پشتیبان نسبت به روش شبکه‌های عصبی مصنوعی عملکرد بهتری دارد. همچنین پال[4] و همکاران (۲۰۱۱)، روش ماشین بردار پشتیبان را به‌منظور پیش‌بینی میزان آب‌شستگی پایه‌ی پل‌ها به‌کار بردند (همل[5]، ۲۰۰۹؛ شولکوف[6] و اسمولا[7]، ۲۰۰۲).

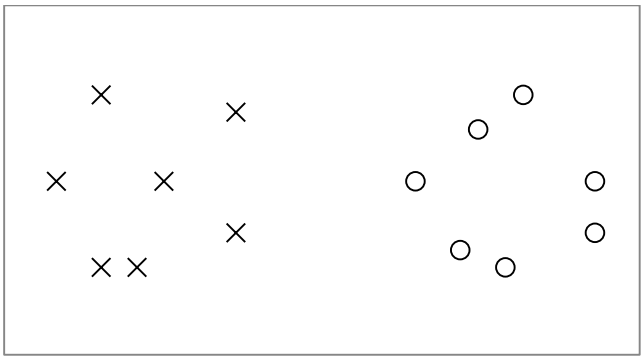
به‌منظور سهولت در نوشتار، از این پس روش ماشین بردار پشتیبان را به اختصار با SVM نمایش می‌دهیم. با توجه به این‌که بیان روش SVM برای حالت رده‌بندی داده‌های کیفی، ساده‌تر و قابل فهم‌تر از حالت پیش‌بینی داده‌های کمی می‌باشد، ابتدا روش SVM برای ساده‌ترین حالت رده‌بندی (داده‌های دو رده‌ای[8]) ارایه می‌گردد و سپس به مدل رگرسیونی آن پرداخته خواهد شد.

[1] Li W.
[2] Kumar R.
[3] Kuncheva L.I.
[4] Pal M.
[5] Hamel L.
[6] Scholkopf B.
[7] Smola A.J.
[8] Class

۲-۳-۱ اساس روش SVM

مبنای کار رده‌بندی در روش SVM، رده‌بندی خطی داده‌ها است. در این روش رده‌بندی، خط ممیزکننده[1] به‌گونه‌ای انتخاب می‌شود که بیشترین فاصله را بین رده‌ها ایجاد کند، که در اصطلاح به این فاصله، حاشیه‌ی اطمینان[2] (حاشیه‌ی امنیت) می‌گویند. در مسایلی هم که داده‌ها به‌صورت خطی جداپذیر نباشد، داده‌ها به فضایی با ابعاد بالاتر نگاشت پیدا می‌کند تا بتوان آن‌ها را به طور خطی جدا کرد.

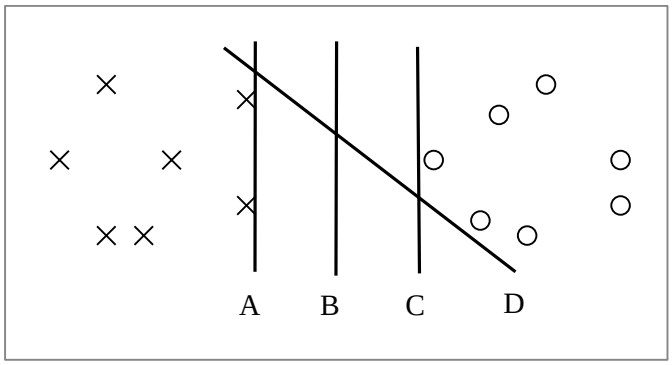
شکل (۲-۷) را در نظر بگیرید که در آن هر یک از نقاط، متعلق به یکی از دو رده‌ی جداپذیری هستند که توسط O و X نشان داده شده‌اند.



شکل (۲-۷)- نمودار پراکنش داده‌های دو رده‌ای

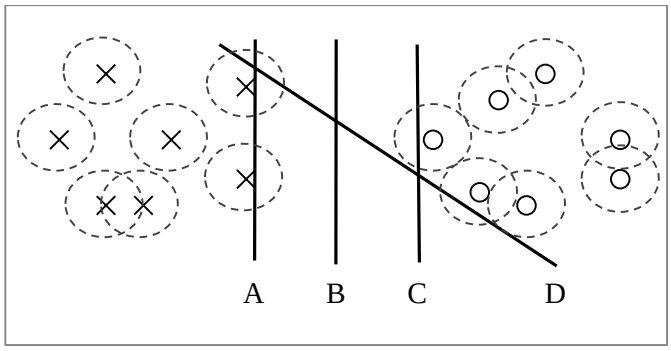
همان‌طور که در شکل (۲-۸) دیده می‌شود، برای رده‌بندی این داده‌ها خطوط ممیزکننده‌ی مختلفی نظیر خطوط A، B، C و D می‌توان رسم کرد که قادر به رده‌بندی داده‌ها هستند. در روش SVM بهترین خط ممیزکننده از بین تمامی خطوط ممکن، انتخاب می‌گردد. اگر این داده‌ها توام با نوفه باشد، عملکرد خطوط ممیزکننده‌ی مذکور، یکسان نخواهد بود.

[1] Classifier
[2] Margin



شکل (۲-۸)- نمونه‌ای از خطوط ممیزکننده برای داده‌های با دو رده‌ی جداپذیر

همان‌طور که در شکل (۲-۹) دیده می‌شود با اضافه شدن نوفه به این داده‌ها، خط B، بهترین خط ممیزکننده‌ی دو رده خواهد بود. دلیل آن هم این است که خط B بیشترین حاشیه امنیت (حاشیه اطمینان) را ایجاد می‌کند.



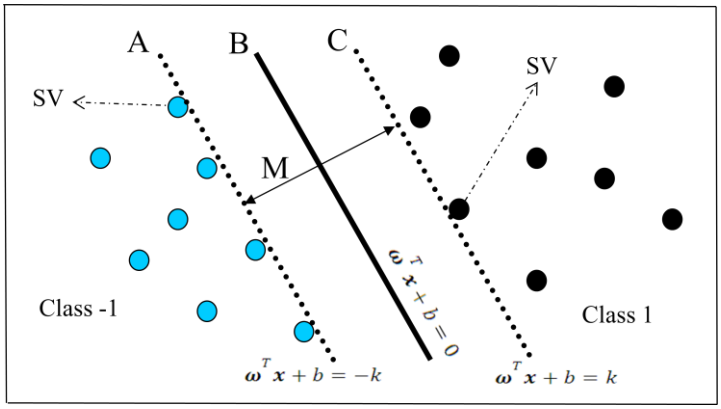
شکل (۲-۹)- عملکرد خطوط ممیزکننده برای داده‌های نوفه‌دار

توجه کنید که در شکل فوق دایره‌های حول داده‌ها، بیانگر ناحیه‌ی داده‌های نوفه‌ای می‌باشد (وپنیک، ۱۹۹۸؛ وپنیک، ۱۹۹۹).

۲-۳-۲ رده‌بندی خطی داده‌های تفکیک‌پذیر دو رده‌ای

شکل (۲-۱۰) را در نظر بگیرید. فرض کنید $(\mathbf{X}_i, Y_i)$; $i = 1,\ldots,n$ مجموعه داده‌های مدل‌ساز باشد که در آن $\mathbf{X}_i = (X_{i1}, X_{i2})$ برداری از دو متغیر توضیحی و $y_i \in \{-1,1\}$ متغیر پاسخ است. همچنین فرض

کنید داده‌ها تفکیک‌پذیر باشند. در روش SVM بهترین خط ممیزکننده بایستی از بین بی‌شمار خط ممیز کننده‌ی ممکن، طوری انتخاب گردد که بیشترین فاصله را بین دو کلاس ایجاد کند و فاصله‌ی آن از دو کلاس یکسان باشد. فرض کنید B، خط ممیزکننده‌ی بهینه باشد که فاصله‌ی آن از خطوط مرزی (A و C) دو کلاس یکسان است. در اصطلاح به داده‌هایی که مماس با خطوط مرزی قرار می‌گیرند بردار پشتیبان[1] (SV) گویند. در روش SVM، خطوط جداکننده‌ی رده‌ها، بر اساس همین نقاط SV شکل می‌گیرد که در بخش‌های بعدی به نقش این نقاط اشاره خواهد شد.



شکل (۲-۱۰)- نمایش خط جدا کننده در داده‌های دو رده‌ای

در واقع در بین تمامی خطوط ممیزکننده‌ی ممکن، B خطی است که بیشترین فاصله را از خطوط مرزی A و C دارد (دارای بیشترین حاشیه امنیت باشد). در شکل (۲-۱۰)، M بیانگر حاشیه امنیت است. توجه شود که در ابعاد بالاتر، خط جداکننده به صفحه و ابرصفحه تبدیل می‌شود.

همان‌طور که در شکل (۲-۱۰) دیده می‌شود اگر معادله‌ی خط جداکننده‌ی دو کلاس به‌صورت $\boldsymbol{\omega}^T\boldsymbol{x} + b = 0$ نمایش داده شود، آن‌گاه معادلات خطوط مرزی دو کلاس به‌صورت $\boldsymbol{\omega}^T\boldsymbol{x} + b = k$ و $\boldsymbol{\omega}^T\boldsymbol{x} + b = -k$ خواهد بود (مقدار k با توجه به فاصله‌ی خطوط A و C از خط B حاصل می‌گردد).

[1] Support Vector

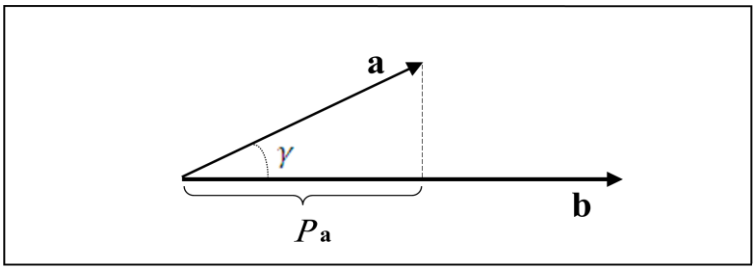
با توجه به این‌که اساس روش SVM به گونه‌ای است که بایستی حاشیه امنیت، M، بیشینه گردد، در معادله‌ی خط جداکننده بایستی این موضوع لحاظ شود. برای به‌دست آوردن مقدار بیشینه‌ی M لازم است که ابتدا قضیه‌ی تصویر[1] بیان شود.

قضیه تصویر

اگر $\mathbf{a}$ و $\mathbf{b}$ بردارهایی در فضای $\Re^n$، با زاویه‌ی بین γ باشند (شکل (۲-۱۱))، آن‌گاه $P_{\mathbf{a}}$ را تصویر $\mathbf{a}$ روی $\mathbf{b}$ نامند و مقدار آن را با استفاده از رابطه‌ی زیر به‌دست می‌آورند.

$$P_{\mathbf{a}} = \|\mathbf{a}\|.\cos\gamma = \frac{\mathbf{a}.\mathbf{b}}{\|\mathbf{b}\|} \qquad (۱۲-۲)$$

که در آن $\|\mathbf{a}\|$ بیانگر اندازه‌ی بردار $\mathbf{a}$ است.

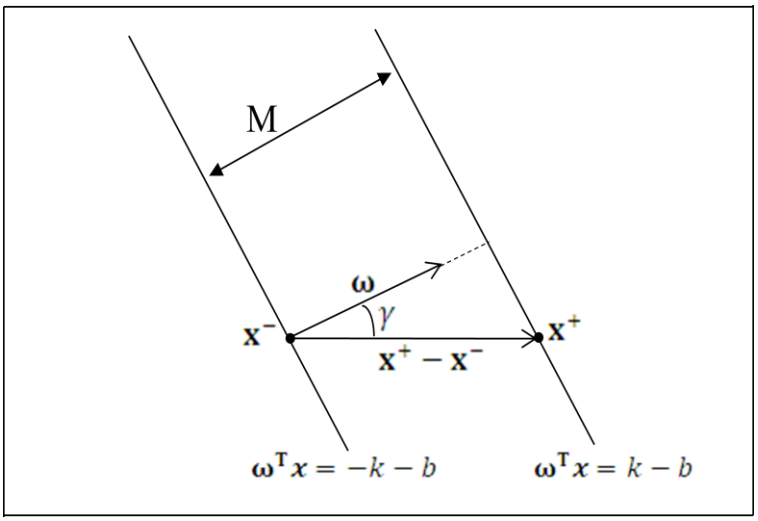


شکل (۲-۱۱)- تصویر بردار $\mathbf{a}$ روی بردار $\mathbf{b}$

در واقع $P_{\mathbf{a}}$ مبین آن است که چه مقدار از $\mathbf{a}$ همسو در راستای $\mathbf{b}$ است (بیژن‌زاده و همکاران، ۱۳۸۹).

حال با استفاده از قضیه تصویر و با توجه به شکل (۲-۱۲) که تلفیقی از شکل (۲-۱۰) و قضیه تصویر است می‌توان مقدار M را به‌دست آورد.

[1] Projection Theorem



شکل (۲-۱۲)- محاسبه‌ی M با استفاده از قضیه تصویر

$$M = \|\boldsymbol{x}^{+} - \boldsymbol{x}^{-}\| cos\gamma \; = \; \frac{\boldsymbol{\omega}\|\boldsymbol{x}^{+} - \boldsymbol{x}^{-}\|}{\|\boldsymbol{\omega}\|} \; = \; \frac{\|\boldsymbol{\omega x}^{+} - \boldsymbol{\omega x}^{-}\|}{\|\boldsymbol{\omega}\|}$$

$$= \frac{\|k-b-(-k-b)\|}{\|\boldsymbol{\omega}\|} \; = \; \frac{2k}{\|\boldsymbol{\omega}\|} \qquad (۲-۱۳)$$

در رابطه‌ی (۲-۱۳)، فاصله‌ی هر نقطه در صفحه، مانند x^{*} تا خط $\boldsymbol{\omega x} + b = 0$ برابر است با $\frac{\left|\boldsymbol{\omega} x^{*} + b\right|}{\|\boldsymbol{\omega}\|}$.

در نتیجه با توجه به ثابت بودن مقدار k، می‌توان گفت

$$\max_{\boldsymbol{\omega},\mathrm{b}}(M) \equiv \max_{\boldsymbol{\omega},\mathrm{b}}\left(\frac{2k}{\|\boldsymbol{\omega}\|}\right) \equiv \min_{\boldsymbol{\omega},\mathrm{b}}\|\boldsymbol{\omega}\|. \qquad (۲-۱۴)$$

لذا برای به‌دست آوردن پارامترهای معادله‌ی خط جدا کننده‌ی دو رده $(\boldsymbol{\omega}, b)$، بایستی مساله‌ی کمینه-سازی (۲-۱۴) مورد بررسی قرار گیرد.

با تقسیم کردن معادلات خطوط مرزی A و C بر k، این معادلات به‌صورت $\boldsymbol{\omega}^{T}.\mathbf{x}+b=1$ و $\boldsymbol{\omega}^{T}.\mathbf{x}+b=-1$ درخواهد آمد (توجه شود که $\boldsymbol{\omega}, b$ پارامترهای جدید می‌باشند). حال، با توجه به شکل (۲-۱۰)، می‌توان گفت

$$\begin{cases}(\boldsymbol{\omega x}_i + b) \geq +1, & \boldsymbol{x}_i \in \text{class}(+1), \\ (\boldsymbol{\omega x}_i + b) \leq -1, & \boldsymbol{x}_i \in \text{class}(-1). \end{cases} \quad (۱۵-۲)$$

لذا مساله‌ی بهینه‌سازی (۲-۱۴) باید تحت شرایط (۲-۱۵) حل گردد. از آن‌جا که کمینه کردن عبارت $\|\boldsymbol{\omega}\|$ هم‌ارز با کمینه کردن عبارت $\frac{1}{2}\|\boldsymbol{\omega}\|^2$ است، می‌توان این مساله‌ی بهینه‌سازی را به‌صورت زیر بازنویسی نمود.

$$\min_{\boldsymbol{\omega},b} \ \frac{1}{2}\|\boldsymbol{\omega}\|^2 ,$$

$$\text{Subject to} \qquad y_i(\boldsymbol{\omega}^T \mathbf{x}_i + b) \geq +1, \qquad i = 1,\ldots,n. \quad (۱۶-۲)$$

مساله‌ی فوق یک مساله‌ی بهینه‌سازی غیرخطی است که شرایط آن، مجموعه‌ای از نامساوی‌های خطی می‌باشد. برای حل این مساله‌ی بهینه‌سازی، می‌توان روش لاگرانژ را به‌کار برد.

$$, \qquad i = 1,\ldots,n, \quad \alpha_i \geq 0 , L = \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + \sum_{i=1}^{n}\alpha_i(1 - y_i(\boldsymbol{\omega}^T\mathbf{x}_i + b)) \quad (۱۷-۲)$$

که در آن، α_iها ضرایب لاگرانژ هستند. با مشتق گیری از L نسبت به $\boldsymbol{\omega}$ و b، داریم.

$$\frac{\partial L}{\partial \boldsymbol{\omega}} = \boldsymbol{\omega} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i = 0 \Rightarrow \boldsymbol{\omega} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i , \quad (۱۸-۲)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n}\alpha_i y_i = 0. \quad (۱۹-۲)$$

با جایگذاری روابط (۲-۱۸) و (۲-۱۹) در L، رابطه-

ی زیر حاصل می‌شود.

$$L(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{n}\alpha_i y_i {\mathbf{x}_i}^T \sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j + \sum_{i=1}^{n}\alpha_i(1 - y_i(\sum_{j=1}^{n}\alpha_j y_j {\mathbf{x}_j}^T \mathbf{x}_i + b))$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j {\mathbf{x}_i}^T \mathbf{x}_j + \sum_{i=1}^{n}\alpha_i , \quad (۲۰-۲)$$

که در آن، $\boldsymbol{\alpha}$ بردار ضرایب لاگرانژ $(\alpha_1, \dots, \alpha_n)$ است.

جواب بهینه‌ی این مساله‌ی بهینه‌سازی بایستی در شرایط KKT[1] (پیوست **ب**) صدق کند که این شرایط برای این مساله به‌صورت زیر است (تیودوریدیس[2] و کوترومباس[3]، ۲۰۰۶).

$$\frac{\partial}{\partial \boldsymbol{\omega}} L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) = 0,$$

$$\frac{\partial}{\partial b} L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) = 0,$$

$$\alpha_i \geq 0, \qquad i = 1, \dots, n,$$

$$\alpha_i [y_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + b) - 1] = 0, \qquad i = 1, \dots, n. \tag{۲۱-۲}$$

مساله‌ی بهینه‌سازی (۲-۲۰) تحت شرایط (۲-۲۱)، و با استفاده از مساله‌ی بهینه‌سازی دوگان کمینه-بیشینه[4] (پیوست **ج**) به یک مساله‌ی بیشینه‌سازی تبدیل می‌گردد. در این مساله، α_iها مجهول هستند که با استفاده از روش‌های عددی به‌دست خواهند آمد. اما از آن‌جایی که به ازای هر مشاهده، یک α_i وجود دارد، در عمل، حل عددی این معادله بسیار پیچیده خواهد بود. اگر α_iهای بهینه با α_i^* نشان داده شود، در این‌صورت با جایگذاری آن در رابطه‌ی (۲-۱۸) برآورد پارامتر $\boldsymbol{\omega}$ به‌دست خواهد آمد.

$$\hat{\boldsymbol{\omega}} = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i. \tag{۲۲-۲}$$

با قرار دادن مقدار $\hat{\boldsymbol{\omega}}$ در معادله‌ی $y_i = \boldsymbol{\omega}^T \mathbf{x}_i + b$، به‌ازای هر نقطه‌ی SV، برای پارامتر b، یک برآورد به‌دست می‌آید. لذا یک برآورد مناسب برای پارامتر b برابر است با

$$\hat{b} = \frac{1}{N_{SV}} \sum_{i \in SV} (y_i - \boldsymbol{\omega}^T x_i), \tag{۲۳-۲}$$

---

1 Karush-Kuhn-Tucker Conditions
2 Theodoridis S.
3 Koutroumbas K.
4 Min-Max Duality

که در آن، N_{SV} تعداد نقاط بردار پشتیبان (SV) می‌باشد.

همان‌طور که پیش از این گفته شد، خطوط جداکننده‌ی مشاهدات بر اساس نقاط بردار پشتیبان شکل می‌گیرند. باتوجه به این‌که به ازای هر مشاهده یک α_i^* وجود دارد، می‌توان گفت مشاهده‌ی i-ام یک SV است اگر و فقط اگر $\alpha_i^* \neq 0$ (گان[1]، ۱۹۹۸؛ هستی و همکاران، ۲۰۰۹؛ تیودوریدیس و کوترومباس، ۲۰۰۶).

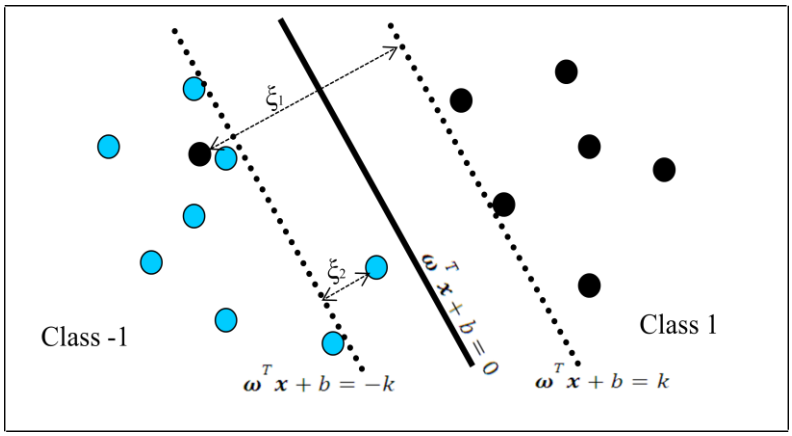
۲-۳-۳ رده‌بندی خطی داده‌های تفکیک‌ناپذیر دو رده‌ای

در زیربخش قبل فرض شد که داده‌ها به‌صورت خطی تفکیک‌پذیر هستند که در عمل به‌ندرت چنین است. همانند قبل، فرض کنید $(\mathbf{X}_i, Y_i)$; $i = 1,\ldots,n$ مجموعه داده‌های مدل‌ساز باشد، که در آن $\mathbf{X}_i = (X_{i1}, X_{i2})$ برداری از متغیرهای توضیحی و $y_i \in \{-1,1\}$ است. همان‌طور که در شکل (۲-۱۳) دیده می‌شود، یکی از داده‌های رده‌ی (1) در رده‌ی (-1) قرار گرفته و یکی از داده‌های رده‌ی (-1) خارج از ناحیه‌ی امن رده‌ی خود قرار گرفته است. در واقع داده‌ی دوم در نوار بین دو رده قرار دارد. در این نوع مجموعه داده‌های تفکیک‌ناپذیر، یک قید دیگر به مساله‌ی بهینه‌سازی اضافه می‌گردد. برای بیان این قید، بایستی ابتدا متغیر کمکی[2] ξ_i را به‌صورت زیر تعریف کرد.

$$\xi_i = \begin{cases} 0, & y_i(\boldsymbol{\omega}^T \mathbf{x}_i + b) \geq 1, \\ d_i, & y_i(\boldsymbol{\omega}^T \mathbf{x}_i + b) < 1, \end{cases} \qquad (۲۴-۲)$$

که در آن، d_i فاصله‌ی مشاهده‌ی i-ام از مرز رده‌ی مربوط به همان مشاهده است. همچنین مقادیر ξ_i برای مشاهداتی که به‌درستی رده‌بندی شده‌اند، صفر است، بدین ترتیب همواره می‌توان گفت $\xi_i \geq 0$.

[1] Gunn S.R
[2] Slack



شکل (۲-۱۳)- نمایش متغیر کمکی در داده‌های تفکیک‌ناپذیر دو رده‌ای

با این توضیحات، مساله‌ی بهینه‌سازی رابطه‌ی (۲-۱۶) به‌صورت زیر خواهد بود.

$$\min_{\boldsymbol{\omega},b} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}\xi_i \ ,$$

$$\text{Subject to} \quad y_i(\boldsymbol{\omega}^T\mathbf{x}_i + b) \geq +1-\xi_i, \qquad \boldsymbol{i = 1,\ldots, n.} \tag{۲-۲۵}$$

که در آن، C پارامتری در دست کاربر، برای تنظیم ارتباط بین خطای برآورد و حاشیه امنیت (M) است. در واقع این پارامتر مشخص می‌کند که برای کاربر، کمینه کردن خطا مهمتر است یا بیشینه کردن حاشیه امنیت. رابطه‌ی (۲-۲۵)، مشابه زیربخش۲-۳-۲، با استفاده از معادلات لاگرانژ، شرایط KKT و مساله‌ی بهینه‌سازی دوگان کمینه-بیشینه به‌صورت زیر در می‌آید.

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i{}^T\boldsymbol{x}_j \ ,$$

$$\text{Subject to} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \qquad C \geq \alpha_i \geq 0. \tag{۲-۲۶}$$

توجه شود که تنها تفاوت رابطه‌ی فوق با رابطه‌ی (۲-۲۱)، در داده‌های تفکیک‌پذیر، وجود یک حد بالا (یعنی C) برای ضرایب لاگرانژ (α_i) است.اگر α_iهای بهینه با α_i^* نشان داده شود، در این‌صورت

$$\hat{\boldsymbol{\omega}} = \sum_{i=1}^{n} \alpha_i^{*} y_i \mathbf{x}_i, \qquad (۲۷-۲)$$

$$\hat{b} = \frac{1}{N_{SV}} \sum_{i \in SV} (y_i - \boldsymbol{\omega}^T \mathbf{x}_i). \qquad (۲۸-۲)$$

با توجه به مطالب فوق، رده‌بندی مشاهده‌ی جدید $\mathbf{x}_0$ به‌صورت زیر انجام می‌گیرد (گان، ۱۹۹۸؛ هستی و همکاران، ۲۰۰۹).

$$if \;\; f(\boldsymbol{x}_0) \geq 0 \quad then \quad \boldsymbol{x}_0 \in class\,(+1), \qquad (۲۹-۲)$$

$$if \;\; f(\boldsymbol{x}_0) < 0 \quad then \quad \boldsymbol{x}_0 \in class\,(-1),$$

که در آن

$$f(\mathbf{x}_0) = \boldsymbol{\omega}^T \mathbf{x}_0 + b = \sum_{i=1}^{n} \alpha_i^{*} y_i \mathbf{x}_i^T \mathbf{x}_0 + b \;. \qquad (۳۰-۲)$$

مثال۱

یک مجموعه داده با ۵ مشاهده‌ی {(1,1),(2,1),(4,-1),(5,-1),(6,1)} را در نظر بگیرید، که در آن C=10 معلوم باشد. ابتدا برای به‌دست آوردن معادله منحنی جداکننده‌ی دو رده، با استفاده از رابطه‌ی (۲-۲۶)، مقادیر α_i به کمک روش‌های عددی بهینه‌سازی محاسبه می‌گردد.

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{5} \alpha_i - \frac{1}{2} \sum_{i=1}^{5} \sum_{j=1}^{5} \alpha_i \alpha_j y_i y_j x_i x_j \,,$$

$$\text{Subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad 10 \geq \alpha_i \geq 0.$$

با استفاده از نرم‌افزار بهینه‌سازی Lingo مقادیر α_i عبارتند از

α_1=7.467, α_2=0, α_3=10, α_4=7.467, α_5=10 .

در نتیجه $\{x_1=1, x_3=4, x_4=5, x_5=6\}$ نقاط SV می‌باشد که با جایگذاری در رابطه‌ی (۲-۲۷) برآورد $\boldsymbol{\omega}$

$$\hat{\boldsymbol{\omega}} = (7.467, 0, -40, -37.335, 60)$$

به‌صورت زیر به‌دست می‌آید.

با جایگذاری مقادیر $\hat{\omega}$ در رابطه‌ی $\omega x + b = 0$، معادله‌ی خط جداکننده عبارتست از

$f(x)= -9.869x+b.$

همچنین با توجه به معادله‌ی $f(x)$ و رابطه‌ی (۲-۲۸) برآورد پارامتر b به‌دست می‌آید.

$$f(x_1) = 1 = -9.869 + b_1 \Rightarrow b_1 = 10.869$$
$$f(x_3) = -1 = 39.476 + b_2 \Rightarrow b_2 = 38.476$$
$$f(x_4) = -1 = 49.345 + b_3 \Rightarrow b_3 = 48.345$$
$$f(x_5) = 1 = 59.214 + b_4 \Rightarrow b_4 = 60.214$$

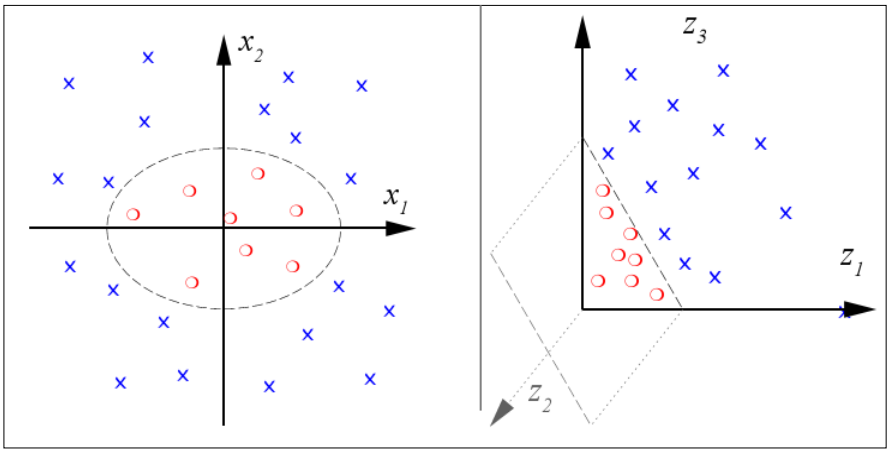
آن‌گاه $\hat{b} = \frac{1}{4}\sum_{i=1}^{4} b_i = 39.476$

بدین ترتیب معادله‌ی نهایی خط جداکننده‌ی دو رده، به‌صورت زیر به‌دست می‌آید.

$f(x)= -9.869x+39.476$

۲-۳-۴ رده‌بندی غیرخطی داده‌های دو رده‌ای

در بسیاری از مسایل، داده‌ها به گونه‌ای هستند که بایستی ابتدا با استفاده از یک نگاشت مناسب، آن‌ها را به یک مجموعه داده‌ی جدید تبدیل نمود و سپس عمل رده‌بندی را بر روی داده‌های جدید انجام داد. برای مثال، قاب چپ شکل (۲-۱۴) را در نظر بگیرید که در آن رده‌بندی داده‌ها با استفاده از یک خط توسط روش SVM امکان‌پذیر نیست. اگر این داده‌ها طی یک نگاشت مناسب به‌صورت قاب راست شکل (۲-۱۴) تبدیل گردند آن‌گاه توسط یک صفحه می‌توان این داده‌ها را رده‌بندی کرد. با انجام این نگاشت روی داده‌ها، داده‌های دو بعدی (قاب چپ) به داده‌های سه‌بعدی (قاب راست) تبدیل شده است.



شکل (۲-۱۴)- نگاشت داده‌ها از فضای درجه دو (قاب چپ) به فضای درجه سه (قاب راست)

۲-۳-۴-۱ تابع هسته

همان‌طور که اشاره شد، روش SVM رده‌بندی را اغلب بر روی تبدیلی از داده‌های اولیه انجام می‌دهد. در این روش، توابعی با نام *توابع هسته* تعریف می‌گردند که این توابع مانند یک نگاشت، بر روی داده‌ها اثر کرده و مجموعه داده‌های جدیدی را تولید می‌کنند. تابع هسته K را می‌توان به‌صورت زیر تعریف کرد.

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j), \qquad (۳۱-۲)$$

که در آن تابع $\phi(.)$ به عنوان یک تابع تبدیل یا نگاشت به روی داده‌ها اثر می‌کند. شایان ذکر است که K نیمه معین مثبت[1] و همچنین متقارن می‌باشد.

مثال۲

بردار $\mathbf{x} = (x_1, x_2)$ و تابع هسته $K(x_i, x_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^2$ را در نظر بگیرید. برای به‌دست آوردن تابع $\phi(.)$ باید عملیات زیر را پیش گرفت.

$$K(x_i, x_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^2$$

[1] Positive Semidefinite

$$= 1 + x_{i1}{}^2 x_{j1}{}^2 + 2x_{i1}x_{j1}x_{i2}x_{i2} + x_{i2}{}^2 x_{j2}{}^2 + 2x_{i1}x_{j1} + 2x_{i2}x_{j2}$$

$$= (1, x_{i1}{}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}{}^2, \sqrt{2}x_{i1}, \sqrt{2}x_{i2})^T . (1, x_{j1}{}^2, \sqrt{2}x_{j1}x_{i2}, x_{j2}{}^2, \sqrt{2}x_{j1}, \sqrt{2}x_{j2})$$

$$= \phi(x_i)^T \phi(x_j),$$

که در آن $\phi(x) = (1, x_1{}^2, \sqrt{2}x_1x_2, x_2{}^2, \sqrt{2}x_1, \sqrt{2}x_2)$.

۲-۳-۴-۲ برآورد پارامتر مدل‌های غیرخطی

در صورت استفاده از توابع هسته در روش SVM، مانند زیربخش ۲-۳-۳، برآورد پارامترهای مدل، با استفاده از حل مسایل بهینه‌سازی به‌دست می‌آید. بدین ترتیب معادله‌ی بهینه‌سازی نهایی به‌صورت زیر در می‌آید.

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) ,$$

$$\text{Subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad C \geq \alpha_i \geq 0. \tag{۳۲-۲}$$

که در آن، C پارامتری برای تنظیم ارتباط بین خطای برآورد و حاشیه امنیت (M) است.

لذا برآورد پارامترهای مدل عبارتند از

$$\hat{\boldsymbol{\omega}} = \sum_{i=1}^{n} \alpha_i{}^* y_i \phi(\mathbf{x}_i), \tag{۳۳-۲}$$

$$\hat{b} = \frac{1}{N_{SV}} \sum_{i \in SV} (y_i - \boldsymbol{\omega}^T \phi(\mathbf{x}_i)), \tag{۳۴-۲}$$

که در آن، α_i^* نشان‌دهنده‌ی α_iهای بهینه و N_{SV} تعداد نقاط SV هستند. با توجه به مطالب فوق، معادله‌ی خط (صفحه) جداکننده به‌صورت زیر خواهد بود.

$$f(x) = \boldsymbol{\omega}^T \phi(x) + b = \sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i^T, x) + b. \quad (۳۵-۲)$$

۲-۳-۴-۳ انواع توابع هسته

در مسایل مختلف، برحسب نیاز، توابع هسته‌ی متعددی مورد استفاده قرار می‌گیرد. چهار تابع هسته که کاربرد بیشتری در عمل دارند، عبارتند از

$$K(x_i, x_j) = x_i^T x_j \quad \text{هسته‌ی خطی} \quad (۳۶-۲)$$

$$K(x_i, x_j) = (\gamma x_i^T x_j + Coef0)^d \quad \text{هسته‌ی چند جمله‌ای از مرتبه‌ی } d \quad (۳۷-۲)$$

$$K(x_i, x_j) = \exp\{-\gamma \|x_i - x_j\|^2\} \quad \text{هسته‌ی نرمال}^1 \quad (۳۸-۲)$$

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + Coef0) \quad \text{هسته‌ی سیگوید}^2 \quad (۳۹-۲)$$

که در آن، Coef0، d و γ پارامترهای قابل تنظیم توسط کاربر هستند.

شایان ذکر است که در واقع هسته خطی منجر به نتایج مدل خطی، رابطه (۲-۳۰)، می‌گردد (گان، ۱۹۹۸؛ هستی و همکاران، ۲۰۰۹).

مثال ۳

مجموعه داده‌های مثال۱ را در نظر بگیرید. فرض کنید C=100 و هسته‌ی چندجمله‌ای درجه ۲ مورد نظر باشد. همچنین فرض کنید $Coef0 = 0$ و $\gamma = 1$. ابتدا مقادیر α_i به کمک روش‌های عددی بهینه‌سازی محاسبه می‌گردد.

1 Radial Basis Function Kernel
2 Sigmoid Kernel

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{5} \alpha_i - \frac{1}{2}\sum_{i=1}^{5}\sum_{j=1}^{5} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i \boldsymbol{x}_j + 1)^2 ,$$

Subject to $\sum_{i=1}^{5} \alpha_i y_i = 0, \qquad 100 \geq \alpha_i \geq 0.$

با استفاده از نرم‌افزار بهینه‌سازی Lingo، مقادیر بهینه‌ی α_i عبارتند از

$\alpha_1=0,\ \alpha_2=2.5,\ \alpha_3=0,\ \alpha_4=7.333,\ \alpha_5=4.833$.

در نتیجه $\{x_2=2, x_4=5, x_5=6\}$ نقاط SV می‌باشد.

با استفاده از رابطه‌ی (۲-۳۵)، معادله‌ی منحنی جداکننده‌ی دو رده به‌صورت زیر به‌دست می‌آید.

$$f(x) = 2.5(1)(2x+1)^2 + 7.333(-1)(5x+1)^2 + 4.833(1)(6x+1)^2 + b$$

$$= 0.663x^2 - 5.334x + b$$

با توجه به رابطه (۲-۳۴) و با قرار دادن نقاط $\{x_2=2, x_4=5, x_5=6\}$ در معادله‌ی فوق، $\hat{b}$ به‌دست می‌آید.

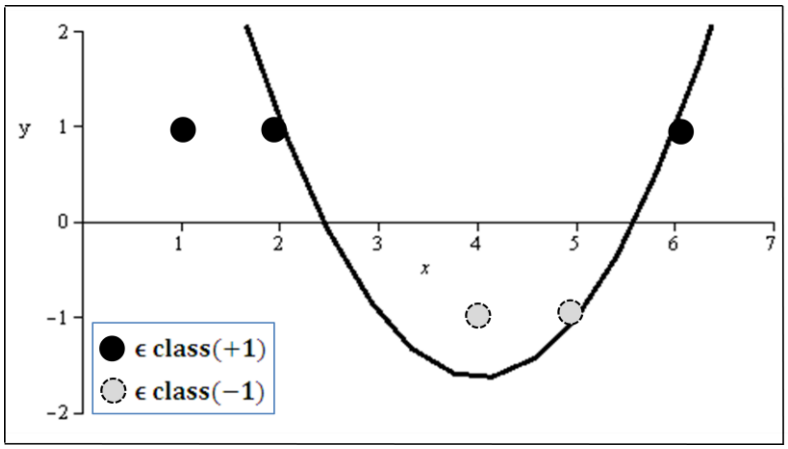
$$f(x_2) = +1 = -8.016 + b_1 \Rightarrow b_1 = 9.016$$
$$f(x_4) = -1 = -10.095 + b_2 \Rightarrow b_2 = 9.095$$
$$f(x_5) = +1 = -8.136 + b_3 \Rightarrow b_3 = 9.136$$

آن‌گاه $\hat{b} = \frac{1}{3}\sum_{i=1}^{4} b_i = 9.08$

لذا معادله‌ی نهایی منحنی جداکننده به‌صورت زیر است.

$$f(x) = 0.663x^2 - 5.334x + 9.08$$

و در نهایت، رده‌بندی داده‌ها به‌صورتی که در شکل (۲-۱۵) نشان داده شده است، خواهد بود.



شکل (۲-۱۵)- رده‌بندی داده‌های یک متغیره به روش SVM با استفاده از هسته چندجمله‌ای درجه ۲

۲-۳-۵ رگرسیون خطی ماشین بردار پشتیبان

مدل کلی رگرسیونی عبارت است از

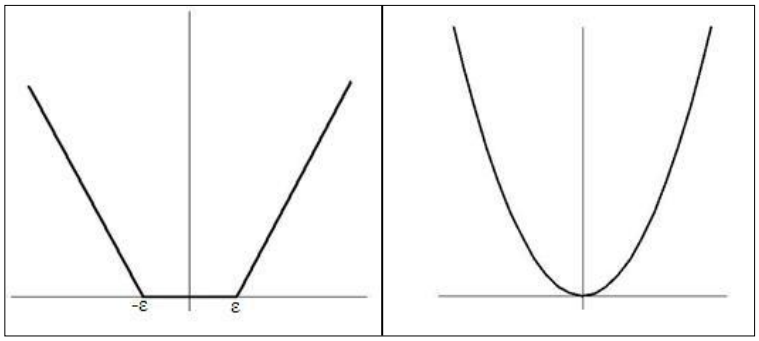
$$f(x) = \boldsymbol{\omega}^T \mathbf{x} + b, \qquad (۴۰-۲)$$

که در آن، پارامتر b، عرض از مبدأ و پارامتر $\boldsymbol{\omega}$، ضریب $\mathbf{x}$ است که هر دو نامعلوم‌اند و برای برآورد این دو پارامتر از روش کمترین توان‌های دوم خطا استفاده می‌شود. در واقع در رگرسیون خطی برای برآورد پارامترها از تابع زیان درجه دوم $\sum_{i=1}^{n}(y_i - f(\boldsymbol{x_i}))^2$ استفاده می‌شود.

فرض کنید n مشاهده با مشخصات $\{(\mathbf{x}_i, y_i); i = 1,..,n\}$ در اختیار باشد. در رگرسیون ماشین بردار پشتیبان برای برآورد پارامترهای مدل، بایستی تابع $\sum_{i=1}^{n} V_\varepsilon(y_i - f(\mathbf{x}_i))$ کمینه گردد که در آن V_ε تابع زیان $\varepsilon - insensitive$ با ضابطه‌ی زیر است.

$$V_\varepsilon(\mathrm{r_i}) = \begin{cases} 0, & |\mathrm{r_i}| < \varepsilon, \\ |\mathrm{r_i}| - \varepsilon, & \mathrm{o.w}, \end{cases} \qquad (۴۱-۲)$$

که در آن، $r_i = y_i - f(\mathbf{x}_i)$ و $\varepsilon \geq 0$ پارامتری است که توسط کاربر تعیین می‌گردد. در شکل (۲-۱۶) می‌توان تفاوت‌های تابع زیان درجه دوم و تابع زیان $\varepsilon - insensitive$ را مشاهده نمود.



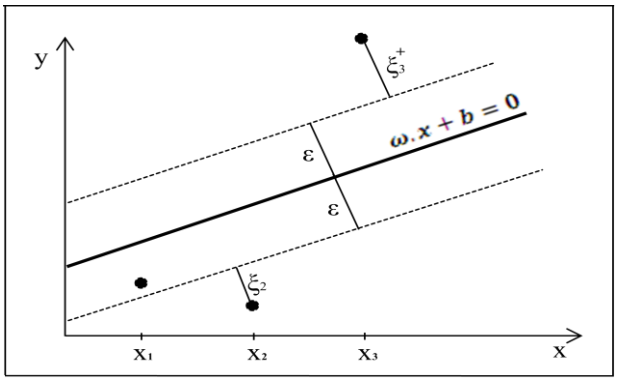
شکل (۲-۱۶)- تابع زیان درجه دوم (قاب راست) و تابع زیان $\varepsilon - insensitive$ (قاب چپ)

فرض کنید معادله‌ی روش SVM برازش‌داده‌شده به‌صورت رابطه‌ی (۲-۴۰) باشد. به‌منظور برآورد پارامترهای مدل، مشابه مدل رده‌بندی، در رگرسیون نیز باید به دنبال کمینه کردن عبارت $\frac{1}{2}\|\omega\|^2$ بود. پیش از آنکه به مساله‌ی بهینه‌سازی پرداخته شود، نیاز است به متغیر کمکی و نقش آن در رگرسیون ماشین بردار پشتیبان اشاره نمود.

شکل (۲-۱۷) را که در آن سه مشاهده‌ی $\{(x_1,y_1),(x_2,y_2),(x_3,y_3)\}$ نشان داده شده‌اند، در نظر بگیرید. فرض کنید $f(x)=\omega x+b$، معادله‌ی خط برازش‌داده‌شده برای y باشد و $\varepsilon \geq 0$ پارامتر موجود در تابع زیان $\varepsilon - insensitive$ در اختیار کاربر باشد. در این‌صورت برای هر یک از مشاهدات یکی از سه وضعیت زیر رخ خواهد داد.

$$\begin{cases} \xi_i = |y_i - f(x_i)| \quad \text{if} \quad y_i - f(x_i) > \varepsilon \\ \text{if} \quad f(x_i) - y_i > \varepsilon \qquad\qquad \xi_i^+ = |y_i - f(x_i)| \\ \xi_i^+, \xi_i = 0 \qquad \text{if} \quad |y_i - f(x_i)| \leq \varepsilon \end{cases} \qquad (۲-۴۲)$$

که در آن، با توجه به وضعیت مشاهده‌ی x_i ، مقادیر ξ_i یا ξ_i^+ به عنوان متغیر کمکی مربوط به مشاهده‌ی i-ام تعریف می‌گردد.



شکل(۲-۱۷)- متغیر کمکی در رگرسیون ماشین بردار پشتیبان

با توجه به شکل(۲-۱۷) در واقع می‌توان گفت که خط $f(x)=\omega x+b$ به‌وسیله‌ی نواری[1] به پهنای 2ε احاطه شده است. اگر مشاهده‌ی x_i داخل این نوار قرار بگیرد، متغیر کمکی مربوط به آن صفر خواهد بود و در غیر این‌صورت با توجه به رابطه‌ی (۲-۴۲)، یکی از مقادیر ξ_i یا ξ_i^+ به عنوان متغیر کمکی مشاهده‌ی x_i تعریف می‌شود.

با توجه به تعاریف فوق، مشابه رابطه‌ی (۲-۲۵)، به‌منظور برآورد پارامترهای مدل، بایستی مساله‌ی بهینه‌سازی زیر حل شود. در واقع تفاوت مساله‌ی بهینه‌سازی زیر با رابطه‌ی (۲-۲۵) در قیدهای مساله‌ی بهینه‌سازی است.

$$\min_{\boldsymbol{\omega},b}\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^+)\ , \qquad (۴۳-۲)$$

$$\text{Subject to}\quad y_i\geq\boldsymbol{\omega}^T\mathbf{x}_i+b-\varepsilon-\xi_i\ ,\quad \forall i=1,\ldots,n\ ,$$

$$y_i\leq\boldsymbol{\omega}^T\mathbf{x}_i+b+\varepsilon+\xi_i^+\ ,\quad \forall i=1,\ldots,n\ ,$$

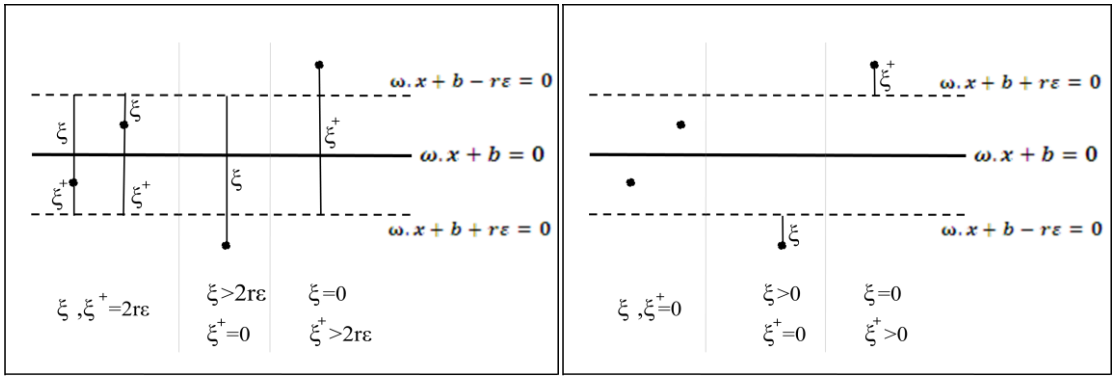
$$\xi_i,\xi_i^+\geq 0\ ,\quad \forall i=1,\ldots,n\ .$$

دو شرط ابتدایی نشان می‌دهد که پهنای نوار(2ε) باید طوری تنظیم شود تا بیشترین تعداد از y_i ها در این محدوده قرار گیرد. در واقع می‌توان گفت این نوار همانند حاشیه امنیت در حالت رده‌بندی داده-

[1] Tube

های کیفی عمل می‌کند. هرچند که افزایش پهنای این نوار موجب می‌گردد که تعداد بیشتری در این محدوده قرار بگیرد ولی این امر نیازمند افزایش خطا (ε) می‌باشد. لذا پارامتر C برای تنظیم ارتباط بین خطای برآورد و پهنای نوار تعریف می‌گردد. در واقع پارامتر C توازنی بین کمینه‌سازی خطا و بیشینه کردن پهنای نوار برقرار می‌کند.

اگر متغیر کمکی، مطابق رابطه‌ی (۲-۴۲) و شکل (۲-۱۷) تعریف گردد متغیر r_i تعریف شده در ساختار تابع خطای $\varepsilon - insensitive$، تاثیری در مساله‌ی بهینه‌سازی رابطه‌ی (۲-۴۳) نخواهد داشت. لذا بایستی متغیر کمکی مطابق شکل (۲-۱۸) تعریف گردد.



شکل (۲-۱۸)- نمایش هندسی متغیر کمکی به ازای r>0 (قاب راست) و r<0 (قاب چپ)

در این‌صورت مساله‌ی بهینه‌سازی به‌صورت زیر در خواهد آمد.

$$\min_{\boldsymbol{\omega},b} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^+) \ , \qquad (۴۴-۲)$$

$$\text{Subject to} \quad y_i \geq \boldsymbol{\omega}^T\mathbf{x}_i + b - r_i\varepsilon - \xi_i \ , \quad \forall i = 1,\ldots,n \ ,$$

$$y_i \leq \boldsymbol{\omega}^T\mathbf{x}_i + b + r_i\varepsilon + \xi_i^+ \ , \quad \forall i = 1,\ldots,n \ ,$$

$$\xi_i + \xi_i^+ \geq h_i \ , \quad \forall i = 1,\ldots,n \ ,$$

$$\xi_i, \xi_i^+ \geq 0 \ , \quad \forall i = 1,\ldots,n \ .$$

که در آن، h_i به‌صورت زیر تعریف می‌شود.

$$h_i = \begin{cases} 0, & r_i \geq 0, \\ \rho - 2\varepsilon r_i, & r_i < 0, \end{cases} \quad (۴۵-۲)$$

که در آن، ρ مقدار ثابتی است که $\rho > 0$ و $\rho \ll min_i|r_i\varepsilon|$.

بدین ترتیب برای برآورد پارامترهای مدل، نیاز به حل مساله‌ی بهینه‌سازی (۲-۴۴) می‌باشد. با استفاده از روش لاگرانژ، رابطه‌ی (۲-۴۴) به‌صورت زیر در می‌آید.

$$L = \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^+) - \sum_{i=1}^{n}\alpha_i(y_i - \boldsymbol{\omega}^T\mathbf{x}_i - b + r_i\varepsilon + \xi_i) \quad (۴۶-۲)$$

$$-\sum_{i=1}^{n}\alpha_i^*(\boldsymbol{\omega}^T\mathbf{x}_i + b - y_i + r_i\varepsilon + \xi_i^+) - \sum_{i=1}^{n}\beta_i\xi_i - \sum_{i=1}^{n}\beta_i^*\xi_i^+ - \sum_{i=1}^{n}\gamma_i(\xi_i + \xi_i^+ - h_i),$$

که در آن، $\alpha_i, \alpha_i^*, \beta_i, \beta_i^*, \gamma_i$ ضرایب لاگرانژ می‌باشد.

با انجام عملیات ریاضی، می‌توان رابطه‌ی (۲-۴۶) را به‌صورت زیر بازنویسی کرد.

$$L = \frac{1}{2}\|\boldsymbol{\omega}\|^2 + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\boldsymbol{\omega}\mathbf{x}_i - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i + b\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) \quad (۴۷-۲)$$

$$-\varepsilon.\sum_{i=1}^{n}r_i(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}(C - \alpha_i - \beta_i - \gamma_i)\xi_i + \sum_{i=1}^{n}(C - \alpha_i^* - \beta_i^* - \gamma_i)\xi_i^+ + \sum_{i=1}^{n}\gamma_i h_i .$$

با مشتق گرفتن از L نسبت به $\boldsymbol{\omega}$، b، ξ_i و ξ_i^+ ، روابط (۲-۴۸) الی (۲-۵۱) حاصل می‌گردد.

$$\frac{\partial L}{\partial \boldsymbol{\omega}} = \boldsymbol{\omega} + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}_i = 0 \ , \quad (۴۸-۲)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \ , \quad (۴۹-۲)$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i - \gamma_i = 0 \ , \qquad (۵۰-۲)$$

$$\frac{\partial L}{\partial \xi_i^{+}} = C - \alpha_i^{*} - \beta_i^{*} - \gamma_i = 0 \ . \qquad (۵۱-۲)$$

با جایگذاری روابط (۲-۴۸) الی (۲-۵۱) در رابطه‌ی (۲-۴۷)، می‌توان مساله را به‌صورت زیر بازنویسی کرد.

$$L = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^{*})(\alpha_j - \alpha_j^{*})\mathbf{x}_i^{T}\mathbf{x}_j - \sum_{i=1}^{n}(\alpha_i - \alpha_i^{*})y_i - \varepsilon\sum_{i=1}^{n}r_i(\alpha_i + \alpha_i^{*}) + \sum_{i=1}^{n}\gamma_i h_i . \qquad (۵۲-۲)$$

مشابه بخش۲-۳-۲، معادله‌ی کمینه‌سازی (۲-۴۴) با استفاده از مساله‌ی بهینه‌سازی دوگان کمینه-بیشینه به صورت معادله‌ی بیشینه‌سازی تحت شرایط زیر در می‌آید.

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^{*}) = 0 \qquad (۵۳-۲)$$

$$\forall i = 1,\ldots,n \quad \alpha_i, \alpha_i^{*}, \gamma_i \geq 0 \qquad (۵۴-۲)$$

در رابطه‌ی (۲-۵۲) به غیر از پارامترهای $\alpha_i, \alpha_i^{*}, \gamma_i$ که مجهول‌اند، مابقی پارامترها معلوم می‌باشد. اگر بتوان این رابطه را تنها به پارامترهای α_i, α_i^{*} وابسته کرد و پارامتر γ_i را حذف کرد، می‌توان با استفاده از روش‌های عددی پیشرفته همانند رابطه‌ی (۲-۲۱)، مساله را حل کرد. به همین منظور با توجه به روابط (۲-۵۰) و (۲-۵۱) و این‌که $\gamma_i \geq 0$، می‌توان گفت

$$\gamma_i = \max_{\gamma \geq 0}\{\gamma \leq C - \alpha_i, \gamma \leq C - \alpha_i^{*}\} = \min\{C - \alpha_i, C - \alpha_i^{*}\} = C - \max\{\alpha_i, \alpha_i^{*}\} \ , \qquad (۵۵-۲)$$

بدین ترتیب می‌توان رابطه‌ی زیر را به عنوان یک تقریب مناسب برای L ارایه کرد.

$$L' = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - \varepsilon.\sum_{i=1}^{n}r_i(\alpha_i + \alpha_i^*) \qquad (۵۶-۲)$$

$$+C\sum_{i=1}^{n}h_i - \sum_{i=1}^{n}h_i(\alpha_i + \alpha_i^*),$$

و در نهایت با توجه به این‌که $C\sum_{i=1}^{n}h_i \geq 0$، می‌توان رابطه‌ی زیر را به عنوان تقریب نهایی برای L ارایه کرد.

$$L'' = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - \sum_{i=1}^{n}(r_i\varepsilon + h_i)(\alpha_i + \alpha_i^*) \qquad (۵۷-۲)$$

با توجه به این‌که معادله‌ی فوق بایستی تحت قیدهای (۲-۵۳) و (۲-۵۴) بیشینه گردد و از آن‌جایی که قید $\gamma_i \geq 0$ هم‌ارز با قید $\max\{\alpha_i, \alpha_i^*\} \leq C$ است، می‌توان مساله‌ی بهینه‌سازی را به‌صورت زیر بازنویسی کرد.

$$\max_{\boldsymbol{\alpha}} L'' = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - \sum_{i=1}^{n}(r_i\varepsilon + h_i)(\alpha_i + \alpha_i^*),$$

$$\text{Subject to} \quad \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \qquad (۵۸-۲)$$

$$0 \leq \alpha_i \leq C,$$

$$0 \leq \alpha_i^* \leq C.$$

در معادله‌ی فوق تنها پارامترهای α_i, α_i^* مجهول می‌باشد لذا این بهینه‌سازی با استفاده از روش‌های عددی قابل حل خواهد بود. اما از آن‌جایی که به ازای هر مشاهده یک α_i, α_i^* وجود دارد، در عمل، حل عددی این معادله بسیار پیچیده است. با این وجود اگر α_i, α_i^*های بهینه به‌ترتیب با $\hat{\alpha}_i, \hat{\alpha}_i^*$ نشان داده شود در این‌صورت با جایگذاری در روابط (۲-۴۸) و (۲-۴۰) برآورد پارامترهای مدل به‌دست خواهد آمد.

$$\hat{\boldsymbol{\omega}} = \sum_{i=1}^{n} (\hat{\alpha}_i^{\,*} - \hat{\alpha}_i)\mathbf{x}_i, \qquad (۵۹-۲)$$

$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \boldsymbol{\omega}^T \mathbf{x}_i). \qquad (۵۹-۲)$$

توجه کنید که مطابق بخش‌های ۲-۳-۲ و ۲-۳-۳ بایستی برقراری شرایط KKT را نیز بررسی کرد. شرایط KKT در این مساله عبارتند از

$$\begin{aligned}
&\forall i = 1,\ldots,n;\\
&(y_i - \boldsymbol{\omega}^T \mathbf{x}_i - b + r_i\varepsilon + \xi_i)\alpha_i = 0,\\
&(\boldsymbol{\omega}^T \mathbf{x}_i + b - y_i + r_i\varepsilon + \xi_i^{\,+})\alpha_i^{\,*} = 0,\\
&\beta_i \xi_i = 0,\\
&\beta_i^{\,*} \xi_i^{\,+} = 0,\\
&(\xi_i + \xi_i^{\,+} - h_i)\gamma_i = 0.
\end{aligned} \qquad (۶۰-۲)$$

با استفاده از شرایط KKT می‌توان گفت.

$$y_i = \boldsymbol{\omega}^T \mathbf{x}_i + b - r_i\varepsilon - \xi_i = 0$$

$$r_i\varepsilon + \xi_i = \boldsymbol{\omega}^T \mathbf{x}_i + b - y_i$$

که با جایگذاری در شرط دوم رابطه‌ی(۲-۶۰) خواهیم داشت

$$(2r_i\varepsilon + \xi_i + \xi_i^{\,+})\alpha_i^{\,*} = 0$$

بدین ترتیب اگر $r_i > 0$ و یا $r_i < 0$ آنگاه بایستی همواره $\alpha_i^{\,*} = 0$، مگر این‌که $\xi_i + \xi_i^{\,+} = -2r_i\varepsilon$، که با شرط ذکرشده در رابطه‌ی (۲-۴۴) در تناقض است (آپولونی[1] و همکاران، ۲۰۱۰؛ گان، ۱۹۹۸؛ هستی و همکاران، ۲۰۰۹).

[1] Apolloni B.

۲-۳-۶ رگرسیون غیرخطی ماشین بردار پشتیبان

همان‌طور که در بخش۲-۳-۴ اشاره شد، برای برازش مدل‌های غیرخطی، بایستی از توابع هسته استفاده کرد. در صورت استفاده از توابع هسته در مدل رگرسیونی روش SVM، برآورد پارامترهای مدل مانند بخش قبل با استفاده از مسایل بهینه‌سازی به‌دست می‌آید. معادله‌ی بهینه‌سازی نهایی به‌صورت زیر خواهد بود.

$$\max_{\boldsymbol{\alpha}} L'' = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - \sum_{i=1}^{n}(r_i\varepsilon + h_i)(\alpha_i + \alpha_i^*).$$

$$\text{Subject to} \quad \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \qquad (۶۱-۲)$$

$$0 \le \alpha_i \le C,$$

$$0 \le \alpha_i^* \le C.$$

در این‌صورت برآورد پارامترهای مدل به‌صورت زیر است (آپولونی و همکاران، ۲۰۱۰).

$$\hat{\boldsymbol{\omega}} = \sum_{i=1}^{n}(\hat{\alpha}_i^* - \hat{\alpha}_i)\phi(\mathbf{x}_i), \qquad (۶۲-۲)$$

$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\omega}^T\phi(\mathbf{x}_i)). \qquad (۶۳-۲)$$

۲-۳-۷ مزیت‌های روش ماشین بردار پشتیبان

الف) این روش حتی در صورت کم بودن تعداد داده‌های آزمایشی، عملکرد مناسبی دارد. نمونه‌ی بارز این ویژگی را می‌توان در مثال ۳ دید.

ب) روش SVM بهترین مدل را در کل فضای مشاهدات پیدا می‌کند و هرگز به‌صورت ناحیه‌ای[1] عمل نمی‌کند.

پ) با توجه به وجود هسته‌های مختلف، روش SVM قابلیت کار کردن در فضای مشاهدات با ابعاد زیاد را دارا می‌باشد.

[1] Locally

فصل سوم

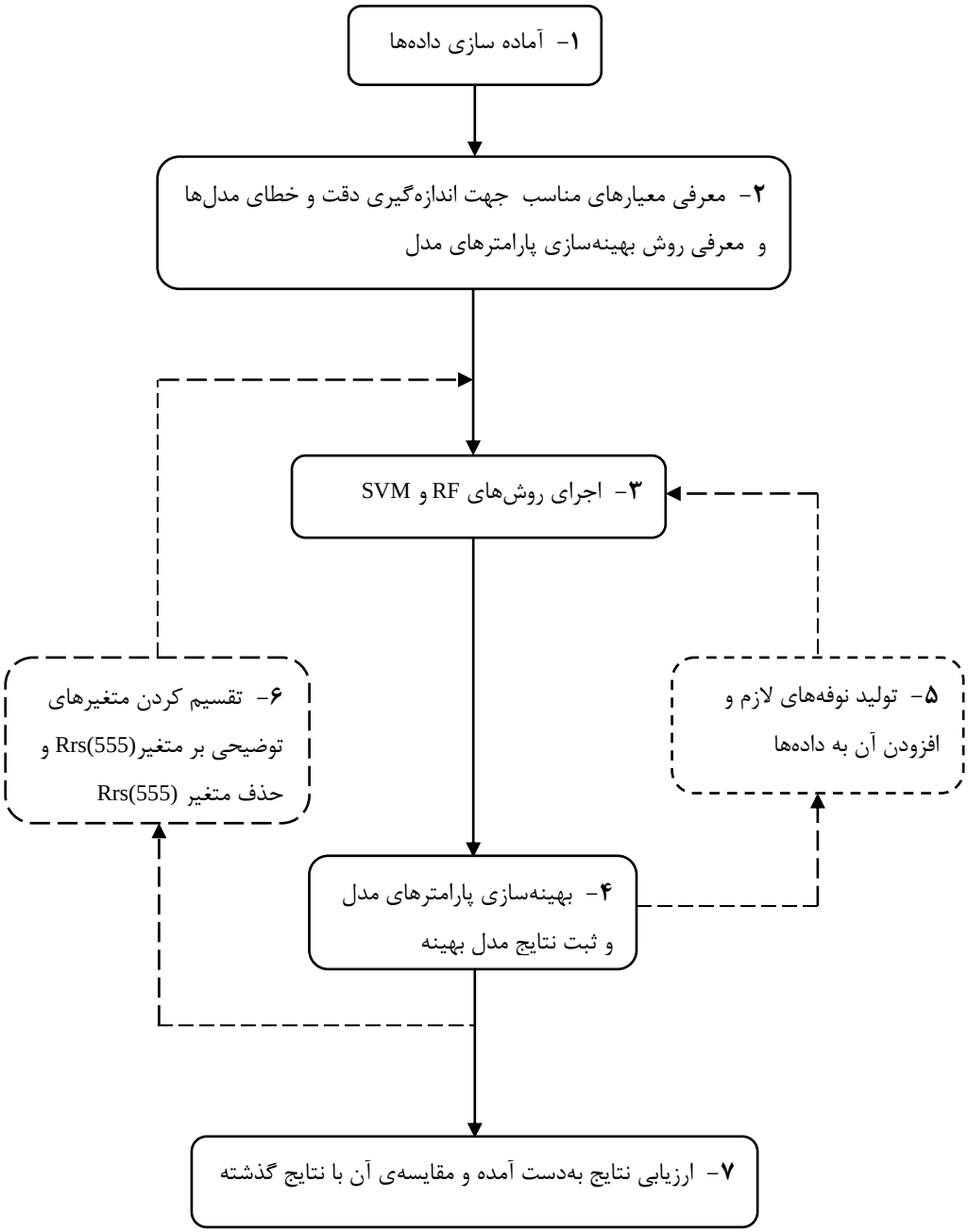
الگوریتم تحقیق

۳-۱ مقدمه

مراحل انجام و شکل‌گیری هر تحقیق و پژوهش را می‌توان به اختصار در چارچوب یک الگوریتم ساده به نمایش درآورد تا مخاطب در کوتاه‌ترین زمان و به ساده‌ترین شکل ممکن، به ساختار کلی مراحل تحقیق پی ببرد. همان‌طور که در فصل اول اشاره شد، هدف نهایی تحقیق در این پایان‌نامه، برآورد برخی از پارامترهای کیفی آب در داده‌های تشعشع طیفی با استفاده از روش‌های نوین یادگیری ماشین است. حال با توجه به این‌که پایگاه داده‌ها، روش‌های جنگل‌های تصادفی و ماشین بردار پشتیبان به‌عنوان ابزارهای این تحقیق معرفی گردیدند، می‌توان اساس مراحل این پژوهش را در قالب الگوریتم تحقیق ارایه داد. در این فصل به تشریح مراحل الگوریتم انجام این تحقیق پرداخته می‌شود.

۳-۲ الگوریتم تحقیق

در شکل (۳-۱)، مراحل کلی تحقیق در قالب الگوریتم به نمایش درآمده است. این الگوریتم شامل کلیه‌ی مراحل و اقداماتی است که برای دستیابی به نتایج نهایی بر روی هر یک از پایگاه داده‌ها به اجرا درآمده است. هر یک از مراحل این الگوریتم، خود شامل جزییاتی است که در ادامه به تفسیر این جزییات پرداخته می‌شود.



شکل (۳-۱)- الگوریتم تحقیق

تشریح مراحل الگوریتم فوق به ترتیب به شرح زیر است.

۳-۲-۱ آماده‌سازی داده‌ها

معمولاً در تحقیقاتی که مدل‌ها و روش‌های آماری در آن‌ها دخیل هستند، در صورت کافی بودن تعداد مشاهدات، داده‌های اولیه به دو دسته‌ی داده‌های مدل‌ساز و داده‌های آزمون تقسیم می‌شود. از داده‌های مدل‌ساز جهت ساخت مدل نهایی استفاده شده و مدل ساخته‌شده، به کمک داده‌های آزمون مورد ارزیابی قرار می‌گیرد. در صورتی که اجرای مدل ساخته‌شده بر روی داده‌های آزمون، منجر به نتایج مطلوبی گردد، ارزش نتایج دو چندان خواهد بود، چرا که داده‌های آزمون در ساخت این مدل نقشی نداشته و تنها جهت اعتبارسنجی مدل به‌کار رفته است. معمولاً فرآیند تقسیم‌بندی داده‌ها به داده‌های مدل‌ساز و آزمون به‌طور تصادفی انجام می‌گیرد. البته تقسیم‌بندی پایگاه داده‌های مورد استفاده در این پژوهش که از جمله پایگاه-های استاندارد داده می‌باشند، از پیش توسط محققین انجام گردیده است.

با توجه به این‌که نوفه، یکی از اجزای جداناپذیر داده‌های تشعشع طیفی است لذا در برآورد پارامترهای کیفی آب، مدلی با ارزش است که بتواند در شرایط نوفه‌ای نیز دارای دقت لازم در برآورد پارامترها باشد. در بسیاری از آزمایشات کامپیوتری، این نوفه‌ها توسط پژوهش‌گر به داده‌ها افزوده می‌شود. از طرفی برای از بین بردن اثر خالص نوفه، نیاز است که پایگاه داده‌ها به اندازه‌ی کافی بزرگ باشند تا بتوان تاثیر نوفه‌ی افزوده‌شده را کاملاً تصادفی دانست. در صورتی که تعداد داده‌ها در یک پایگاه داده، به اندازه‌ی کافی نباشد، محققین برای دستیابی به نتایج مطلوب و قابل اطمینان، مشاهدات هر یک از پایگاه-های داده را به تعداد خاصی تکرار می‌کنند. برای مثال در پایگاه داده‌ی NOMAD که دارای ۱۰۴۸ داده-ی مدل‌ساز و ۱۰۴۸ داده‌ی آزمون است، پس از تکرار ۲۰ مرتبه‌ای هر یک از مشاهدات، پایگاه داده‌ی جدید شامل ۲۰۹۶۰ داده‌ی مدل‌ساز و ۲۰۹۶۰ داده‌ی آزمون خواهد بود.

۳-۲-۲ معرفی معیارهای مناسب جهت اندازه‌گیری دقت و خطای مدل‌ها و معرفی روش بهینه‌سازی پارامترهای مدل

به‌منظور ارزیابی توانایی مدل‌ها در برآورد متغیر پاسخ مورد نظر، نیاز به معرفی برخی از معیارها است تا بتوان نتایج مدل‌های مختلف را با یکدیگر مقایسه کرد. در واقع این معیارها با اندازه‌گیری خطای برآورد یا دقت برآورد یک مدل، نتایج حاصل را قابل ارزیابی و قیاس‌پذیر با سایر نتایج می‌کنند. محققین در پایگاه-های مختلف داده، از معیارهای متفاوتی استفاده می‌کنند. در این زیربخش، چند معیار پرطرفدار را معرفی می‌کنیم.

۳-۲-۲-۱ ضریب تعیین

ضریب تعیین[1] یک معیار آماری می‌باشد که بیانگر آن است که چه درصدی از تغییرات متغیر پاسخ را می‌توان توسط مدل ساخته‌شده توجیه کرد. ضریب تعیین را معمولاً با R^2 نشان می‌دهند و مقدار آن بین صفر و یک است. در واقع R^2 معیاری است که بیانگر دقت مدل موردنظر در برآورد متغیر پاسخ می‌باشد. فرمول محاسباتی R^2 به‌صورت زیر است.

$$R^2 = \frac{SS_M}{SS_T}, \qquad (۱-۳)$$

که در آن، SS_M مجموع توان‌های دوم مدل برآورد شده و SS_T مجموع توان‌های دوم کل، با استفاده از روابط زیر به‌دست می‌آیند.

$$SS_M = \sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2, \qquad (۲-۳)$$

$$SS_T = \sum_{i=1}^{N}(Y_i - \overline{Y})^2, \qquad (۳-۳)$$

[1] Coefficient of Determination

که در آن، Y_i مقدار متغیر پاسخ در مشاهده‌ی i–ام، $\widehat{Y}_i$ مقدار برآورد متغیر پاسخ در مشاهده‌ی i–ام و $\overline{Y}$ میانگین حسابی مقادیر متغیر پاسخ در N مشاهده می‌باشد.

۳-۲-۲-۲ میانگین توان دوم خطا

یکی از راه‌های اندازه‌گیری خطای برآورد مدل، استفاده از معیار میانگین توان دوم خطا[1] است که آن‌را به اختصار MSE می‌نامند و مقدار آن از رابطه‌ی زیر به‌دست می‌آید.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2, \qquad (۳-۴)$$

که در آن، Y_i مقدار متغیر پاسخ در مشاهده‌ی i–ام است و $\widehat{Y}_i$ مقدار برآورد متغیر پاسخ در مشاهده‌ی i–ام می‌باشد.

۳-۲-۲-۳ جذر میانگین توان دوم خطا

در برخی از موارد، برای اندازه‌گیری خطای برآورد مدل از جذر MSE استفاده می‌شود که آن‌را در اصطلاح جذر میانگین توان دوم خطا[2] می‌نامند. این معیار را که به اختصار با علامت RMSE نشان می‌دهند، با استفاده از رابطه‌ی زیر می‌توان محاسبه کرد.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2}. \qquad (۳-۵)$$

۳-۲-۲-۴ میانگین قدر مطلق خطای نسبی

در برخی از پژوهش‌ها، نیاز است که خطای برآورد مدل، به‌صورت نسبی اندازه‌گیری شود. در این نوع تحقیقات، مقدار خطای مطلق نمی‌تواند معیار مناسبی برای بیان خطا باشد. مثلاً اگر مقدار واقعی یک

[1] Mean Squared Error
[2] Root of Mean Squared Error

پارامتر ۱۰۰ واحد و مقدار برآورد آن ۷۰ واحد باشد مقدار خطای نسبی ۳۰ درصد خواهد بود. در حالی که اگر مقدار واقعی پارامتر ۱۰ واحد و مقدار برآورد آن ۲۰ واحد باشد مقدار خطای نسبی ۲۰۰ درصد خواهد بود. این در حالی است که در مثال اول خطای مطلق ۳۰ واحد و در مثال دوم ۱۰ واحد است.

یکی از معیارهای اندازه‌گیری خطای برآورد مدل که در زمینه‌ی پژوهش‌های زیست‌محیطی هم کاربرد فراوانی دارد، میانگین قدر مطلق خطای نسبی[1] است. این معیار را به اختصار با علامت MPAE نشان می‌دهند و با استفاده از رابطه‌ی (۳-۶) محاسبه می‌شود.

$$\mathrm{MPAE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|Y_i - \hat{Y}_i|}{Y_i}\right) \times 100, \qquad (۳-۶)$$

که در آن، Y_i مقدار متغیر پاسخ در مشاهده‌ی i-ام است و $\hat{Y}_i$ مقدار برآورد متغیر پاسخ در مشاهده‌ی i-ام می‌باشد.

۳-۲-۲-۵ معرفی روش بهینه‌سازی پارامترهای مدل

تمامی پارامترهای مدل حاصل از روش SVM از نوع مقادیر پیوسته می‌باشند. در این روش تغییرات خطای مدل (یا دقت مدل) به ازای افزایش این پارامترها، اغلب دارای روند خاصی (صعودی یا نزولی) نمی‌باشد. در چنین مواردی پژوهش‌گران برای بهینه‌سازی پارامترهای مدل از روش‌های مختلفی استفاده می‌کنند. در این پایان‌نامه، برای بهینه‌سازی پارامترهای مدل SVM از الگوریتم بهینه‌سازی هوک و جیوز[2]، (۱۹۶۱) استفاده شده است. برای درک روش بهینه‌سازی این الگوریتم به مثال زیر توجه کنید.

فرض کنید در مدلی سه پارامتر A، B و C وجود داشته باشد. در این روش، بهینه‌سازی ابتدا با ثابت نگه‌داشتن دو پارامتر B و C در دو مقدار دلخواه $B=b_0$ و $C=c_0$ و با تغییر دادن مقدار پارامتر A، مقدار

[1] Mean Percentage of Absolute Error
[2] Hooke and Jeeves Algorithm

بهینه‌ی این پارامتر ($A=a_1$) به‌دست می‌آید. در مرحله‌ی بعد با ثابت نگه‌داشتن $A=a_1$ و $C=c_0$ مقدار بهینه‌ی پارامتر B به‌دست می‌آید که آن‌را $B=b_1$ می‌نامیم. سپس با ثابت نگه‌داشتن $A=a_1$ و $B=b_1$ مقدار بهینه‌ی پارامتر C به‌دست می‌آید و این مقدار را $C=c_1$ می‌نامیم. مراحل فوق تکرار می‌شود تا مقادیر بهینه شده‌ی پارامترها به‌هنگام[1] شود. این فرآیند تا زمانی که پارامترهای بهینه‌ی به‌هنگام‌شده ثابت بمانند، ادامه پیدا خواهد کرد (نلز[2]، ۲۰۰۱).

۳-۲-۳ اجرای روش‌های RF و SVM

در این گام از تحقیق، جهت اجرای روش‌های RF و SVM به‌ترتیب از بسته‌های[3] randomForest و e1071 در نسخه‌ی ۲,۹,۲ نرم افزار منبع-باز[4] R استفاده شده است.

۳-۲-۴ بهینه‌سازی پارامترهای مدل و ثبت نتایج مدل بهینه

همان‌طور که در فصل دوم اشاره شد هر یک از روش‌های RF و SVM در ساختار خود دارای پارامترهای متعددی هستند که تعیین مقدار این پارامترها در اختیار کاربر است. یکی از مهم‌ترین مراحل تحقیق، انتخاب صحیح پارامترهای مدل برای رسیدن به نتایج مطلوب است. معمولاً در فرآیند بهینه‌سازی پارامترهای مدل، از انواع نمودارهای روند تغییرات و الگوریتم‌های بهینه‌سازی استفاده می‌شود. در این پایان‌نامه نیز هر دو روش مذکور به‌کار گرفته شده است. توجه شود که در برخی موارد در بهینه کردن یک پارامتر، یک رابطه‌ی مستقیم بین پیچیدگی مدل[5] و عملکرد آن وجود دارد. به‌طوری که بهبود عملکرد مدل نیازمند پیچیدگی بیشتر در مدل است. لذا کاربر انتخاب می‌کند که حاضر است چه مقدار پیچیدگی مدل را بپذیرد تا به دقت مدنظرش برسد. در واقع ممکن است کاربر از مدل مطلوب‌تر صرف‌نظر کند تا

[1] Update
[2] Nelles O.
[3] Package
[4] Open Source
[5] Model Complexity

مدلی ساده‌تر و کم‌هزینه‌تر (در اینجا هزینه می‌تواند زمان اجرای مدل باشد) داشته باشد. بهینه‌سازی پارامترهای مدل RF با استفاده از نمودارهای تغییرات انجام شده است. همچنین همان‌طور که در زیربخش ۳-۲-۲-۵ اشاره شد، در بهینه‌سازی پارامترهای مدل SVM از الگوریتم هوک و جیوِز استفاده شده است. پس از آن‌که در یک پایگاه داده، مدل بهینه پیدا شد، نتایج مورد نیاز این مدل (معیارهای دقت و خطا و مقادیر پیش‌بینی شده‌ی متغیر پاسخ) ثبت می‌شود.

۳-۲-۵ تولید نوفه‌های لازم و افزودن آن به داده‌ها

با توجه به اثری که اتمسفر بر امواج تابشی خورشید می‌گذارد (مطابق شکل (الف-۲) در پیوست الف)، نوفه یکی از اجزای تفکیک‌ناپذیر داده‌های تشعشع طیفی است. مدل بهینه در یک پایگاه داده در صورتی می‌تواند کارا باشد که بتواند در شرایط نوفه‌ای نیز عملکرد مناسبی از خود نشان دهد. لذا علاوه بر این‌که نتایج مدل بر روی داده‌های خام ارزیابی می‌شود، لازم است که در شرایط نوفه‌ای بودن داده‌ها نیز، نتایج مورد بررسی قرار گیرند. بدین منظور باید با استفاده از توزیع‌های احتمالی معین، داده‌ها را نوفه‌دار کرد. در این زمینه، اکثر پژوهش‌گران نوفه‌ی لازم را با استفاده از توزیع نرمال تولید می‌کنند. با این حال با توجه به اینکه برخی از محققین توزیع یکنواخت را به‌کار برده‌اند، در این پایان‌نامه هر دو نوع نوفه به‌کار می‌رود تا بتوان نتایج را در حالت جامع‌تری با نتایج گذشته مورد مقایسه قرار داد. تجربه‌ی مهندسین و محققین در این زمینه نشان داده است که نوفه‌های لازم، بایستی در یک چارچوب خاص تولید شوند تا بهتر بتوانند رفتار تاثیرات اتمسفری را در برگیرند. نوفه‌های مورد نیاز و نحوه‌ی تولید آن‌ها به قرار زیر است.

۳-۲-۵ -۱ نوفه‌ی نرمال[1]

پژوهش‌گران در این زمینه، داده‌هایی با نوفه‌های ۵، ۱۰ و ۲۰ درصدی شبیه‌سازی شده توسط توزیع نرمال را مطابق جدول (۳-۱)، تولید می‌کنند (دورند[2] و همکاران، ۲۰۰۰).

جدول (۳-۱)- روابط تولید نوفه‌ی نرمال

نوع نوفه	رابطه‌ی تولید نوفه
نوفه ۵٪ نرمال	$N= (0.05\ r + 1) \times F$
نوفه ۱۰٪ نرمال	$N= (0.1\ r + 1) \times F$
نوفه ۲۰٪ نرمال	$N= (0.2\ r + 1) \times F$

در جدول (۳-۱)، r مقدار نوفه‌ی تولیدشده با استفاده از توزیع N(0,1) ، F مقدار داده‌ی اولیه و N مقدار نهایی داده‌ی نوفه‌ای است.

توجه کنید که اثر اتمسفر بر روی امواج تابشی و بازتابشی خورشید است که به عنوان متغیرهای توضیحی لحاظ شده‌اند. لذا این نوفه‌ها تنها بر روی متغیرهای توضیحی اعمال می‌شود و متغیر پاسخ تغییری نخواهد کرد. در واقع نحوه‌ی تولید داده‌های نوفه‌دار، مثلاً در پایگاه داده‌ی NOMAD، به این صورت است: با توجه به این‌که تعداد مشاهدات داده‌های مدل‌ساز (پس از ۲۰ برابر کردن مشاهدات) ۲۰۹۶۰ می‌باشد و هر مشاهده دارای ۶ متغیر توضیحی است، باید ۱۲۵۷۶۰ (یعنی۲۰۹۶۰×۶) عدد تصادفی با توزیع N(0,1) تولید و با قرار دادن در روابط جدول (۳-۱)، داده‌های نوفه‌ای را تولید کرد. از همین ۱۲۵۷۶۰ عدد تصادفی، برای ساخت داده‌های نوفه‌ای آزمون نیز استفاده می‌شود.

[1] Normal Noise
[2] Durand D.

۳-۲-۵ -۲ نوفه‌ی یکنواخت[1]

داده‌هایی با نوفه‌های ۵، ۱۰، ۲۰ و ۳۰ درصدی شبیه‌سازی شده توسط توزیع یکنواخت، مطابق جدول (۳-۲)، تولید می‌شوند (دورند و همکاران، ۲۰۰۰).

جدول (۳-۲)- روابط تولید نوفه‌ی یکنواخت

نوع نوفه	رابطه‌ی تولید نوفه
نوفه ۵٪ یکنواخت	$N= (0.05\ r + 0.975) \times F$
نوفه ۱۰٪ یکنواخت	$N= (0.1\ r + 0.95) \times F$
نوفه ۲۰٪ یکنواخت	$N= (0.2\ r + 0.9) \times F$
نوفه ۳۰٪ یکنواخت	$N= (0.3\ r + 0.85) \times F$

در جدول (۳-۲)، r مقدار نوفه‌ی تولیدشده با استفاده از توزیع U(0,1)، F مقدار داده‌ی اولیه و N مقدار نهایی داده‌ی نوفه‌ای است.

توجه کنید که بدین ترتیب پس از تولید داده‌هایی با نوفه‌های نرمال و یکنواخت، در واقع یک پایگاه داده تبدیل به ۸ پایگاه داده خواهد شد (بدون نوفه، نوفه‌های ۵، ۱۰ و ۲۰ درصدی نرمال و نوفه‌های ۵، ۱۰، ۲۰ و ۳۰ درصدی یکنواخت) که در هر یک از این پایگاه‌های داده، بایستی به‌طور جداگانه بهینه‌سازی پارامترها انجام شده و نتایج ثبت گردند.

۳-۲-۶ تقسیم متغیرهای توضیحی بر متغیر $R_{rs}(555)$ و حذف متغیر $R_{rs}(555)$

یکی از مسایلی که در زمینه‌ی برآورد پارامترهای کیفی آب در داده‌های تشعشع طیفی برای پژوهش‌گران حایز اهمیت است، قابلیت روش به‌کار رفته در شرایطی است که متغیرهای توضیحی به‌صورت غیر خطی

[1] Uniform Noise

در مدل حضور داشته باشند. تجربه‌ی پژوهش‌گران در این زمینه نشان داده است که اگر مقدار متغیرهای توضیحی مشاهدات را بر متغیر $R_{rs}(555)$ تقسیم کرده و متغیر $R_{rs}(555)$ حذف گردد، پایگاه داده‌ی جدید می‌تواند جهت مدل‌سازی مورد استفاده قرار گیرد.

با در نظر گرفتن مطالب این زیربخش و زیربخش ۳-۲-۵-۲ می‌توان نتیجه گرفت که هر پایگاه داده پس از تولید نوفه و همچنین تقسیم متغیرهای توضیحی بر متغیر $R_{rs}(555)$ ، به ۱۶ پایگاه داده‌ی مجزا تبدیل خواهد شد. هر یک از این پایگاه‌ها نیازمند بهینه‌سازی پارامترهای مدل به‌صورت جداگانه خواهد بود.

۳-۲-۷ ارزیابی نتایج به‌دست آمده و مقایسه‌ی آن با نتایج گذشته

با توجه به معرفی معیارهای اندازه‌گیری دقت و خطای مدل‌ها، می‌توان نتایج حاصل از برازش مدل‌های مختلف بر روی هر پایگاه داده را با استفاده از این معیارها مورد ارزیابی قرار داد. در این پایان‌نامه، ضمن ارزیابی نتایج دو روش RF و SVM در برآورد پارامترهای کیفی آب، به بررسی نتایج حاصل از این دو روش با نتایج سایر روش‌ها که توسط محققین در گذشته به‌کار گرفته شده‌اند، پرداخته می‌شود.

فصل چهارم

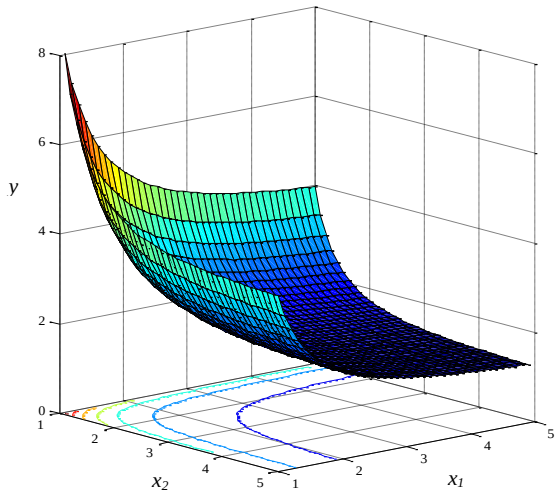
ارایه و تحلیل نتایج تجربی

در این فصل به‌منظور ارزیابی و بررسی عملکرد و توانایی‌های دو روش RF و SVM، ابتدا یک مدل ساختگی با داده‌های شبیه‌سازی‌شده (بخش۴-۱) ارایه گردیده است و سپس برای برآورد پارامترهای کیفی آب، داده‌های معرفی‌شده در فصل دوم مورد بررسی قرار می‌گیرد. همچنین نتایج به‌دست آمده در هر مورد با نتایجی که قبلاً توسط سایر روش‌ها به‌دست آمده است، مقایسه و ارزیابی می‌شوند.

۴-۱ برآورد یک مدل غیرخطی با استفاده از روش‌های RF و SVM

مدل غیر خطی (۴-۱) را به عنوان یک تابع آزمون در نظر بگیرید. نمودار این تابع در شکل (۴-۱) آمده است. به‌منظور بررسی عملکرد دو روش RF و SVM ، تعداد ۵۰ مشاهده برای ۴ متغیر توضیحی (x_1, x_2, x_3, x_4) تولید و متغیر پاسخ (y)، بر اساس رابطه‌ی (۴-۱) محاسبه شده است (سوگِنو[1] و یاسوکاوا[2]، ۱۹۹۳).

$$y = (1 + {x_1}^{-2} + {x_2}^{-1.5})^2, \qquad 1 \le x_1, x_2 \le 5, \tag{۴-۱}$$



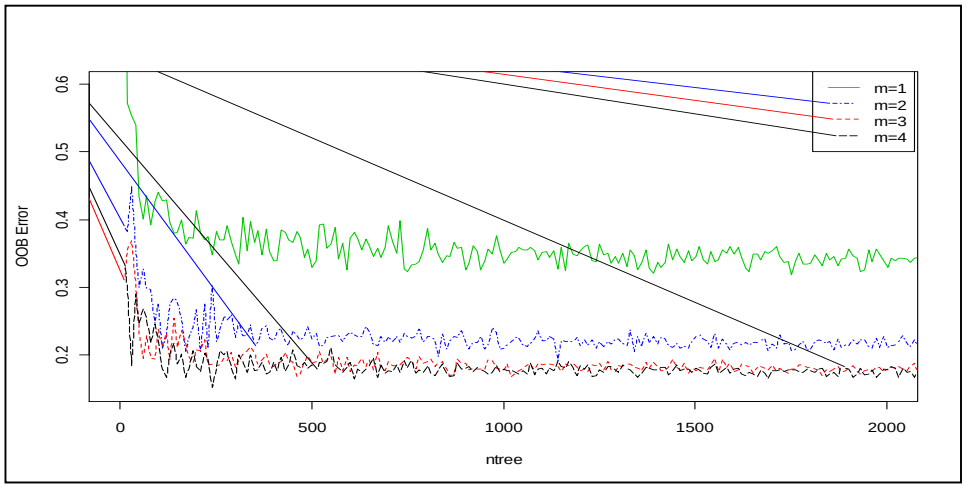
شکل (۴-۱)- نمایش سه بعدی مدل غیرخطی (۴-۱)

[1] Sugeno M.
[2] Yasukawa T.

توجه کنید، دو متغیر x_3 و x_4 در ساختن متغیر پاسخ هیچ نقشی ندارند. با توجه به ویژگی تشخیص اهمیت متغیرها در مدل RF ، بایستی دید که آیا اهمیت نداشتن متغیرهای x_3 و x_4 در برآورد متغیر پاسخ، توسط روش RF تشخیص داده می‌شود یا خیر.

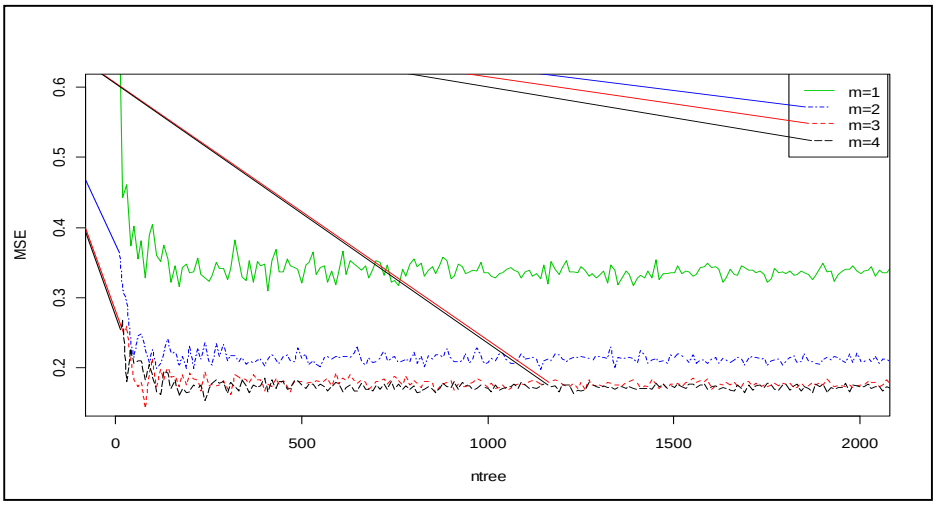
۴-۱-۱ نتایج روش RF

برای دستیابی به بهترین مدل، باید پارامترهای *ntree* و *m* بهینه گردند. بدین منظور، خطای *OOB* به‌ازای مدل‌های مختلف (ترکیبات مختلفی از m=1,2,3,4 و $ntree$=10, 20, … , 2100) محاسبه شده است (شکل (۴-۲)). همان‌طور که ملاحظه می‌شود، برای *m*های مختلف، مقادیر خطا به ازای *ntree* >1000 به پایداری رسیده است. با توجه به شکل می‌توان گفت به ازای m=1 ، مدل دارای بیشترین خطاست، و این امر کاملاً طبیعی است زیرا در هر گره از هر درخت، فقط از یک متغیر، برای افراز فضای ۴ بعدی متغیرهای توضیحی استفاده گردیده است. این در حالی است که به ازای m=2 مقدار این خطا تا حد چشمگیری کاهش می‌یابد. مقدار خطا به ازای m=3,4 تفاوت معنی‌داری نداشته ولی مقادیر آن اندکی کمتر از مقادیر حالت m=2 است.



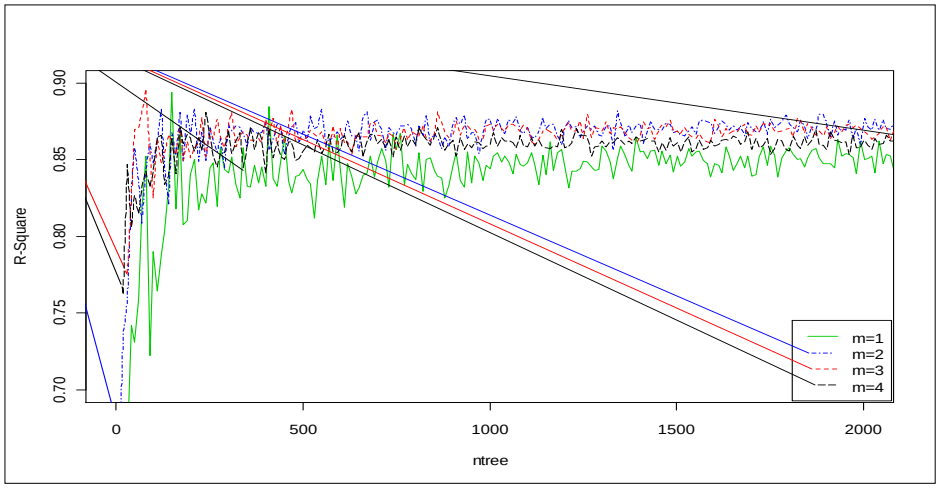
شکل (۴-۲)- روند خطای *OOB* بر حسب پارامتر *ntree* به ازای m=1,2,3,4 برای تابع آزمون (۴-۱)

شکل (۴-۳)، بیانگر مقدار MSE بر حسب پارامتر *ntree* و به ازای m=1,2,3,4 می‌باشد. همان‌طور که دیده می‌شود این نمودار هم نتایج شکل(۴-۲) را کاملاً تایید می‌کند.



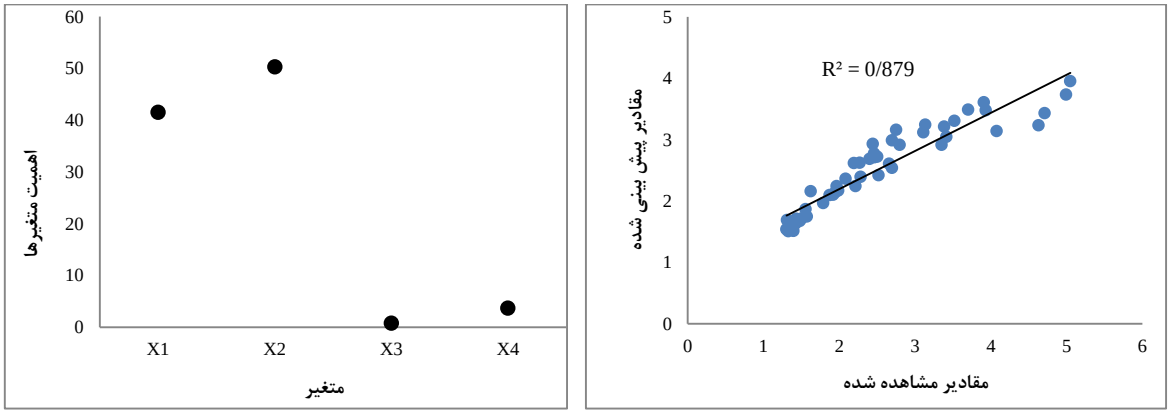
شکل (۴-۳)- روند MSE بر حسب پارامتر *ntree* به ازای m=1,2,3,4 برای تابع آزمون (۴-۱)

همچنین در شکل (۴-۴)، نمودار ضریب تعیین (R^2) در تنظیم‌های مختلف، بر حسب پارامتر *ntree* و به ازای m=1,2,3,4 رسم شده است. همانگونه که پیداست، نوسانات R^2 برای m=2,3,4، به‌ازای *ntree*>1000 ناچیز بوده و از آن‌جایی که خطا، به‌ازای m=2,3,4 و *ntree*>1000 تفاوت چشمگیری ندارد، می‌توان m=2 و *ntree*=1000 را به عنوان مقادیر بهینه‌ی پارامترهای روش RF انتخاب کرد.



شکل (۴-۴)- روند R^2 بر حسب پارامتر *ntree* به ازای m=1,2,3,4 برای تابع آزمون (۴-۱)

با اعمال این تنظیمات، مقادیر معیارهای ارزیابی عبارتند از: MSE=0.20 ، OOB Error=0.21 و R^2=0.88. همچنین در نمودار پراکنش مقادیر مشاهده‌شده در مقابل مقادیر پیش‌بینی‌شده (قاب راست شکل (۴-۵))، نقاط حول خط y=x قرار گرفته‌اند که نشان از عملکرد مناسب مدل بهینه دارد.



شکل (۴-۵)- نمودار مقادیر مشاهده‌شده در مقابل مقادیر پیش‌بینی‌شده (قاب راست) و میزان اهمیت متغیرها (قاب چپ) به ازای m=2 و $ntree$=1000 برای تابع آزمون (۴-۱)

قاب چپ در شکل (۴-۵)، میزان اهمیت متغیرها را در مدل بهینه نشان می‌دهد. ناچیز بودن اهمیت متغیرهای x_3 و x_4، حاکی از آن است که روش RF به‌خوبی توانسته است متغیرهای غیرتاثیرگذار در متغیر پاسخ را شناسایی کند.

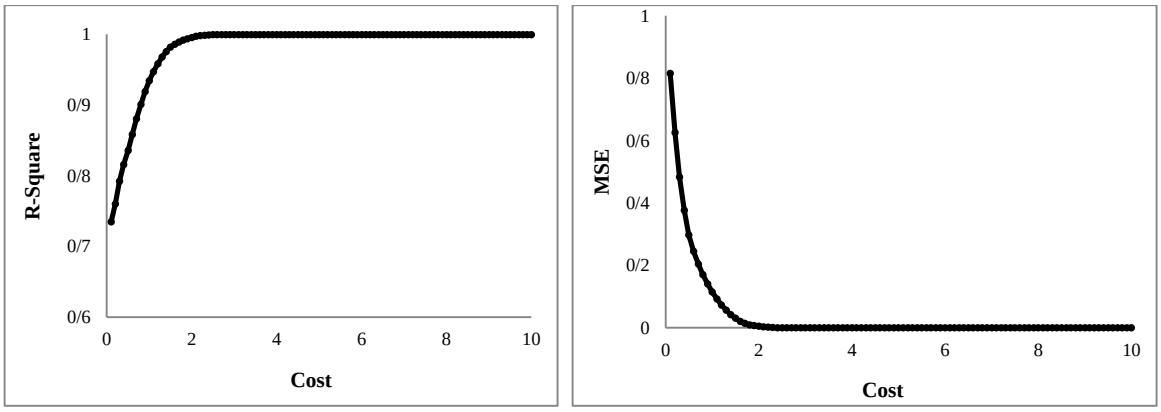
۴-۱-۲ نتایج روش SVM

در فصل دوم، چهار نوع تابع هسته در ساختار روش SVM معرفی گردید که هر یک از آن‌ها در ساختار خود دارای پارامترهای متعددی می‌باشد. برای دستیابی به بهترین مدل، علاوه بر انتخاب بهترین تابع هسته، بایستی پارامترهای آن هسته نیز بهینه شوند. بدین منظور، نتایج حاصل از اعمال هر یک از توابع هسته، مورد بررسی قرار می‌گیرد.

۴-۱-۲-۱ هسته‌ی نرمال

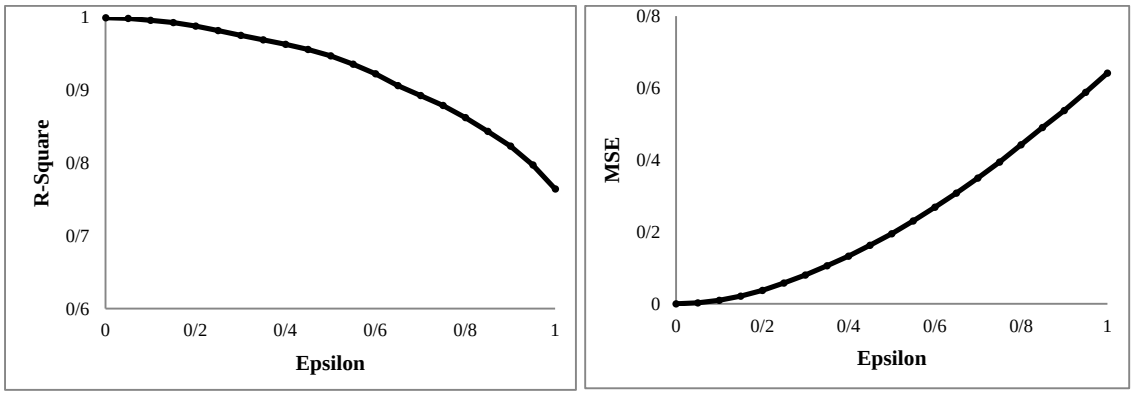
در مدل حاصل از هسته‌ی نرمال (RBF)، علاوه بر پارامترهای epsilon و cost، پارامتر gamma نیز وجود دارد. برای یافتن ترکیب بهینه از مقادیر این پارامترها، از الگوریتم بهینه‌سازی هوک و جیوِز استفاده شده است.

ابتدا با انتخاب مقادیر اولیه‌ی epsilon=0.5 و gamma=1 (این مقادیر به‌صورت اختیاری توسط کاربر انتخاب می‌شوند)، مقدار پارامتر cost در این مرحله بهینه می‌گردد. در قاب‌های چپ و راست شکل (۴-۶)، به‌ترتیب نمودارهای R^2 و MSE به ازای مقادیر مختلف پارامتر cost به نمایش درآمده است. با توجه به شکل، ملاحظه می‌شود که نوسانات مقادیر این دو معیار به ازای cost>7 به پایداری می‌رسد (کمینه‌ی MSE و بیشینه‌ی R^2).



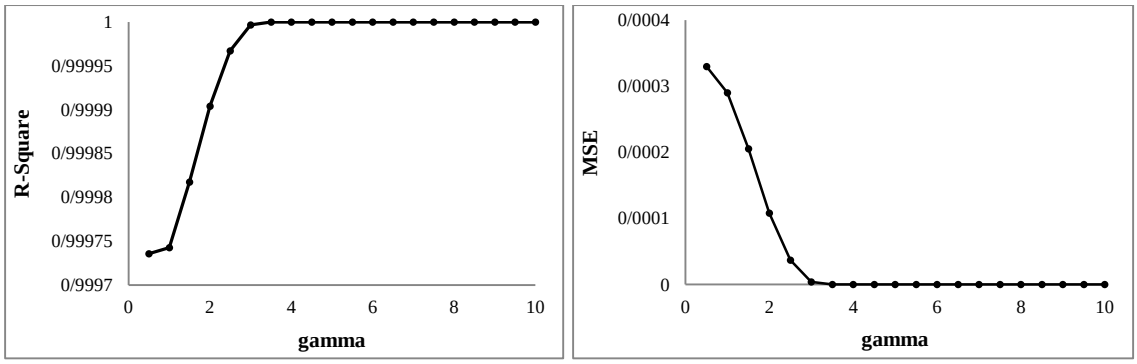
شکل (۴-۶)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر cost، به ازای epsilon=0.5 و gamma=1 در هسته‌ی نرمال برای تابع آزمون (۴-۱)

اکنون بایستی با ثابت نگه‌داشتن cost=7 و gamma=1، مقدار پارامتر epsilon بهینه شود. در قاب‌های چپ و راست شکل (۴-۷)، به‌ترتیب نمودارهای R^2 و MSE به ازای مقادیر مختلف پارامتر epsilon دیده می‌شود. ملاحظه می‌شود که پارامتر epsilon با MSE نسبت مستقیم و با R^2 نسبت عکس دارد و لذا مقدار بهینه‌ی epsilon برابر با صفر به‌دست می‌آید.



شکل (۴-۷)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر epsilon، به ازای cost=7 و gamma=1 در هسته‌ی نرمال برای تابع آزمون (۴-۱)

در مرحله‌ی بعد با ثابت نگه‌داشتن cost=7 و epsilon=0، مقدار پارامتر gamma بهینه می‌گردد. در قاب‌های چپ و راست شکل (۴-۸)، به‌ترتیب نمودارهای R^2 و MSE به ازای مقادیر مختلف پارامتر gamma نشان داده شده است. می‌توان دید که به ازای gamma>4.9، مقدار MSE در مقدار کمینه‌ی خود و R^2 در مقدار بیشینه‌ی خود به پایداری می‌رسد و R^2 به ازای این مقدار، یک (۱۰۰٪) می‌شود.



شکل (۴-۸)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر gamma، به ازای cost=7 و epsilon=0 در هسته‌ی نرمال برای تابع آزمون (۴-۱)

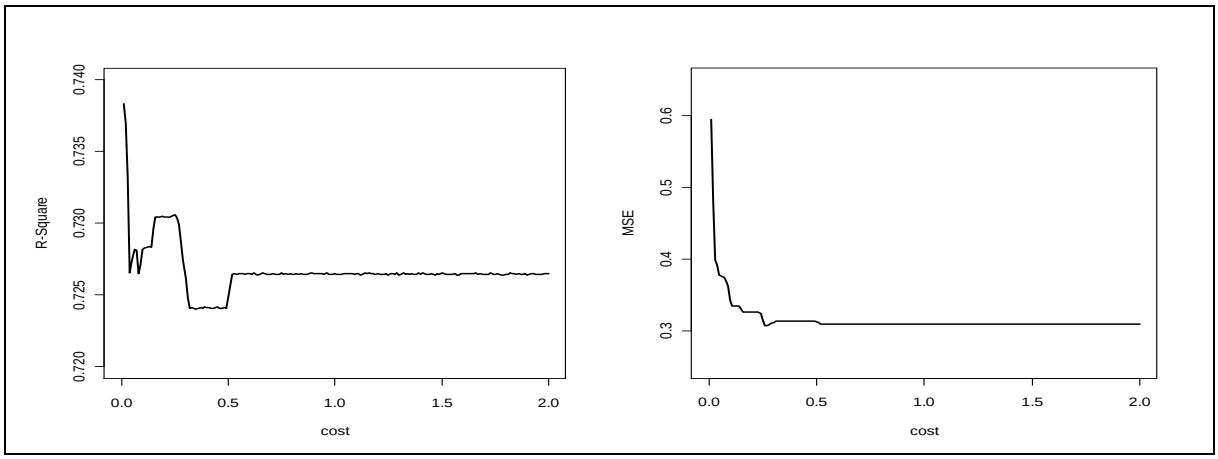
این فرآیند مطابق روش هوک و جیوز بایستی تا زمانی تکرار گردد که دیگر تغییری در برآورد پارامترها حاصل نشود. با تکرار این روند در این مساله، تغییری در نتایج به‌دست آمده حاصل نگردید. لذا

cost=7 ، epsilon=0 و gamma=4.9 به‌عنوان مقادیر بهینه‌ی پارامترهای هسته‌ی نرمال انتخاب می‌شوند. با اعمال این تنظیمات، برای مدل بهینه‌ی SVM، داریم:

MSE=0.000000066, R^2=1.

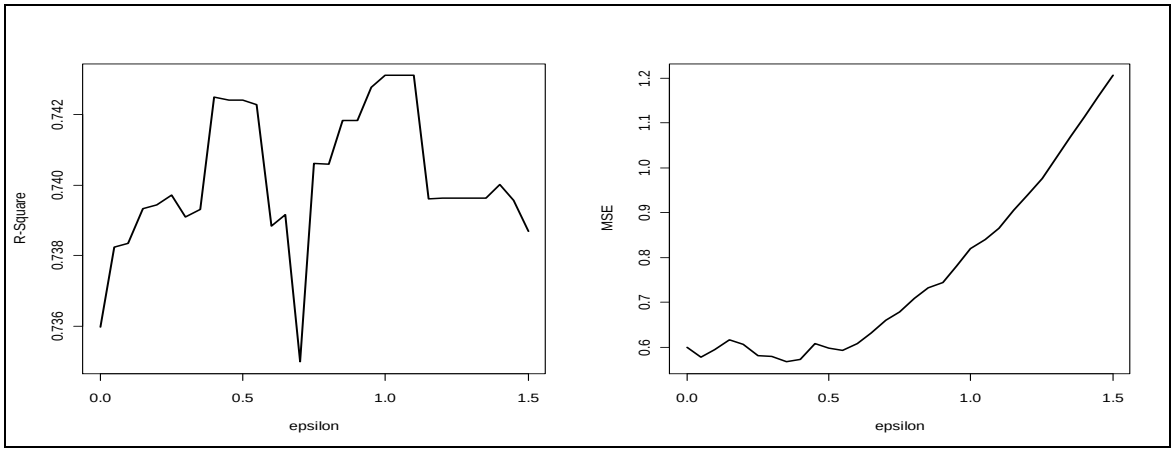
۴-۱-۲-۲ هسته خطی

در تابع هسته‌ی خطی پارامترهای cost و epsilon وجود دارد. برای بهینه‌سازی مدل ابتدا با انتخاب مقدار اولیه‌ی epsilon=0، مقدار پارامتر cost بهینه می‌گردد که مقدار بهینه‌ی این پارامتر با توجه به معیارهای MSE و R^2 کمی متفاوت خواهد بود. با توجه به شکل (۴-۹) ملاحظه می‌شود که مقدار نوسانات R^2 حدود ۱٪ است. بنابراین بر اساس معیار MSE، مقدار بهینه‌ی پارامتر cost=0.3 در نظر گرفته می‌شود.



شکل (۴-۹)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر cost در هسته‌ی خطی برای تابع آزمون (۴-۱)

در ادامه با ثابت نگه داشتن cost=0.3، مقدار بهینه‌ی پارامتر epsilon به‌دست می‌آید که با توجه به شکل (۴-۱۰)، می‌توان مقدار بهینه را epsilon=0.5 در نظر گرفت. در مجموع می‌توان گفت که هسته خطی در این داده‌ها منجر به نتایج چندان مناسبی نمی‌گردد و در ایده‌آل‌ترین حالت مقدار R^2 به ۷۵٪ هم نمی‌رسد و MSE هم حدوداً ۰/۴ خواهد شد.

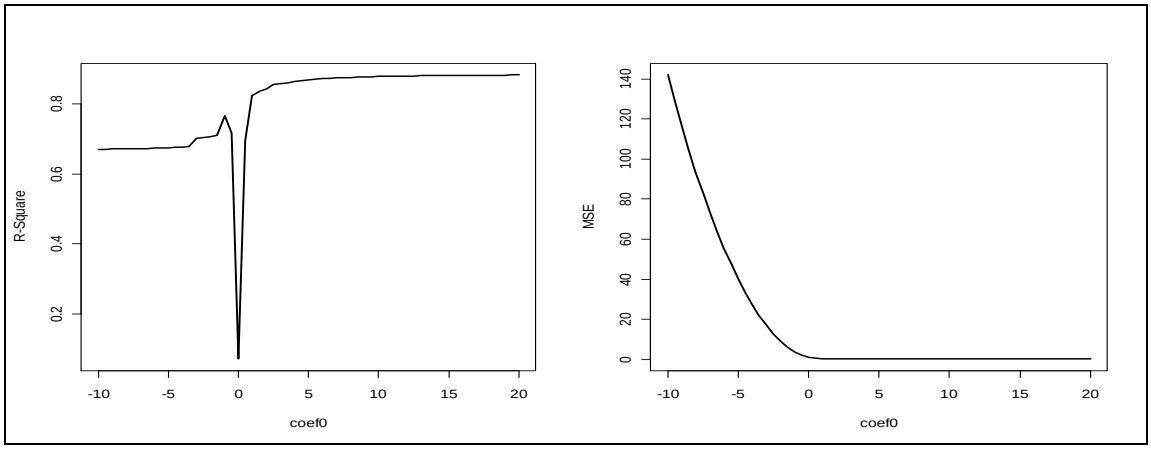


شکل (۴-۱۰)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر epsilon در هسته خطی برای تابع آزمون (۴-۱)

۴-۱-۲-۳ هسته چندجمله‌ای

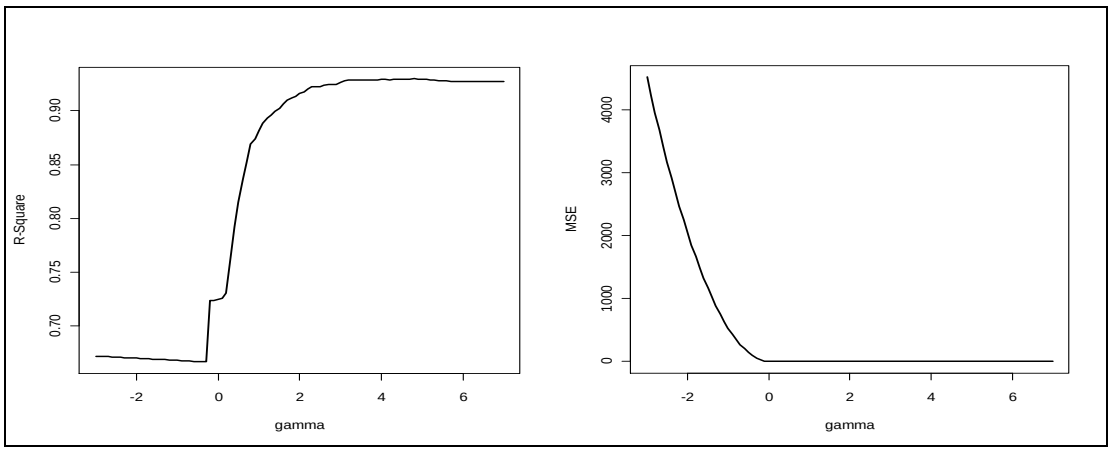
این هسته شامل پارامترهای cost، epsilon، gamma، coef0 و degree می‌باشد، که در آن پارامتر degree تعیین‌کننده‌ی درجه‌ی چندجمله‌ای در این هسته است. در واقع اگر به این پارامتر مقدار ۱ داده شود، تابع هسته چندجمله‌ای به تابع هسته خطی تبدیل می‌گردد. در این بخش، نتایج به‌ازای degree=2,3 محاسبه می‌گردد.

ابتدا با قرار دادن degree=2 و با انتخاب مقادیر اولیه‌ی epsilon=0، cost=0.01 و gamma=1، مقدار پارامتر coef0 در این مرحله بهینه می‌گردد. در قاب‌های چپ و راست شکل (۴-۱۱)، به‌ترتیب نمودارهای R^2 و MSE به ازای مقادیر مختلف پارامتر coef0 به نمایش درآمده است. با توجه به این نمودارها، مقدار این دو معیار به ازای coef0>10 پایدار می‌شود (کمینه‌ی MSE و بیشینه‌ی R^2).



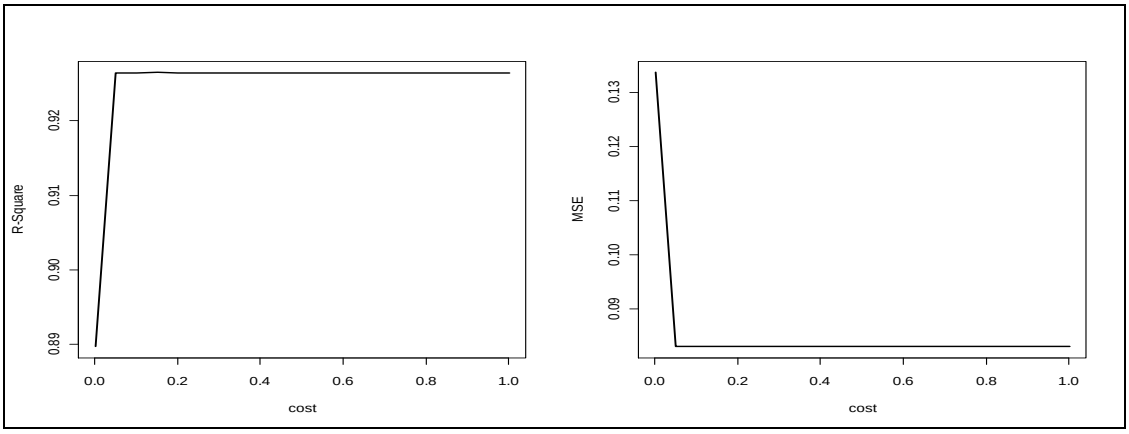
شکل (۴-۱۱)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر coef0 در هسته چندجمله‌ای درجه ۲ برای تابع آزمون (۴-۱)

در ادامه با ثابت نگه‌داشتن epsilon=0، cost=0.01 و coef0=10، پارامتر gamma بهینه می‌شود. با توجه به شکل (۴-۱۲)، مقدار R^2 و MSE به ازای gamma>4 پایدار شده است.



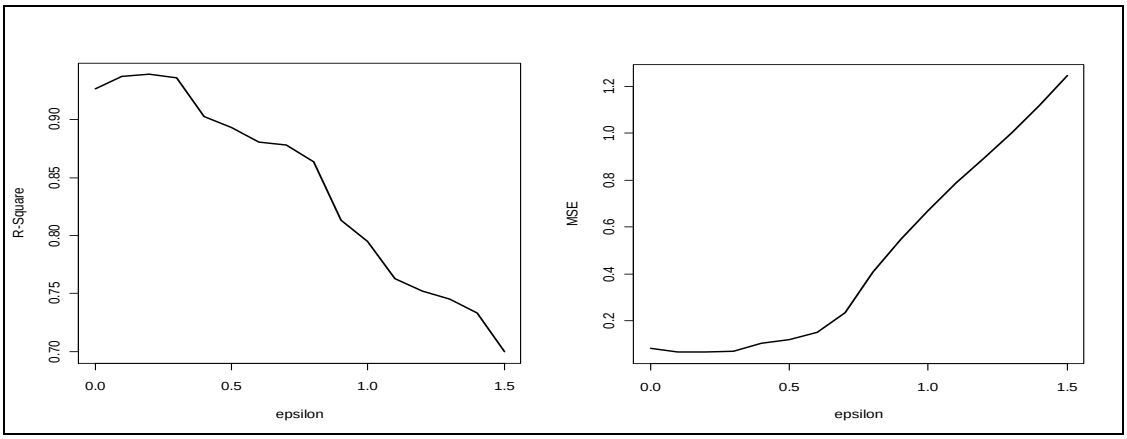
شکل (۴-۱۲)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر gamma در هسته چندجمله‌ای درجه ۲ برای تابع آزمون (۴-۱)

در گام بعد، با ثابت نگه‌داشتن epsilon=0، gamma=4 و coef0=10، پارامتر cost بهینه می‌شود که با توجه به شکل (۴-۱۳)، مقدار بهینه‌ی این پارامتر به ازای cost>0.1 رخ می‌دهد.



شکل (۴-۱۳)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر cost در هسته چندجمله‌ای درجه ۲ برای تابع آزمون (۴-۱)

همچنین با ثابت نگه داشتن cost=0.1، gamma=4 و coef0=10، مقدار بهینه‌ی پارامتر epsilon به ازای ۰/۲ رخ می‌دهد. در این نقطه، بیشینه‌ی R^2 و کمینه‌ی MSE اتفاق افتاده است (شکل (۴-۱۴)).



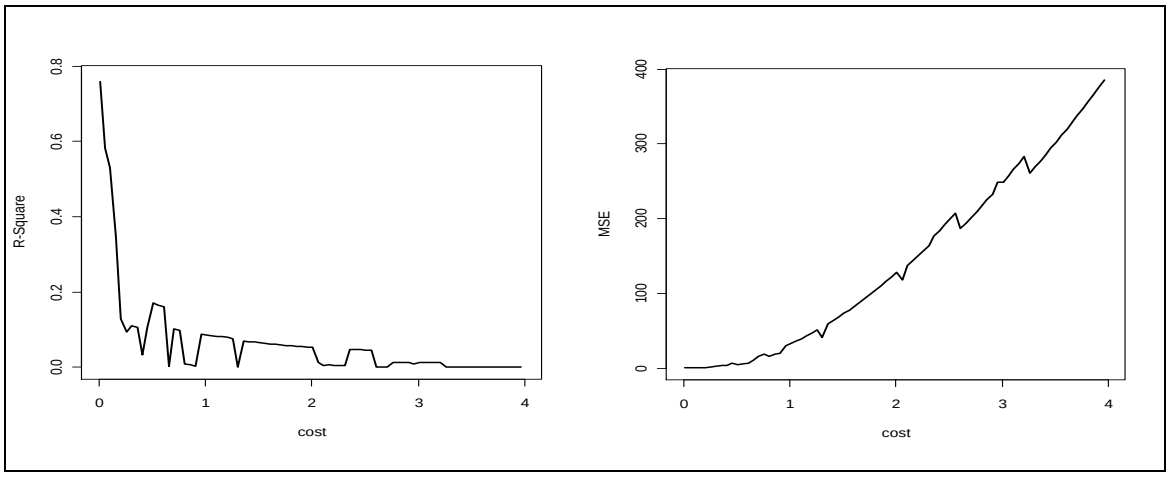
شکل (۴-۱۴)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر epsilon در هسته چندجمله‌ای درجه ۲ برای تابع آزمون (۴-۱)

با ادامه دادن روند فوق به‌منظور به‌هنگام کردن پارامترهای بهینه با روش هوک و جیوز، در نهایت مدلی با پارامترهای cost=0.01، coef0=10، epsilon=0.2 و gamma=4 به‌عنوان مدل بهینه در degree=2 انتخاب می‌شود، که در آن MSE=0.065 و R^2=0.94 می‌باشد.

با انتخاب چندجمله‌ای مرتبه‌ی سوم (degree=3) و تکرار مراحل فوق به‌منظور بهینه‌سازی پارامترهای مدل، مدلی با مشخصات cost=3، coef0=10، epsilon=0 و gamma=5 به‌عنوان مدل بهینه در degree=3 انتخاب می‌شود، که در آن MSE=0.006 و $R^2 \approx 1$ می‌باشد.

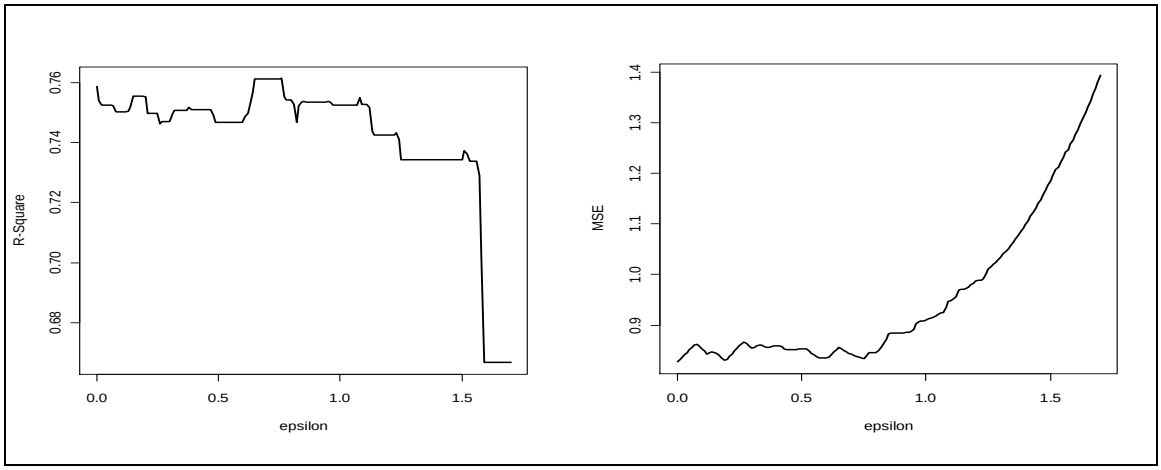
۴-۱-۲-۴ هسته سیگموید

این هسته شامل چهار پارامتر cost، epsilon، gamma و coef0 می‌باشد. ابتدا با انتخاب مقادیر اولیه‌ی epsilon=0، gamma=1 و coef0=0، مقدار پارامتر cost بهینه می‌گردد. با توجه به شکل (۴-۱۵)، می‌توان گفت که مقدار پارامتر cost با معیار MSE نسبت مستقیم و با معیار R^2 نسبت عکس دارد. لذا مقدار بهینه‌ی این پارامتر در cost=0.01 رخ می‌دهد.



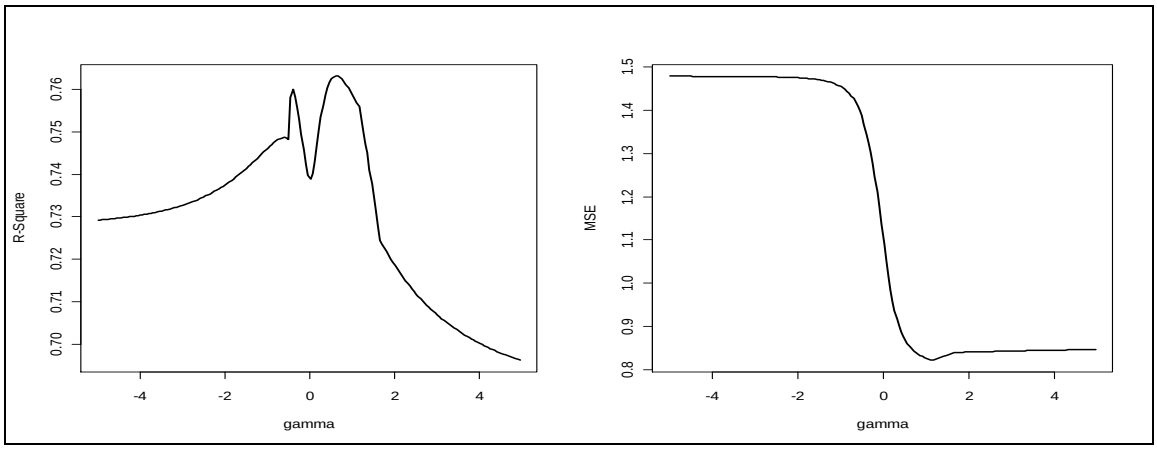
شکل (۴-۱۵)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر cost در هسته سیگموید برای تابع آزمون (۴-۱)

در گام بعد، با ثابت نگه‌داشتن سه پارامتر cost=0.01، gamma=1 و coef0=0، مقدار پارامتر epsilon بهینه می‌شود. با توجه به شکل (۴-۱۶)، روند کلی MSE و R^2 نسبت به epsilon به‌ترتیب صعودی و نزولی است. لذا می‌توان گفت که مقدار بهینه‌ی این پارامتر در نقطه‌ی epsilon=0 رخ می‌دهد.



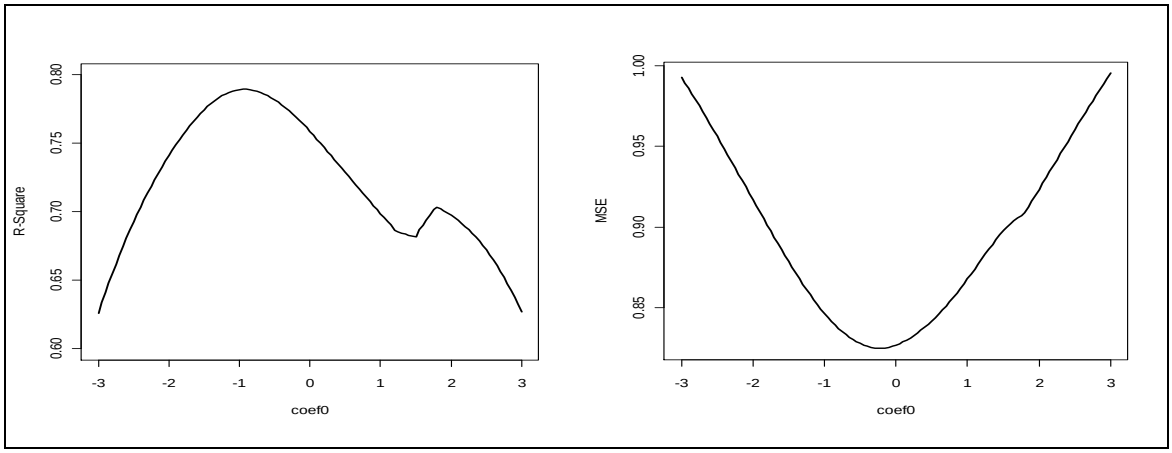
شکل(۴-۱۶)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر epsilon در هسته سیگموید برای تابع آزمون (۱-۴)

در مرحله‌ی بعد، با ثابت نگه‌داشتن سه پارامتر cost=0.01، epsilon=0 و coef0=0، تأثیر پارامتر gamma بر معیارهای موردنظر بررسی می‌گردد. با توجه به شکل (۴-۱۷)، روند نمودار MSE به ازای gamma>1 پایدار می‌شود. لذا با توجه به این‌که بیشینه‌ی R^2 به ازای gamma=1 رخ می‌دهد، می‌توان این نقطه را به‌عنوان مقدار بهینه انتخاب کرد.



شکل (۴-۱۷)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر gamma در هسته سیگموید برای تابع آزمون (۱-۴)

در گام بعد، با ثابت نگه‌داشتن سه پارامتر cost=0.01، epsilon=0 و gamma=1، مقدار پارامتر coef0 بهینه می‌گردد. با توجه به شکل (۴-۱۸)، می‌توان گفت که مقدار بهینه‌ی این پارامتر در نقطه‌ی coef0=-1 رخ می‌دهد. زیرا در این نقطه مقدار کمینه‌ی MSE و مقدار بیشینه‌ی R^2 اتفاق می‌افتد.



شکل (۴-۱۸)- روند MSE (قاب راست) و R^2 (قاب چپ) بر حسب پارامتر coef0 در هسته سیگموید برای تابع آزمون (۴-۱)

با ادامه دادن روند فوق به‌منظور به‌هنگام کردن پارامترهای بهینه، نتایج بهتری حاصل نمی‌گردد و در نهایت مدلی با پارامترهای cost=0.01، coef0=-1، epsilon=0 و gamma=1 به‌عنوان مدل بهینه انتخاب می‌شود، که در آن MSE=0.85 و R^2=0.79 می‌باشد.

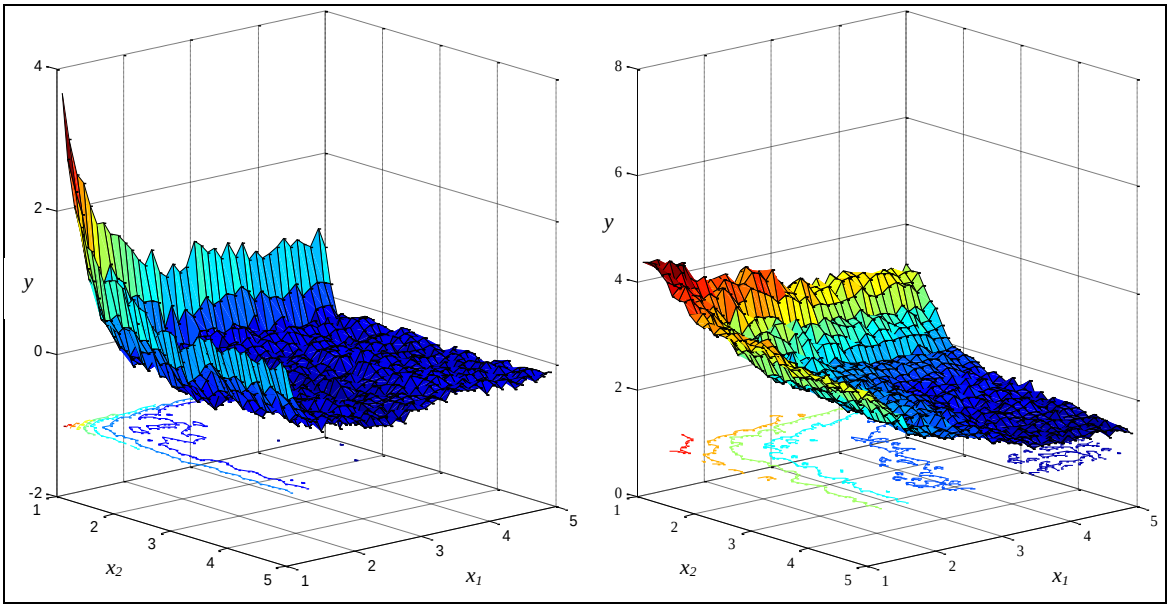
۴-۱-۳ خلاصه‌ی نتایج دو روش RF و SVM در برآورد مدل غیرخطی

با توجه به نتایج به‌دست آمده از دو روش RF و SVM در برآورد مدل غیرخطی مذکور می‌توان نتایج را در قالب جدول زیر ترسیم کرد. با توجه به جدول(۴-۱) می‌توان گفت که روش SVM در حل این مساله نسبت به روش RF دارای عملکرد بهتری می‌باشد. همچنین در روش SVM، بهترین نتایج با استفاده از هسته‌های نرمال و چندجمله‌ای به‌دست آمده است که منجر به بیشترین دقت و کمترین خطا شده است.

جدول (۴-۱)- خلاصه‌ی نتایج دو روش RF و SVM برای تابع آزمون (۴-۱)

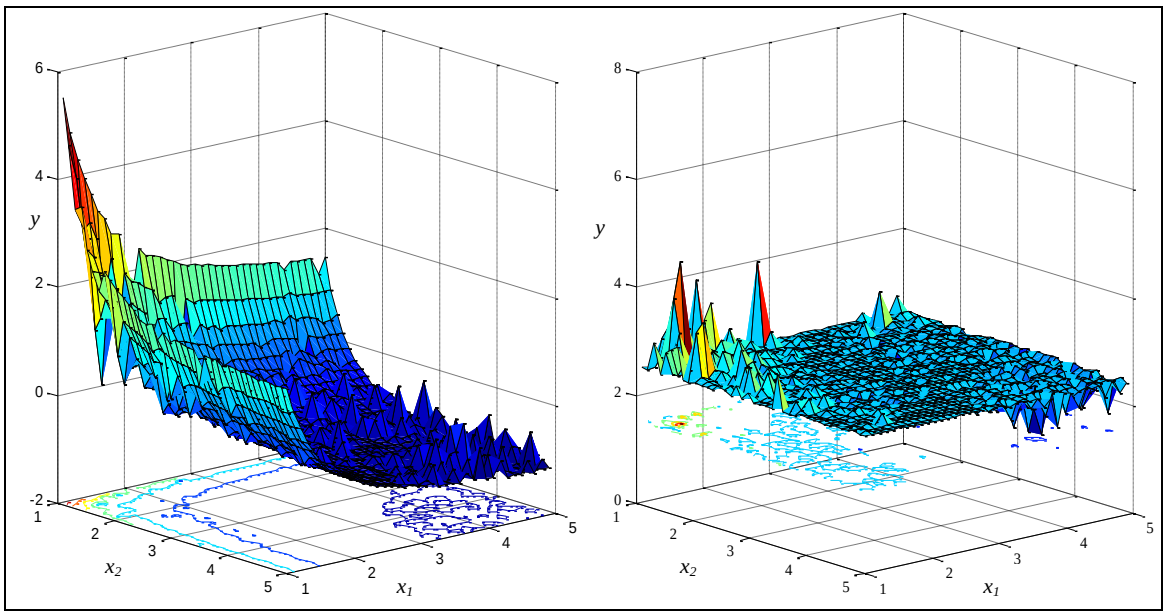
مدل	مشخصات مدل	MSE	R^2
RF	ntree=1000, m=2	۰/۲	٪۸۸
SVM	**RBF**, epsilon=0, cost=7, gamma=4.9	۰	۱
	Linear, epsilon=0.5, cost=0.3	۰/۴	٪۷۵
	Polynomial(degree=2), epsilon=0.2, cost=0.01, gamma=4, coef0=10	۰/۰۶۵	٪۹۴
	Polynomial(degree=3), epsilon=0, cost=3, gamma=5, coef0=10	۰/۰۰۶	٪۹۹/۹
	Sigmoid, epsilon=0, cost=0.01, gamma=1, coef0=-1	۰/۸۵	٪۷۹

به‌منظور بررسی دقیق‌تر این نتایج، برای یک مجموعه نقاط شبکه‌ای، شامل ۲۵۰۰ نقطه در فضای دو بعدی متغیرهای x_1 و x_2 (با فرض ثابت بودن مقادیر x_3 و x_4) مقدار تابع (۴-۱) محاسبه گردید و برای این داده‌های آزمون، مدل‌های بهینه‌ی دو روش RF و SVM مورد ارزیابی گرفت. در شکل (۴-۱۹)، رویه‌ی پاسخ (قاب راست) و خطای (قاب چپ) حاصل از برآورد مدل بهینه‌ی روش RF به‌نمایش در آمده است. با توجه به این شکل، روش RF به‌ازای مقادیر کوچک x_1 و x_2 عملکرد مناسبی نداشته است و بیشترین خطا در این نواحی حاصل گردیده است.



شکل (۴-۱۹)- نمایش سه‌بعدی برآورد (قاب راست) و خطای (قاب چپ) مدل بهینه‌ی روش RF برای تابع آزمون (۴-۱)

همچنین شکل (۴-۲۰)، رویه‌ی پاسخ (قاب راست) و خطای (قاب چپ) حاصل از برآورد مدل بهینه‌ی روش SVM را به‌نمایش در آورده است. با مقایسه‌ی قاب راست شکل (۴-۲۰) با شکل (۴-۱)، ملاحظه می‌شود که روش SVM به‌خوبی نتوانسته است رویه‌ی پاسخ را برآورد کند و این ضعف در اکثر نواحی x_1 و x_2 دیده می‌شود.



شکل (۴-۲۰)- نمایش سه‌بعدی برآورد (قاب راست) و خطای (قاب چپ) مدل بهینه‌ی روش SVM برای تابع آزمون (۴-۱)

با در نظر گرفتن نتایج داده‌های مدل‌ساز (جدول (۴-۱))، اگرچه روش RF، دارای عملکرد ضعیف‌تری نسبت به روش SVM است، اما مقایسه‌ی دو شکل (۴-۱۹) و (۴-۲۰) عملکرد بهتر روش RF را در برآورد رویه‌ی پاسخ تابع آزمون (۴-۱) برای داده‌های آزمون نشان می‌دهد. همچنین مقایسه‌ی قاب راست دو شکل (۴-۱۹) و (۴-۲۰) با شکل (۴-۱) نیز همین مطلب را تایید می‌کند. علاوه بر این، روش RF، این قابلیت را دارد تا متغیرهای توضیحی بی‌اثر را شناسایی کند که روش SVM فاقد چنین قابلیتی است.

۴-۲ تخمین غلظت کلروفیل-a در پایگاه داده NOMAD

در این زیربخش، عملکرد دو روش RF و SVM به‌منظور برآورد غلظت کلروفیل-a در پایگاه داده NOMAD مورد ارزیابی قرار گرفته است. این پایگاه داده که در فصل دوم به‌طور کامل معرفی گردید، شامل ۶ متغیر توضیحی بوده که هر یک معرف میزان انعکاس نور خروجی از سطح آب در طول موج خاصی می‌باشد. علاوه بر این، بنا به دلایل ذکرشده در بخش ۳-۲-۶، این بررسی در حالتی که متغیرهای توضیحی بر متغیر $R_{rs}(555)$ تقسیم شده است نیز انجام گردیده است. همان‌طور که در بخش ۳-۲-۵ اشاره شد، داده‌های انعکاس طیفی همواره تحت‌الشعاع اثرات ذرات و ناخالصی‌های مختلف موجود در اتمسفر قرار می‌گیرند. لذا روش برآوردیابی مناسب، روشی خواهد بود که در شرایط نوفه‌ای نیز عملکرد مناسبی داشته باشد. به‌همین منظور تمامی ارزیابی‌های صورت گرفته فوق‌الذکر، بر روی پایگاه داده‌ی نوفه‌ای NOMAD نیز انجام شده است. در واقع نتایج تجربی به‌دست آمده از این پایان‌نامه بر روی پایگاه داده NOMAD را می‌توان به چهار بخش تقسیم کرد.

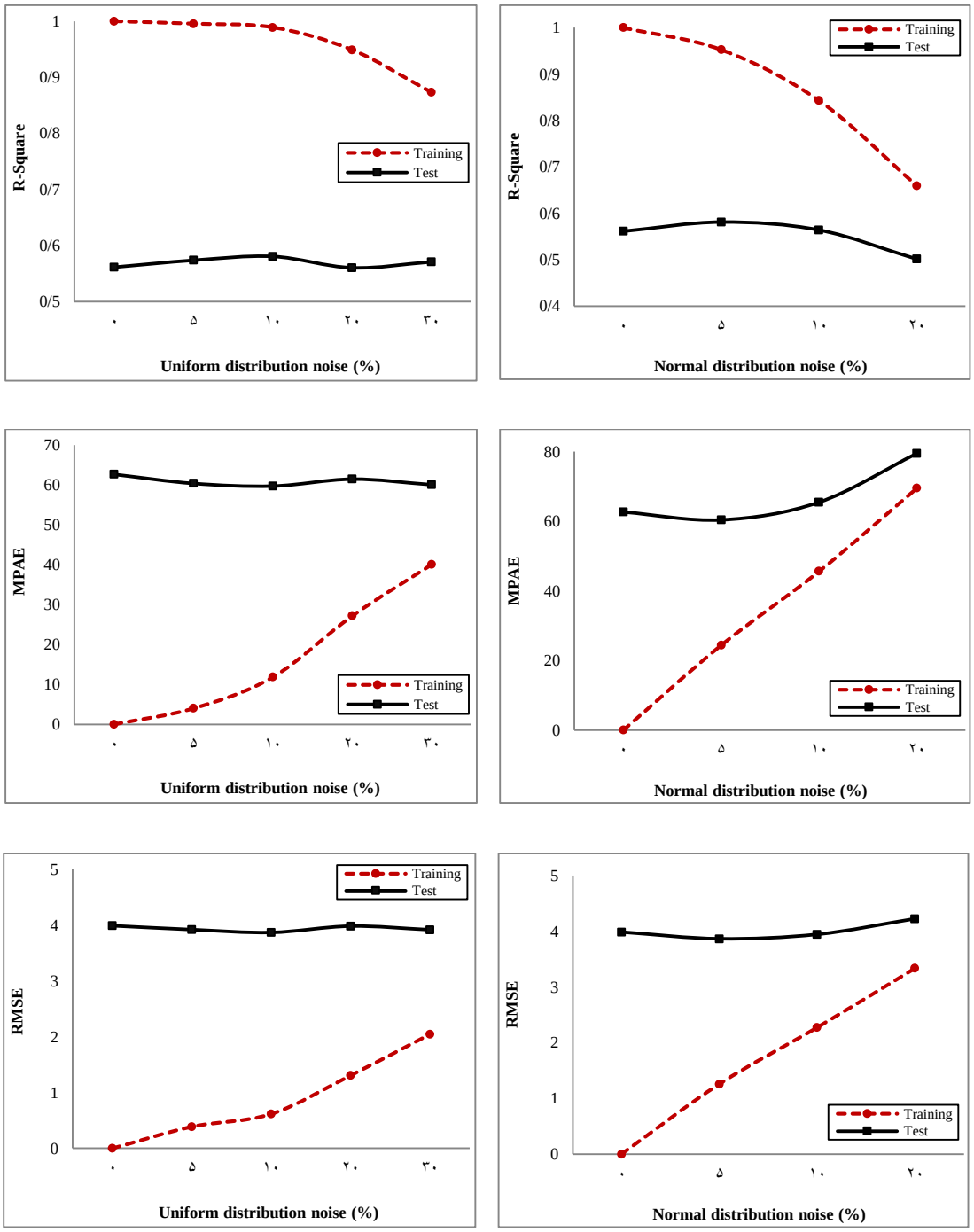
تخمین غلظت کلروفیل-a
- **۱)** با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ با استفاده از روش RF
- **۲)** با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ با استفاده از روش SVM
- **۳)** با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ با استفاده از روش RF
- **۴)** با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ با استفاده از روش SVM

با توجه به سایر تحقیقات انجام‌شده بر روی پایگاه داده‌ی NOMAD، در برآورد غلظت کلروفیل-a، می‌توان گفت که معیار مورد توجه محققین برای این داده‌ها، MPAE می‌باشد. لذا در این پایگاه داده، مراحل بهینه‌سازی پارامترهای مدل‌ها، با در نظر گرفتن کمینه‌ی MPAE دنبال شده‌اند.

۴-۲-۱ تخمین غلظت کلروفیل-a توسط روش RF با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ در پایگاه داده NOMAD

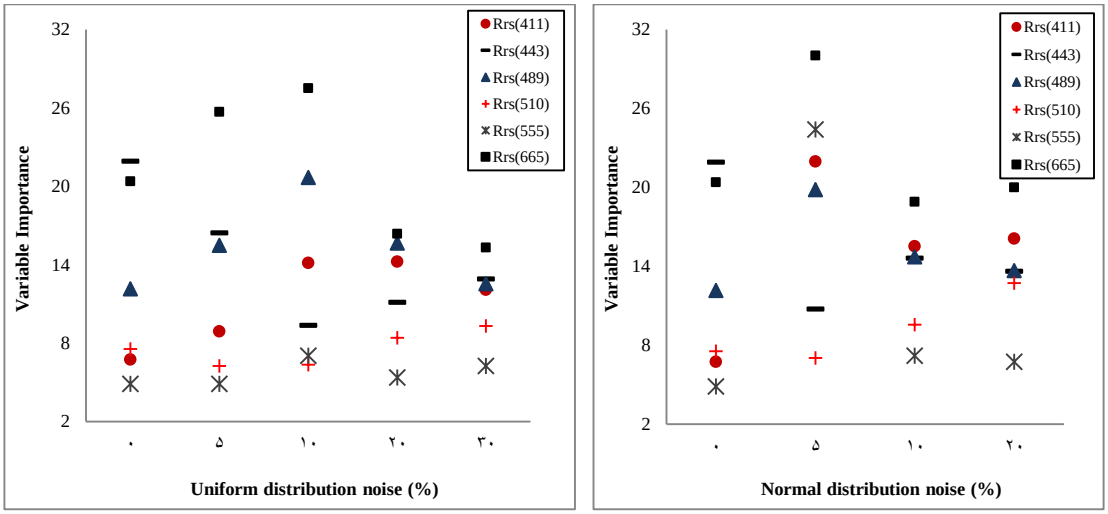
در شکل (۴-۲۱) نتایج روش RF در برآورد غلظت کلروفیل-a با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ در پایگاه داده NOMAD به نمایش درآمده است. در این شکل، مقدار سه معیار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز و آزمون نشان داده شده است. با توجه به این‌که برای هر سه معیار R^2، MPAE و RMSE، نمودار داده‌های آزمون دچار نوسانات اندکی شده است، می‌توان گفت که روش RF در برآورد غلظت کلروفیل-a، نسبت به افزودن نوفه، پایدار عمل می‌کند. البته با مقایسه‌ی قاب‌های چپ و قاب‌های راست می‌توان گفت که این پایداری به‌ازای افزایش مقدار نوفه‌ی یکنواخت بیشتر از نوفه‌ی نرمال است. همچنین با توجه به نمودارهای هر سه معیار، می‌توان گفت که با افزایش مقدار نوفه، نتایج داده‌های مدل‌ساز و داده‌های آزمون به سمت یکدیگر همگرا می‌شوند. در واقع با افزایش مقدار نوفه در داده‌های مدل‌ساز، این داده‌ها حالت جامع‌تری پیدا می-کند و مدل ساخته‌شده با استفاده از این داده‌ها بهتر می‌تواند رفتار داده‌های آزمون را پوشش دهد و در نهایت نتایج برآورد در داده‌های مدل‌ساز و آزمون به سمت یکدیگر همگرا شده است.

معیار MPAE که برای این داده‌ها حایز اهمیت است، به ازای مقادیر مختلف نوفه‌ی یکنواخت، در داده‌های آزمون همواره حدود ۶۰ است. این در حالی است که با افزایش مقدار نوفه‌ی نرمال، مقدار این معیار در داده‌های آزمون از ۶۰ تا ۸۰ در حال افزایش است.



شکل (۴-۲۱)- نتایج روش RF در تخمین غلظت کلروفیل-a برحسب متغیرهای $R_{rs}(\lambda)$ در پایگاه داده NOMAD به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

همان‌طور که در فصل دوم گفته شد، یکی از ویژگی‌های روش RF، تعیین اهمیت متغیرهای توضیحی است. شکل (۴-۲۲) میزان اهمیت هر یک از متغیرهای توضیحی پایگاه داده NOMAD را در مدل بهینه، به ازای مقادیر مختلف نوفه نشان می‌دهد. این بررسی به‌طور جداگانه به ازای نوفه‌های نرمال (قاب راست) و نوفه‌های یکنواخت (قاب چپ) صورت گرفته است. با توجه به این شکل، می‌توان متغیر $R_{rs}(665)$ را به‌عنوان با اهمیت‌ترین متغیر در داده‌های نوفه‌ای معرفی کرد. این در حالی است که در داده‌های بدون نوفه، اهمیت این متغیر در میان متغیرهای توضیحی در رتبه‌ی دوم قرار می‌گیرد (پس از متغیر $R_{rs}(443)$). در حالت کلی نیز می‌توان متغیر $R_{rs}(665)$ را به‌عنوان مهم‌ترین متغیر توضیحی در برآورد غلظت کلروفیل-a در این پایگاه داده معرفی کرد. ضمن این‌که متغیرهای $R_{rs}(555)$ و $R_{rs}(510)$ در بیشتر موارد به‌عنوان کم-اهمیت‌ترین متغیرها شناخته می‌شوند.



شکل (۴-۲۲)-میزان اهمیت متغیرهای $R_{rs}(\lambda)$ در برآورد غلظت کلروفیل-a پایگاه داده NOMAD به ازای مقادیر مختلف نوفه‌های نرمال (قاب راست) و یکنواخت (قاب چپ)

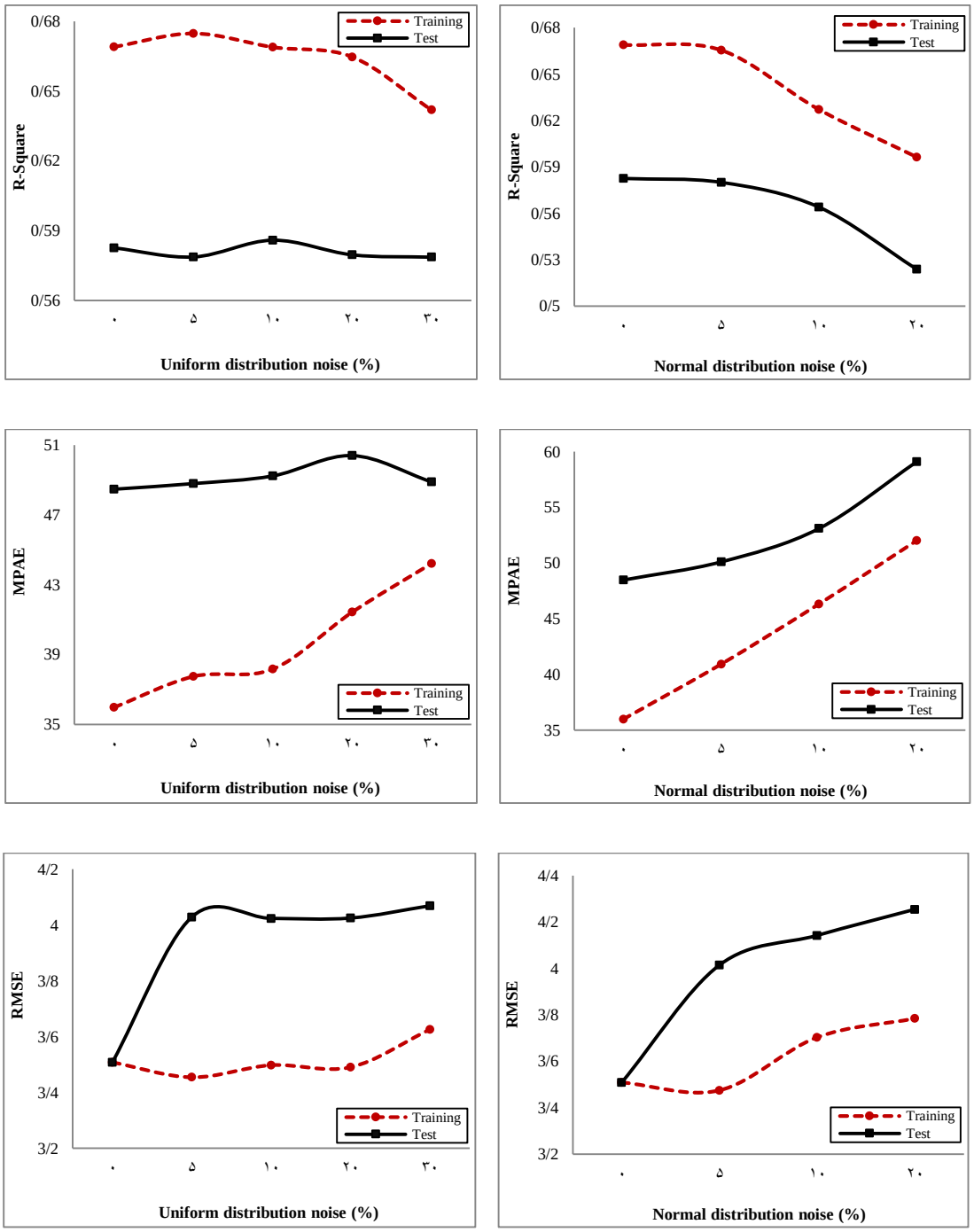
با مقایسه‌ی قاب‌های راست و چپ شکل (۴-۲۲) می‌توان دریافت که متغیر $R_{rs}(555)$ تنها به‌ازای نوفه‌ی نرمال ۵٪ دارای اهمیت بالایی است (دومین متغیر با اهمیت). همچنین به نظر می‌رسد که با افزایش مقدار نوفه، اولویت اهمیت متغیر $R_{rs}(411)$ افزایش می‌یابد.

شایان ذکر است که اندازه‌ی اهمیت یک متغیر، به‌تنهایی قابل تفسیر نمی‌باشد و تنها جهت رتبه‌بندی اهمیت متغیرهای توضیحی به‌کار می‌رود.

۴-۲-۲ تخمین غلظت کلروفیل-a توسط روش SVM با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ در پایگاه داده NOMAD

شکل (۴-۲۳) نتایج روش SVM را در برآورد غلظت کلروفیل-a بر حسب متغیرهای $R_{rs}(\lambda)$ در پایگاه داده NOMAD نشان می‌دهد. در این شکل، مقدار سه معیار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز و آزمون نشان داده شده است. با توجه به روند تغییرات هر سه معیار، می‌توان گفت که روش SVM در برآورد غلظت کلروفیل-a، نسبت به افزودن نوفه، چندان پایدار عمل نمی‌کند. البته با در نظر گرفتن تنها نمودار معیارهای R^2 و MPAE، می‌توان گفت که روش SVM نسبت به افزایش مقدار نوفه‌ی یکنواخت در این دو معیار پایدار است. همچنین با افزایش مقدار نوفه در داده‌های مدل‌ساز، این داده‌ها بهتر می‌تواند داده‌های آزمون را پوشش دهد و مدل ساخته شده با استفاده از داده‌های مدل‌ساز نوفه‌ای بهتر می‌تواند رفتار داده‌های آزمون را برآورد کند و در نهایت نتایج برآورد در داده‌های مدل‌ساز، تا حد محسوسی به سمت نتایج برآورد داده-های آزمون همگرا شده است. میزان برآورد معیار MPAE حاصل‌شده از نتایج این روش، به ازای نوفه‌های یکنواخت، در داده‌های آزمون حدود ۵۰ می‌باشد. این در حالی است که با افزایش مقدار نوفه‌ی نرمال، مقدار این معیار از ۴۸ تا ۵۸ در حال افزایش است.

شایان ذکر است که نتایج بهینه در این پایگاه داده، با به‌کارگیری هسته‌ی نرمال در روش SVM به-دست آمده است.

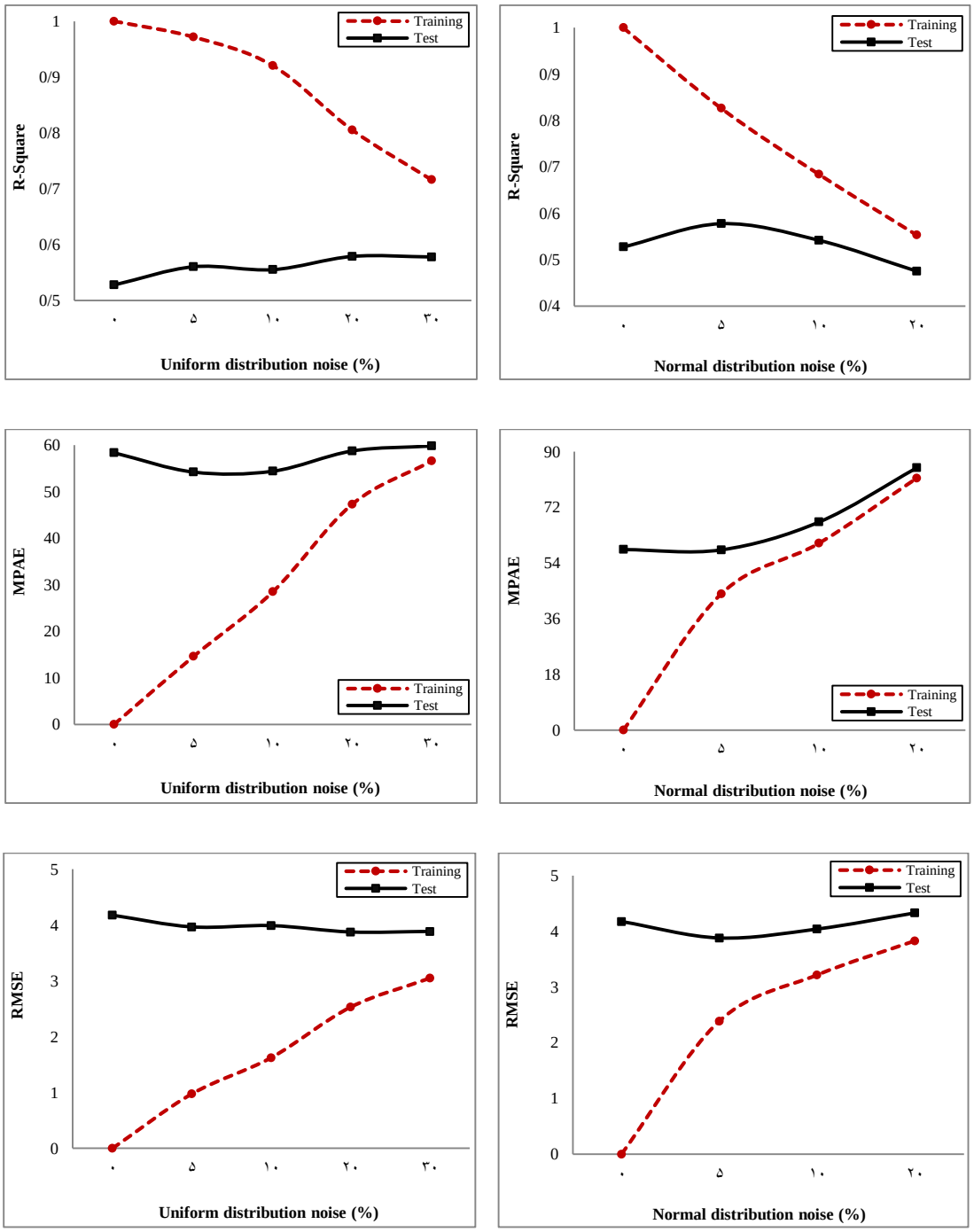


شکل (۴-۲۳)- نتایج روش SVM در تخمین غلظت کلروفیل-a برحسب متغیرهای $R_{rs}(\lambda)$ در پایگاه داده NOMAD به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

۴-۲-۳ تخمین غلظت کلروفیل-a با به‌کارگیری متغیرهای $R_{rs}(\lambda)/R_{rs}(555)$ در پایگاه داده NOMAD با استفاده از روش RF

شکل (۴-۲۴) نتایج روش RF در برآورد غلظت کلروفیل-a بر حسب متغیرهای $R_{rs}(\lambda)/R_{rs}(555)$ در پایگاه داده NOMAD را به نمایش درآورده است. در این شکل، مقدار سه معیار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز و آزمون نشان داده شده است. با توجه به این‌که در هر سه معیار R^2، MPAE و RMSE، نمودار داده‌های آزمون دچار نوسانات اندکی شده است، می‌توان گفت که روش RF در برآورد غلظت کلروفیل-a، نسبت به افزودن نوفه، تا حد زیادی پایدار است. البته با مقایسه‌ی قاب‌های چپ با قاب‌های راست، می‌توان گفت که این پایداری به‌ازای افزایش مقدار نوفه‌ی یکنواخت بیشتر از نوفه‌ی نرمال است. همچنین با توجه به نمودارهای هر سه معیار، می‌توان گفت که با افزایش مقدار نوفه، نتایج داده‌های مدل‌ساز و داده‌های آزمون به سمت یکدیگر همگرا می‌شود. در واقع با افزایش مقدار نوفه در داه‌های مدل‌ساز، مدل ساخته‌شده با استفاده از این داده‌ها بهتر می‌تواند رفتار داده‌های آزمون را پوشش دهد و در نهایت نتایج برآورد در داده-های مدل‌ساز و آزمون به سمت یکدیگر همگرا شده است.

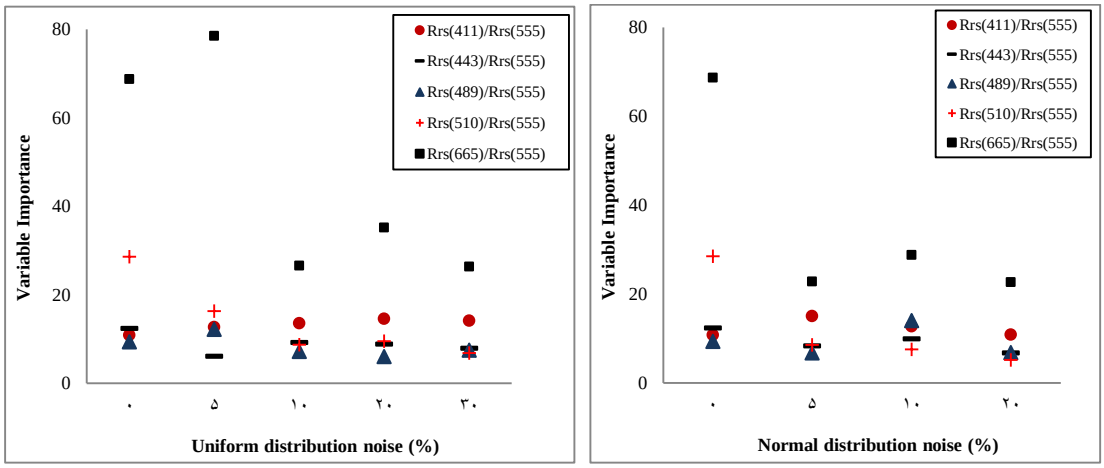
مقدار معیار MPAE به‌دست آمده از برآورد غلظت کلروفیل-a در داده‌های آزمون به ازای نوفه‌های یکنواخت، بین ۵۴ و ۵۹ در نوسان است، در حالی که مقدار این معیار، با افزایش میزان نوفه‌ی نرمال در داده‌های آزمون، بین ۵۸ تا ۸۱ در نوسان می‌باشد.



شکل (۴-۲۴)- نتایج روش RF در تخمین غلظت کلروفیل-a بر حسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده NOMAD به ازای مقادیر مختلف نوفه‌ی نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

شکل (۴-۲۵) میزان اهمیت هر یک از متغیرهای توضیحی $R_{rs}(\lambda)/ R_{rs}(555)$ را به ازای مقادیر مختلف نوفه‌ی نرمال (قاب راست) و یکنواخت (قاب چپ) نشان می‌دهد. با توجه به شکل (۴-۲۵) می‌توان

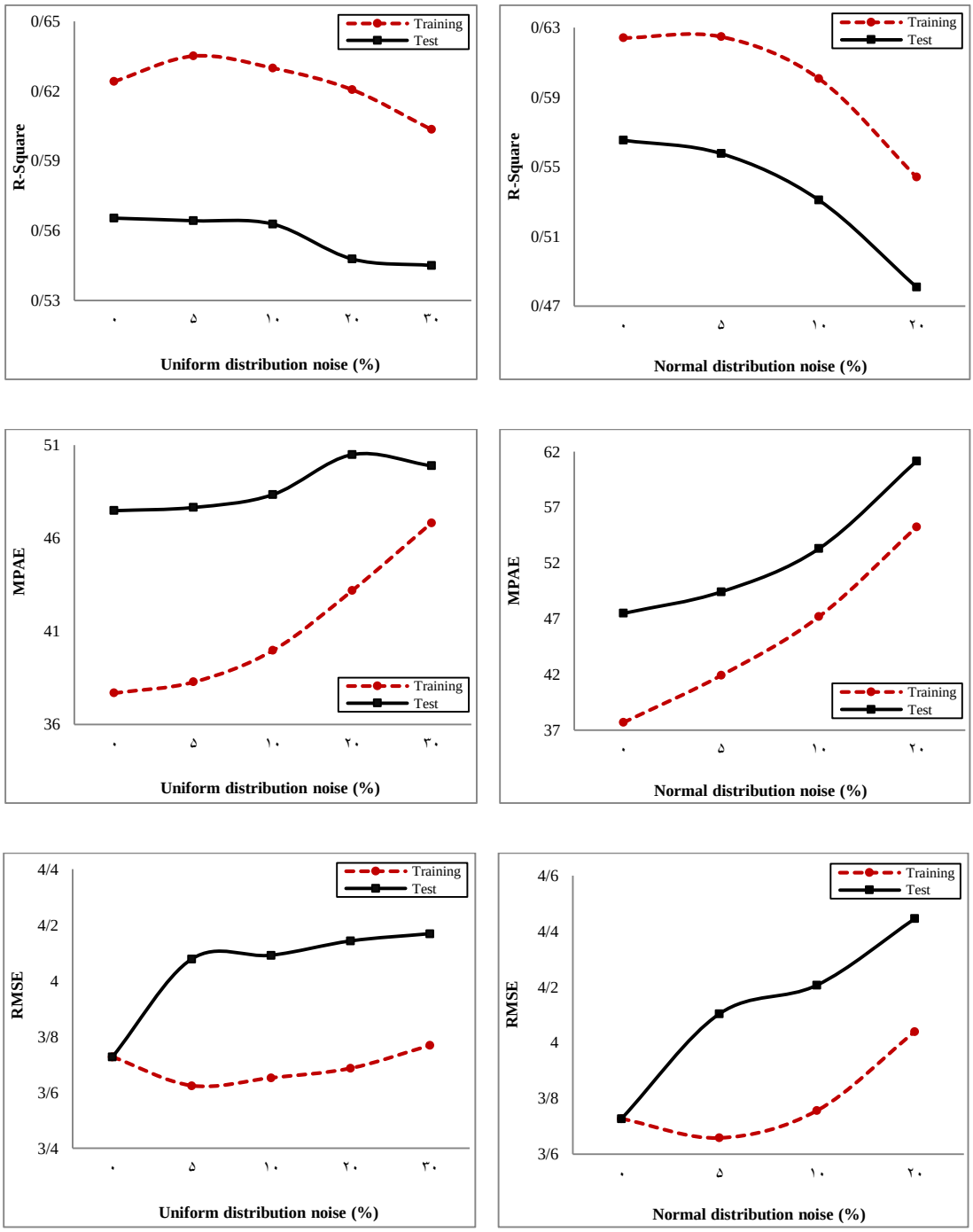
متغیر $R_{rs}(665)/ R_{rs}(555)$ را به‌عنوان با اهمیت‌ترین متغیر معرفی نمود. به‌طور کلی می‌توان گفت که با افزایش مقدار نوفه، از میزان اهمیت این متغیر کاسته می‌شود اما این متغیر همچنان در رتبه‌ی نخست مهم‌ترین متغیرها باقی می‌ماند. همچنین، می‌توان گفت که پس از متغیر $R_{rs}(665)/ R_{rs}(555)$، سایر متغیرها از لحاظ اهمیت، برتری محسوسی نسبت به یکدیگر ندارند.



شکل (۴-۲۵)-میزان اهمیت متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در برآورد غلظت کلروفیل-a پایگاه داده NOMAD به ازای مقادیر مختلف نوفه‌ی نرمال (قاب راست) و یکنواخت (قاب چپ)

۴-۲-۴ تخمین غلظت کلروفیل-a با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده NOMAD با استفاده از روش SVM

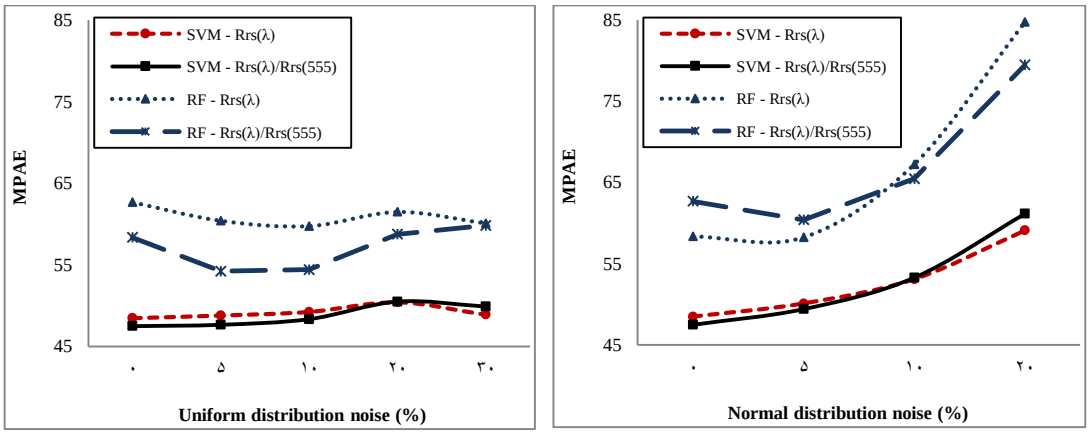
در شکل (۴-۲۶) نتایج به‌دست آمده از روش SVM در برآورد غلظت کلروفیل-a با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در این داده‌ها نشان داده شده است. با توجه به روند تغییرات هر سه معیار R^2، MPAE و RMSE، می‌توان گفت که روش SVM در برآورد غلظت کلروفیل-a، نسبت به افزودن نوفه، چندان پایدار عمل نمی‌کند. همچنین مقدار معیار MPAE به‌دست آمده از برآورد غلظت کلروفیل-a در داده‌های آزمون به ازای مقادیر مختلف نوفه‌ی یکنواخت، بین ۴۷ و ۵۰ در نوسان است. در حالی‌که مقدار این معیار، با افزایش میزان نوفه‌ی نرمال در داده‌های آزمون، از ۴۷ تا ۶۱ در حال افزایش است. نتایج بهینه‌ی در این پایگاه داده، به‌ازای هسته‌ی نرمال به‌دست آمده است.



شکل (۴-۲۶)- نتایج روش SVM در تخمین غلظت کلروفیل-a برحسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده NOMAD به ازای مقادیر مختلف نوفه‌ی نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

۴-۲-۵ مقایسه‌ی نتایج به‌دست آمده در پایگاه داده NOMAD

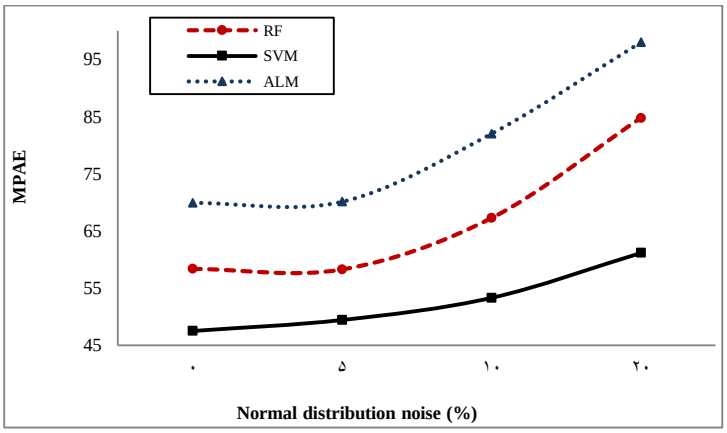
با مقایسه‌ی شکل‌های (۴-۲۱) با (۴-۲۴) و همچنین مقایسه‌ی شکل‌های (۴-۲۳) با (۴-۲۶) این نکته مشخص می‌گردد که به‌طور کلی، با به‌کارگیری هرکدام از روش‌های RF و SVM در برآورد غلظت کلروفیل-a در پایگاه داده NOMAD، استفاده از متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ منجر به نتایج بهتری نسبت به استفاده از متغیرهای $R_{rs}(\lambda)$ به‌عنوان متغیرهای توضیحی می‌گردد. شکل (۴-۲۷) روند معیار MPAE را برحسب مقدار نوفه‌ی نرمال (قاب راست) و نوفه‌ی یکنواخت (قاب چپ) در هر دو روش نشان می‌دهد. با مقایسه‌ی قاب چپ با قاب راست، این نتیجه حاصل می‌گردد که هر دو روش RF و SVM نسبت به افزایش نوفه‌ی یکنواخت دارای پایداری بیشتری نسبت به افزایش نوفه‌ی نرمال هستند. همچنین در تمامی موارد عملکرد روش SVM نسبت به روش RF بهتر بوده است.



شکل (۴-۲۷)- مقایسه‌ی نتایج روش‌های RF و SVM در تخمین غلظت کلروفیل-a برحسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ و$R_{rs}(\lambda)$ در داده‌های آزمون پایگاه داده NOMAD به ازای مقادیر مختلف نوفه‌ی نرمال (قاب راست) و یکنواخت (قاب چپ)

به منظور ارزیابی بهتر نتایج حاصل از این دو روش بر روی این پایگاه داده، نتایج روش ALM (طاهری شهرآیینی و همکاران، ۲۰۰۹) را با نتایج حاصل از این دو روش مورد مقایسه قرار می‌دهیم. با توجه به این‌که طاهری شهرآیینی و همکاران (۲۰۰۹)، در پژوهش خود تنها نوفه‌ی نرمال را به‌منظور تخمین غلظت کلروفیل-a بر حسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ به‌کار برده‌اند، لذا شکل (۴-۲۸) بر

همین اساس تنظیم گردیده است. این شکل عملکرد سه روش RF، SVM و ALM را در برآورد غلظت کلروفیل-a داده‌های آزمون پایگاه داده NOMAD نشان می‌دهد. می‌توان ملاحظه کرد که روش SVM به ازای هر مقدار از نوفه‌ی نرمال نسبت به دو روش دیگر دارای خطای کمتری در برآورد پارامتر مورد نظر است. همچنین روش RF نیز از لحاظ دقت برآورد در رتبه‌ی دوم قرار می‌گیرد. همچنین ملاحظه می‌شود که پایداری روش SVM نسبت به افزایش مقدار نوفه، نسبت به دو روش دیگر بیشتر می‌باشد.



شکل (۴-۲۸)- مقایسه‌ی نتایج روش‌های RF ، SVM و ALM در تخمین غلظت کلروفیل-a برحسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده NOMAD به ازای مقادیر مختلف نوفه‌ی نرمال

۴-۳ تخمین غلظت رنگدانه در پایگاه داده SeaBAM

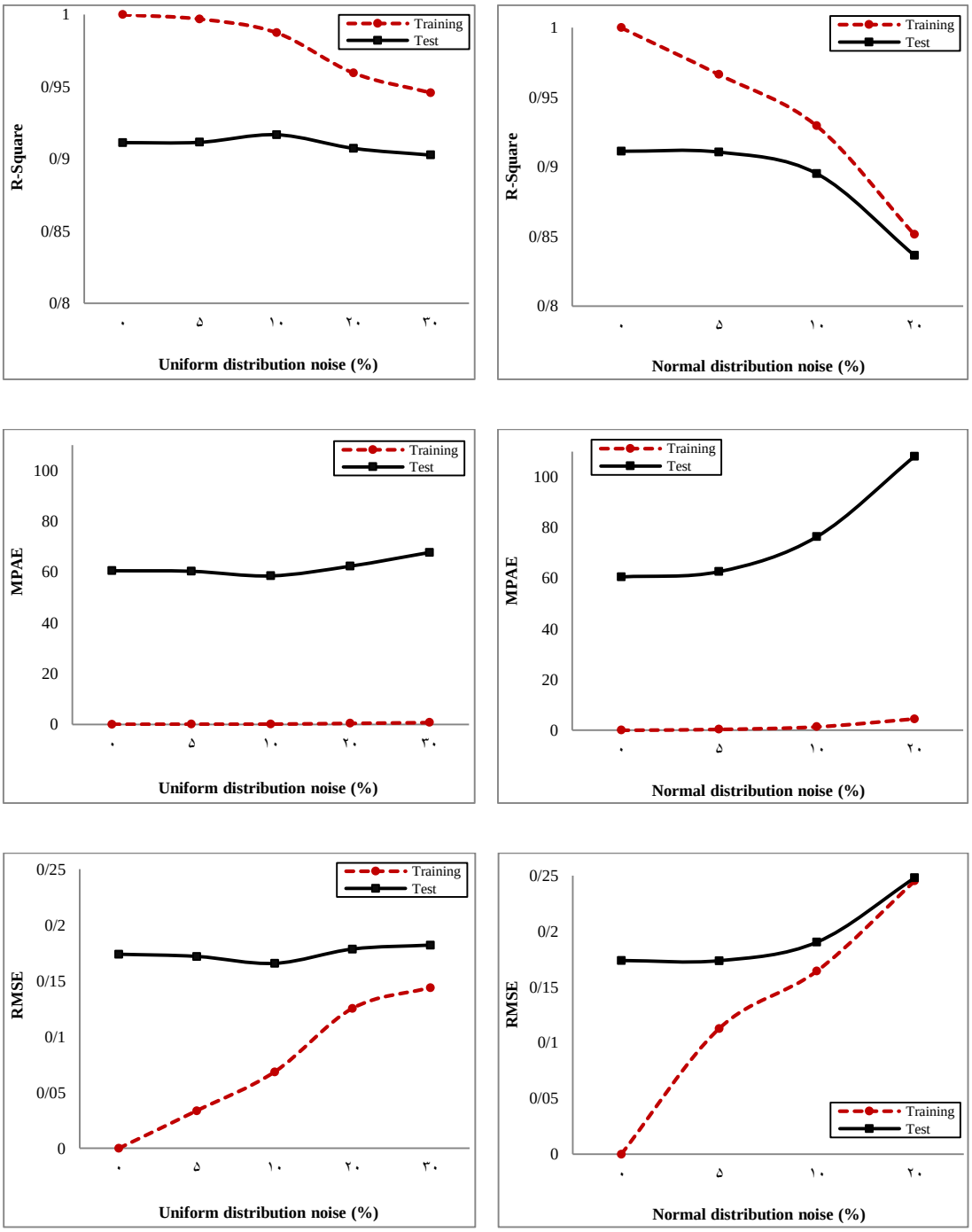
در این زیربخش، عملکرد دو روش RF و SVM به‌منظور برآورد غلظت رنگدانه در پایگاه داده SeaBAM مورد ارزیابی قرار گرفته است. این پایگاه داده شامل ۵ متغیر توضیحی است که هر یک معرف میزان انعکاس نور خروجی از سطح آب در طول موج خاصی می‌باشد. علاوه بر این، این بررسی در حالتی که متغیرهای توضیحی بر متغیر $R_{rs}(555)$ تقسیم شده است، نیز انجام گردیده است. شایان ذکر است که تمامی این ارزیابی‌ها به ازای افزودن نوفه‌های مختلف نیز انجام شده است. در واقع نتایج تجربی به‌دست آمده از این پایان‌نامه بر روی پایگاه داده SeaBAM را می‌توان به چهار بخش تقسیم کرد.

تخمین غلظت رنگدانه
- **۱)** با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ با استفاده از روش RF
- **۲)** با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ با استفاده از روش SVM
- **۳)** با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ با استفاده از روش RF
- **۴)** با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ با استفاده از روش SVM

با بررسی سایر تحقیقات انجام شده بر روی پایگاه داده‌ی SeaBAM، می‌توان دریافت که معیارهای قابل توجه پژوهش‌گران در این پایگاه داده، R^2 و RMSE می‌باشد. لذا در این پایان‌نامه مراحل بهینه‌سازی پارامترهای مدل در این پایگاه داده، با در نظر گرفتن کمینه‌ی RMSE و بیشینه‌ی R^2 انجام می‌گیرد.

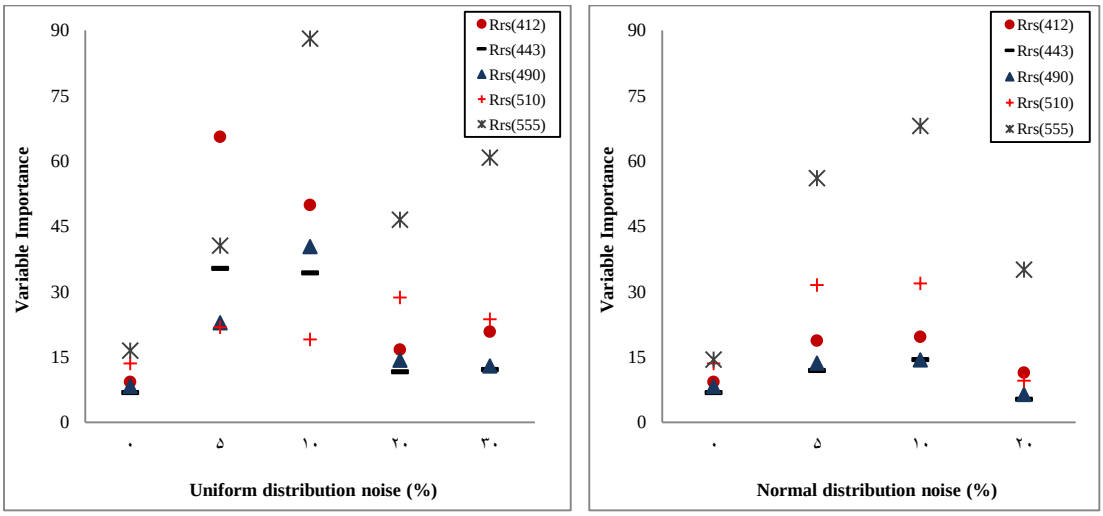
۴-۳-۱ تخمین غلظت رنگدانه توسط روش RF با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ در پایگاه داده SeaBAM

در شکل (۴-۲۹) نتایج به‌دست آمده از روش RF در برآورد غلظت رنگدانه برحسب متغیرهای $R_{rs}(\lambda)$ در پایگاه داده SeaBAM نشان داده شده است. با توجه به روند تغییرات هر سه معیار R^2، MPAE و RMSE، می‌توان گفت که روش RF در برآورد غلظت رنگدانه، نسبت به افزایش مقدار نوفه‌ی یکنواخت، تا حد محسوسی پایدار عمل می‌کند. این در حالی است که این پایداری به ازای افزایش نوفه‌ی نرمال وجود ندارد. همچنین با در نظر گرفتن دو معیار R^2 و RMSE، ملاحظه می‌شود که با افزایش مقدار نوفه (نرمال و یکنواخت)، نتایج داده‌های مدل‌ساز به سمت نتایج داده‌های آزمون همگرا می‌گردد. مشاهده می‌شود که معیار RMSE، به ازای افزایش مقدار نوفه‌ی یکنواخت، از ۰/۱۷ به ۰/۱۸ افزایش یافته است و به ازای افزایش مقدار نوفه‌ی نرمال، از ۰/۱۷ به ۰/۲۵ کاهش یافته است. بدین ترتیب می‌توان گفت که عملکرد این روش در مقادیر زیاد نوفه، به ازای نوفه‌ی یکنواخت بهتر از نوفه‌ی نرمال است.



شکل (۴-۲۹)- نتایج روش RF در تخمین غلظت رنگدانه برحسب متغیرهای $R_{rs}(\lambda)$ در پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

در این بخش با بررسی نتایج روش RF، میزان اهمیت هر یک از متغیرهای توضیحی در این پایگاه داده مشخص گردیده است. شکل (۴-۳۰) میزان اهمیت هر یک از متغیرهای $R_{rs}(\lambda)$ پایگاه داده SeaBAM را در برآورد غلظت رنگدانه به ازای مقادیر مختلف نوفه نشان می‌دهد. این بررسی به‌طور جداگانه به ازای نوفه‌های نرمال (قاب چپ) و یکنواخت (قاب راست) صورت گرفته است. با توجه به شکل، ملاحظه می‌شود که در اکثر موارد، متغیر $R_{rs}(555)$ به‌عنوان مهم‌ترین متغیر توضیحی در مدل معرفی شده است. در واقع تنها در داده‌های با ۵٪ نوفه‌ی یکنواخت، این متغیر در رتبه‌ی دوم با اهمیت‌ترین متغیرها قرار می‌گیرد. همچنین ملاحظه می‌گردد که به‌طور کلی می‌توان متغیرهای $R_{rs}(443)$ و $R_{rs}(490)$ را به‌عنوان کم‌اهمیت‌ترین متغیرها در برآورد غلظت رنگدانه در این پایگاه داده شناسایی کرد. این در حالی است که اهمیت این دو متغیر به ازای نوفه‌های ۵٪ و ۱۰٪ یکنواخت تا حدی در اولویت بالاتری قرار گرفته است. همچنین با مقایسه‌ی قاب‌های چپ و راست شکل (۴-۳۰)، ملاحظه می‌شود که اهمیت متغیر $R_{rs}(412)$ تا حدی به ازای نوفه‌های یکنواخت بیشتر از نوفه‌های نرمال است.

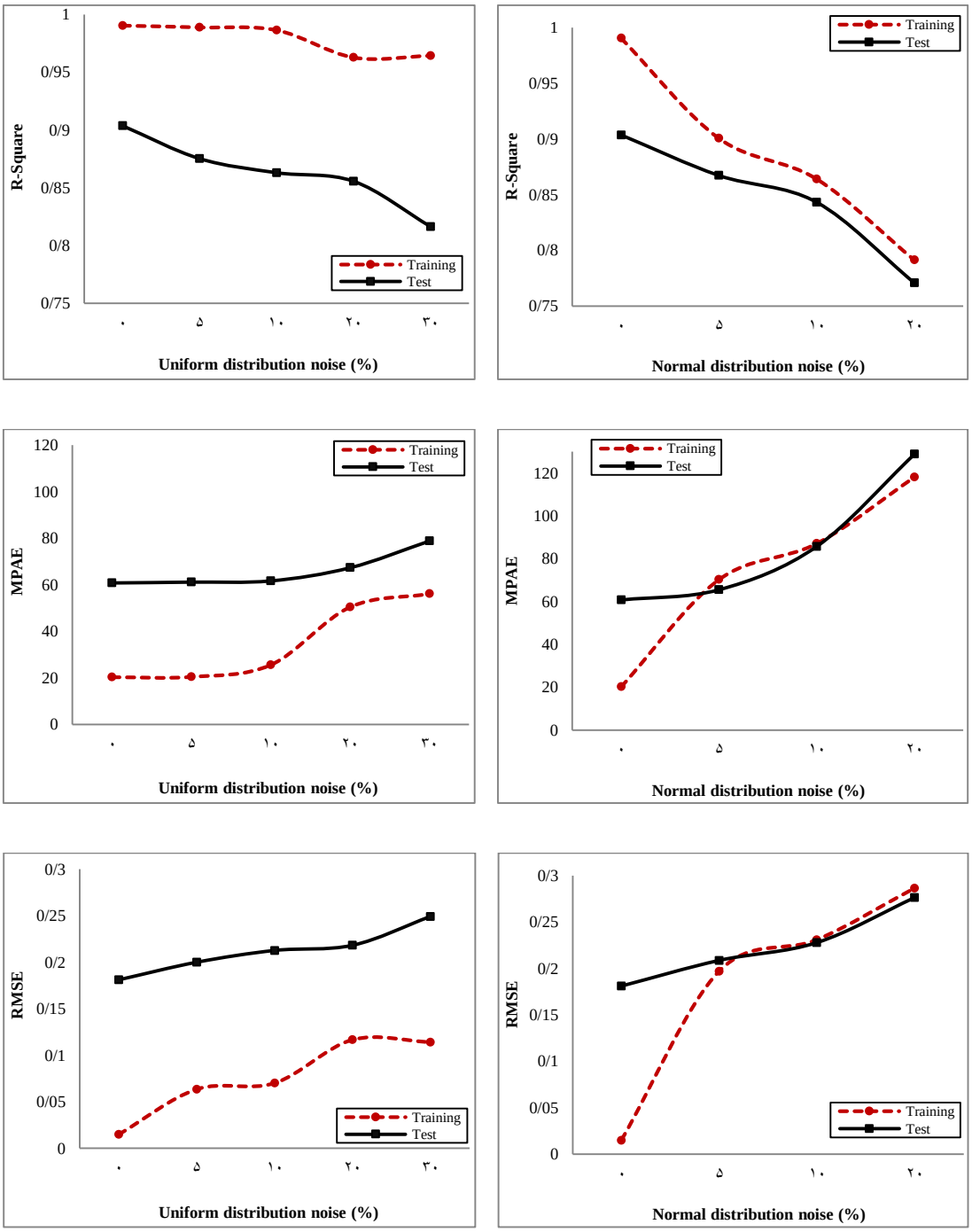


شکل (۴-۳۰)-میزان اهمیت متغیرهای $R_{rs}(\lambda)$ در برآورد غلظت رنگدانه پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب راست) و یکنواخت (قاب چپ)

۴-۳-۲ تخمین غلظت رنگدانه توسط روش SVM با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ در پایگاه داده SeaBAM

شکل (۴-۳۱) نتایج به‌دست آمده از روش SVM در برآورد غلظت رنگدانه با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ در پایگاه داده SeaBAM را به نمایش در آورده است. در این شکل، مقدار سه معیار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز و آزمون نشان داده شده است. با مقایسه‌ی قاب‌های چپ با قاب‌های راست می‌توان گفت که روش SVM در برآورد غلظت رنگدانه، به‌ازای افزایش مقدار نوفه‌ی یکنواخت، نسبت به افزایش مقدار نوفه‌ی نرمال، پایدارتر عمل می‌کند. در واقع با افزایش مقدار نوفه‌ی نرمال، پایداری قابل توجهی در نتایج دیده نمی‌شود. همچنین افزایش مقدار نوفه‌ی نرمال، موجب نزدیک شدن نتایج داده‌های مدل‌ساز و آزمون به یکدیگر شده است. این در حالی است که این همگرایی به‌ازای افزایش نوفه‌ی یکنواخت وجود ندارد. همچنین با توجه به قاب راست-پایین می‌توان گفت که مقدار RMSE در داده‌های آزمون و داده‌های مدل‌ساز بسیار نزدیک به هم هستند. با توجه به شکل (۴-۳۱)، با افزایش مقدار نوفه، دقت برآورد در داده‌های آزمون کاهش یافته است که این کاهش به ازای نوفه‌ی نرمال چشمگیرتر از نوفه‌ی یکنواخت است. همچنین RMSE به ازای افزایش مقدار نوفه، افزایش یافته است که نشان دهنده‌ی افزایش خطا به ازای افزایش مقدار نوفه می‌باشد. به‌طور کلی مقدار ضریب تعیین در داده‌های آزمون به ازای افزایش مقدار نوفه‌ی نرمال، از ۹۰٪ به ۷۷٪ کاهش می‌یابد. این در حالی است که به ازای افزایش مقدار نوفه‌ی یکنواخت، این معیار از ۹۰٪ به ۸۲٪ کاهش یافته است.

توجه کنید که نتایج بهینه در این پایگاه داده، با به‌کارگیری هسته‌ی نرمال در روش SVM به‌دست آمده است.

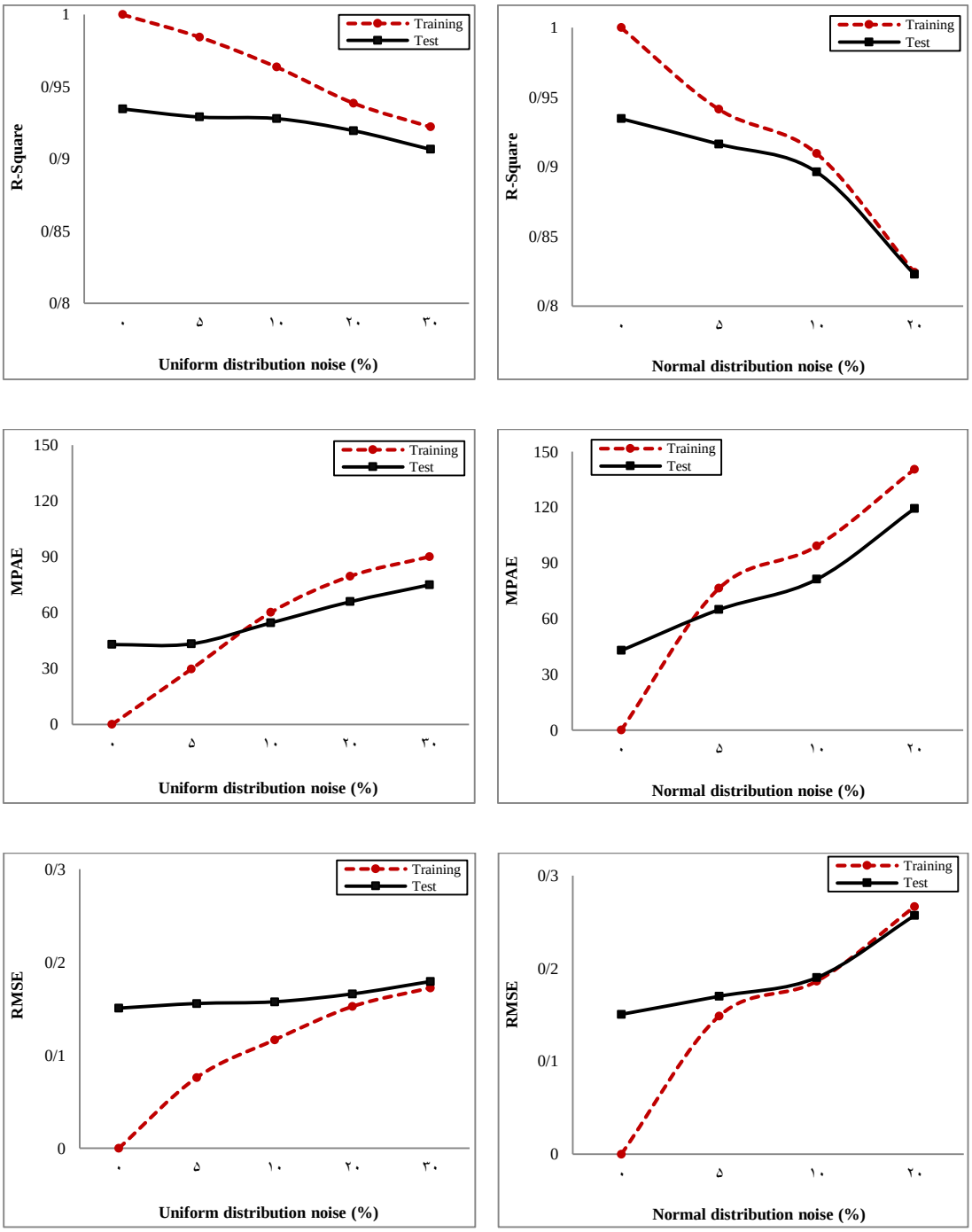


شکل (۴-۳۱)- نتایج روش SVM در تخمین غلظت رنگدانه برحسب متغیرهای $R_{rs}(\lambda)$ در پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

۴-۳-۳ تخمین غلظت رنگدانه با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده SeaBAM با استفاده از روش RF

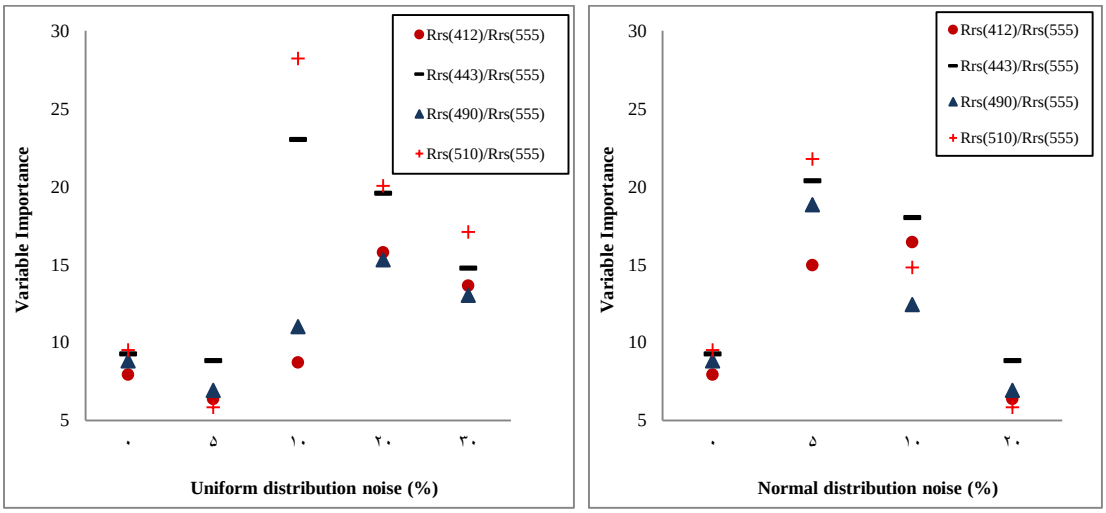
شکل (۴-۳۲) نتایج روش RF را در برآورد غلظت رنگدانه برحسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده SeaBAM به نمایش در آورده است. در این شکل، مقدار سه معیار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز و آزمون نشان داده شده است. با توجه به نمودارهای هر سه معیار R^2، MPAE و RMSE، می‌توان گفت که روش RF در برآورد غلظت رنگدانه، به‌ازای افزایش مقدار نوفه‌ی یکنواخت، پایدار عمل می‌کند. این در حالی است که این پایداری به ازای افزایش مقدار نوفه‌ی نرمال وجود ندارد. همچنین ملاحظه می‌شود که افزایش مقدار نوفه (نرمال و یکنواخت)، موجب نزدیک شدن نتایج داده‌های مدل‌ساز و آزمون به یکدیگر شده است. در واقع می‌توان گفت که نتایج داده‌های مدل‌ساز و داده‌های آزمون، به ازای افزایش نوفه، به یکدیگر همگرا می‌شوند.

همچنین ملاحظه می‌گردد که مقدار معیار R^2 در برآورد غلظت رنگدانه‌ی داده‌های آزمون به ازای افزایش مقدار نوفه‌ی یکنواخت، از ۹۳٪ به ۹۰٪ کاهش می‌یابد. در حالی که مقدار این معیار، با افزایش میزان نوفه‌ی نرمال در داده‌های آزمون، از ۹۳٪ تا ۸۲٪ در حال کاهش است. به‌طور کلی می‌توان گفت که در مقادیر زیاد نوفه، روش RF به‌ازای نوفه‌ی یکنواخت بهتر از نوفه‌ی نرمال عمل کرده است.



شکل (۴-۳۲)- نتایج روش RF در تخمین غلظت رنگدانه برحسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

با بررسی میزان اهمیت متغیرهای توضیحی در این پایگاه داده، رتبه‌بندی اهمیت هر یک از متغیرها به‌ازای مقادیر مختلف نوفه مشخص گردید. شکل (۴-۳۳) نتایج بررسی اهمیت متغیرهای توضیحی این پایگاه داده را به ازای مقادیر مختلف نوفه‌های نرمال و یکنواخت نشان می‌دهد. با توجه به شکل، ملاحظه می‌شود که نمی‌توان متغیر خاصی را با قاطعیت از لحاظ اهمیت معرفی نمود. با این وجود با توجه به وضعیت نسبی اهمیت سایر متغیرها، به‌طور کلی، می‌توان گفت که متغیرهای $R_{rs}(443)/R_{rs}(555)$ و $R_{rs}(510)/R_{rs}(555)$ تا حدی از اهمیت بیشتری نسبت به سایر متغیرها برخوردار می‌باشند. همچنین به‌نظر می‌رسد که به‌ازای نوفه‌ی یکنواخت، شدت میزان اهمیت متغیرها نسبت به یکدیگر بیشتر است. این در حالی است که به‌ازای نوفه‌ی نرمال، مقدار مقیاس اهمیت متغیرهای مختلف، تا حدی به هم نزدیک می‌باشند.



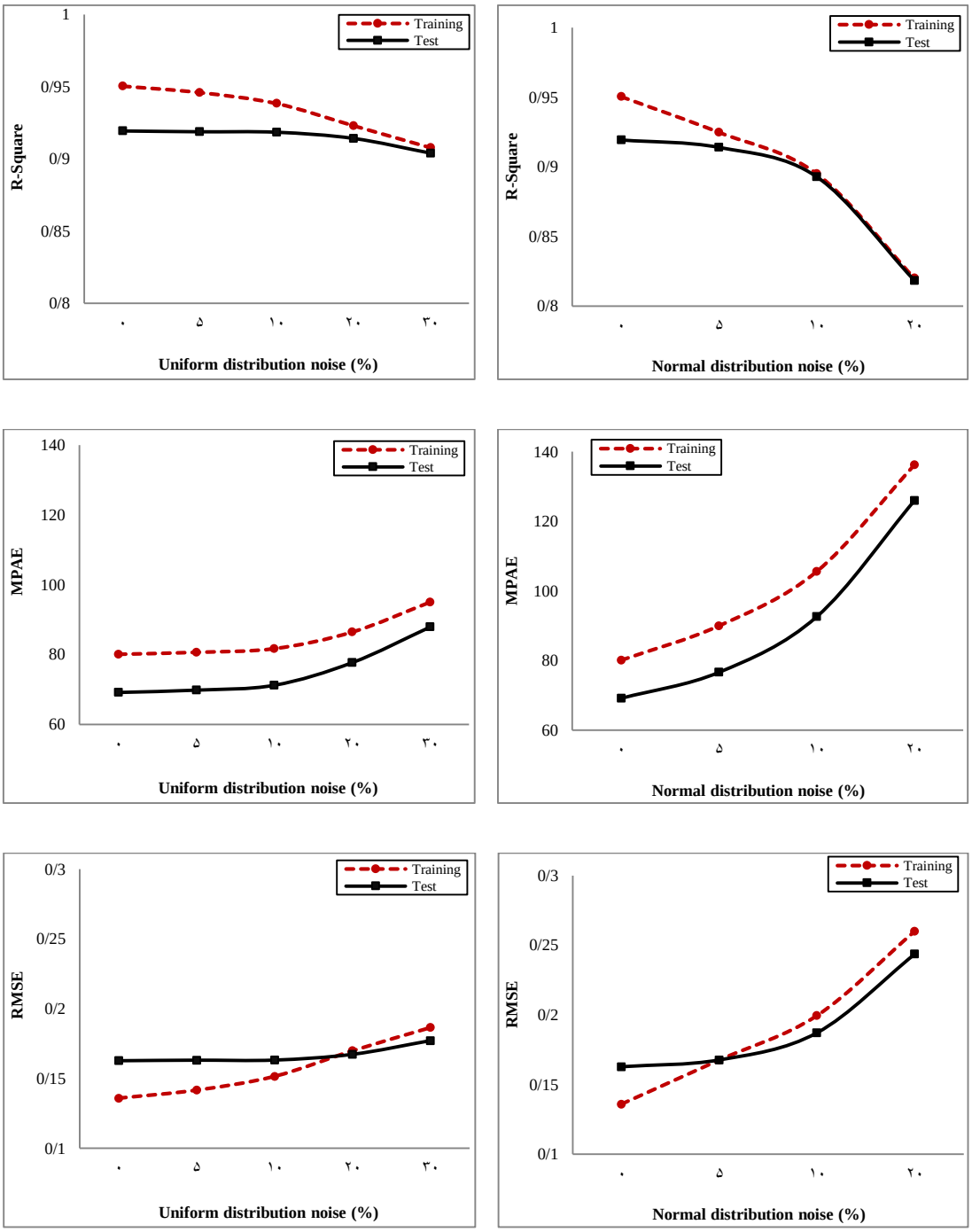
شکل (۴-۳۳)-میزان اهمیت متغیرهای $R_{rs}(\lambda)/R_{rs}(555)$ در برآورد غلظت رنگدانه پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب راست) و یکنواخت (قاب چپ)

ژانگ و همکاران (۲۰۰۳) نشان دادند که مدل‌سازی با استفاده از متغیر $R_{rs}(443)/ R_{rs}(555)$ می‌تواند منجر به نتایج بسیار مناسبی گردد. همچنین متغیر $R_{rs}(443)/ R_{rs}(555)$ که به‌عنوان نسبت آبی به قرمز شناخته می‌شود، برای استخراج غلظت کلروفیل در آب‌های اُلیگوترف[1] (غلظت کلروفیل پایین) به‌دفعات توسط محققین مختلف استفاده شده است (مثل هیم[2] و همکاران، ۲۰۰۵؛ ایلوز[3] و همکاران، ۲۰۰۳؛ اریلی و همکاران، ۱۹۹۸؛ فیشر و کرونفلد[4]، ۱۹۹۰ و گوردون[5] و همکاران، ۱۹۸۳). بنابراین روش RF به‌نحو مطلوبی توانسته است که متغیرها را از لحاظ اهمیت رتبه‌بندی کند. در واقع این روش، توانایی تشخیص متغیرهای مهم، از بین مجموعه متغیرهای معرفی‌شده به مدل را دارا می‌باشد.

۴-۳-۴ تخمین غلظت رنگدانه با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده SeaBAM با استفاده از روش SVM

شکل (۴-۳۴) نتایج روش SVM را در برآورد غلظت رنگدانه بر حسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده SeaBAM به نمایش در آورده است. در این شکل، مقدار سه معیار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز و آزمون نشان داده شده است. با توجه به قاب‌های چپ، ملاحظه می‌شود که روش SVM در برآورد غلظت رنگدانه، نسبت به افزایش مقدار نوفه‌ی یکنواخت، پایدار عمل می‌کند. به‌طور کلی می‌توان گفت که در مقادیر زیاد نوفه، روش SVM به‌ازای نوفه‌ی یکنواخت بهتر از نوفه‌ی نرمال عمل کرده است. شایان ذکر است که نتایج بهینه در این پایگاه داده، با به‌کارگیری هسته‌ی نرمال در روش SVM به‌دست آمده است.

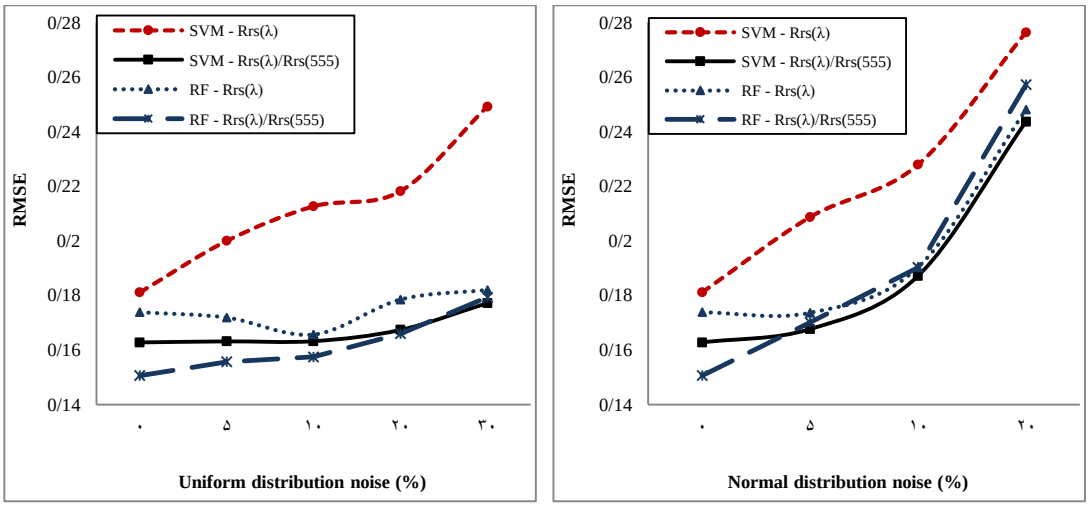
[1] Oligotroph
[2] Heim B.
[3] Iluz D.
[4] Kronfeld U.
[5] Gordon H. R.



شکل (۴-۳۴)- نتایج روش SVM در تخمین غلظت رنگدانه بر حسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

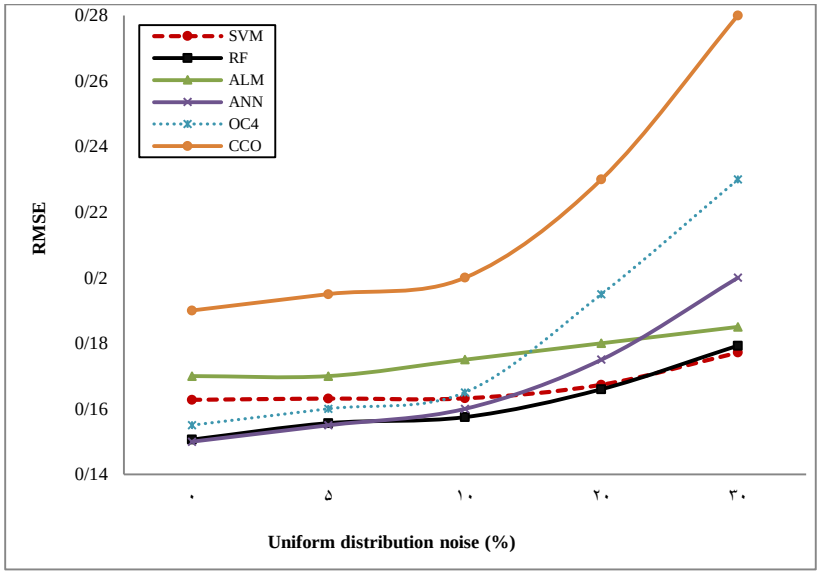
۴-۳-۵ مقایسه‌ی نتایج به‌دست آمده در پایگاه داده SeaBAM

شکل (۴-۳۵) روند معیار RMSE را برحسب مقدار نوفه‌ی نرمال (قاب راست) و نوفه‌ی یکنواخت (قاب چپ) به‌منظور برآورد غلظت رنگدانه به‌ازای داده‌های آزمون پایگاه داده SeaBAM نشان می‌دهد. ملاحظه می‌شود که در هر دو روش RF و SVM، همواره به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ به‌عنوان متغیرهای توضیحی، منجر به نتایج بهتری نسبت به استفاده از متغیرهای $R_{rs}(\lambda)$ می‌گردد. با مقایسه‌ی قاب چپ با قاب راست در شکل (۴-۳۵)، مشاهده می‌شود که هر دو روش به‌کار گرفته شده در برآورد غلظت رنگدانه، به‌ازای افزایش مقدار نوفه‌ی یکنواخت، نسبت به افزایش مقدار نوفه‌ی نرمال، پایدارتر عمل می‌کنند. همچنین با توجه به این شکل، می‌توان گفت که به‌طور کلی عملکرد روش RF بهتر از روش SVM بوده است. با توجه به ۴ نمودار RMSE ثبت شده در هریک از قاب‌های شکل زیر، مشاهده می‌شود که روش SVM با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ منجر به ضعیف‌ترین نتیجه شده است.



شکل (۴-۳۵)- مقایسه‌ی نتایج روش‌های RF و SVM در تخمین غلظت رنگدانه برحسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ و$R_{rs}(\lambda)$ در داده‌های آزمون پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب راست) و یکنواخت (قاب چپ)

همان‌طور که در فصل اول گفته شد، پیش از این برخی از محققین با استفاده از سایر روش‌ها اقدام به تخمین غلظت رنگدانه در این پایگاه داده کرده‌اند. این تحقیقات با افزودن نوفه‌ی یکنواخت و با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ انجام گرفته‌اند. از این‌رو، شکل (۴-۳۶) عملکرد روش‌های RF، SVM، ALM، ANN، OC4 و CCO را در برآورد غلظت رنگدانه در پایگاه داده SeaBAM با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ نشان می‌دهد. با توجه به این شکل، ملاحظه می‌شود که دو روش RF و SVM به ازای افزایش مقدار نوفه‌ی یکنواخت، بسیار پایدار عمل کرده‌اند. این در حالی است که سه روش ANN ، OC4 و CCO این پایداری را ندارند. همچنین مشاهده می‌شود که روش RF در بین این ۶ روش، تقریباً همواره بهترین عملکرد را دارا می‌باشد. ملاحظه می‌گردد که دو روش RF و SVM، به‌ازای مقادیر زیاد نوفه، در برآورد غلظت رنگدانه، دارای بهترین عملکرد می‌باشند.



شکل (۴-۳۶)- ارزیابی نتایج روش‌های RF، SVM، ALM، ANN، OC4 و CCO در تخمین غلظت رنگدانه برحسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده SeaBAM، به ازای مقادیر مختلف نوفه‌ی یکنواخت

۴-۴ تخمین غلظت رنگدانه در پایگاه داده SeaBAM با استفاده از پایگاه داده Synthetic

در بخش قبل به‌منظور برآورد غلظت رنگدانه در پایگاه داده SeaBAM، داده‌های این پایگاه به‌طور تصادفی به دو دسته‌ی داده‌های مدل‌ساز و داده‌های آزمون تقسیم شدند. سپس عملکرد مدل‌های برازش داده‌شده بر روی داده‌های مدل‌ساز، توسط داده‌های آزمون، مورد بررسی قرار گرفت و در نهایت میزان دقت و خطای برآورد مدل‌ها در داده‌های آزمون به‌دست آمد. در این زیربخش، به‌منظور برآورد غلظت رنگدانه در پایگاه داده SeaBAM، از پایگاه داده Synthetic به عنوان داده‌های مدل‌ساز استفاده شده است. درواقع ابتدا با استفاده از داده‌های موجود در پایگاه داده Synthetic، مدل‌های حاصل از روش‌های RF و SVM بهینه می‌شود. سپس مدل‌های بهینه، به‌منظور برآورد غلظت رنگدانه در پایگاه داده SeaBAM مورد آزمون قرار می‌گیرد. شایان ذکر است که هر دو پایگاه داده شامل ۵ متغیر توضیحی می‌باشند که هر متغیر، معرف میزان انعکاس نور خروجی از سطح آب در طول موج خاصی می‌باشد. همچنین، این بررسی در حالتی که متغیرهای توضیحی بر متغیر $R_{rs}(555)$ تقسیم شده است، نیز انجام گردیده است (مشابه زیربخش‌های ۴-۲ و ۴-۳). همچنین تمامی این ارزیابی‌ها به ازای افزودن مقادیر مختلف نوفه‌های نرمال و یکنواخت نیز انجام شده است. در واقع نتایج تجربی حاصل از این بخش را می‌توان به چهار قسمت تقسیم کرد.

تخمین غلظت رنگدانه:

۱) با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ با استفاده از روش RF

۲) با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ با استفاده از روش SVM

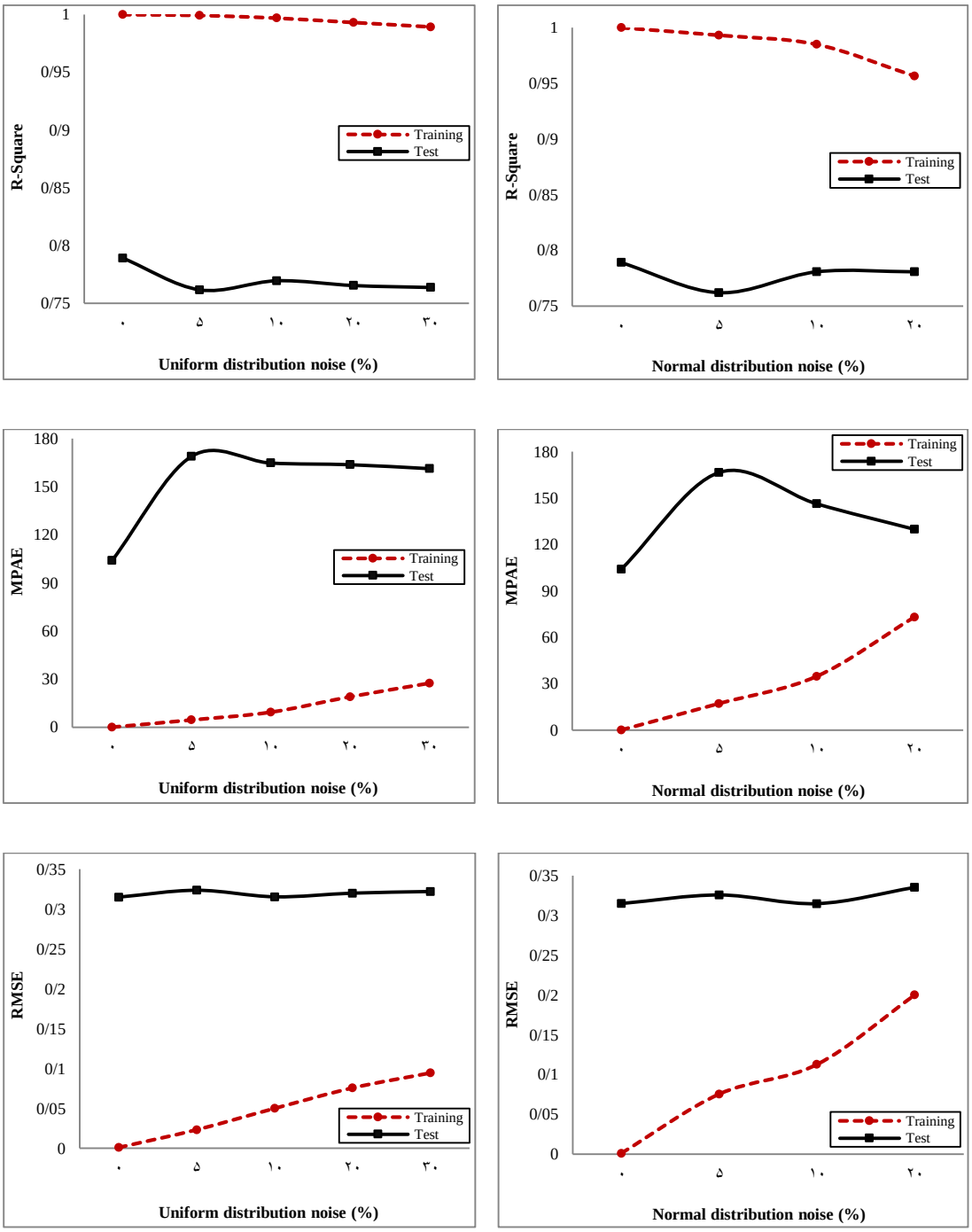
۳) با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ با استفاده از روش RF

۴) با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ با استفاده از روش SVM

با بررسی سایر تحقیقات انجام‌شده بر روی این دو پایگاه داده، می‌توان دریافت که در این دو پایگاه داده، R^2 و RMSE، معیارهای قابل توجه پژوهش‌گران می‌باشد و تمرکز محققین بر روی این دو معیار است. لذا در این پایان‌نامه نیز مراحل بهینه‌سازی پارامترهای مدل در این دو پایگاه داده، با در نظر گرفتن کمینه‌ی RMSE و بیشینه‌ی R^2 انجام می‌گیرد.

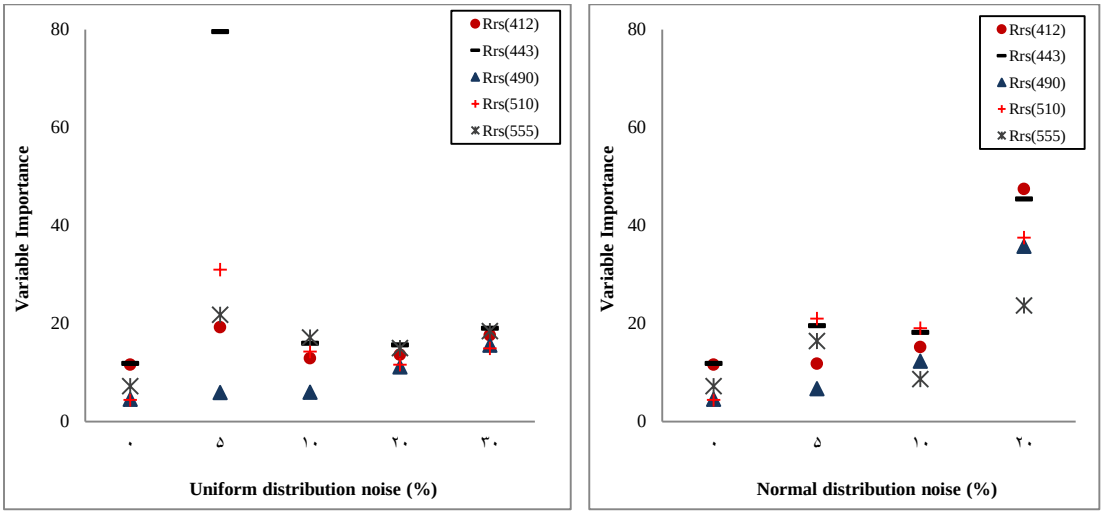
۴-۴-۱ تخمین غلظت رنگدانه در پایگاه داده SeaBAM با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ با استفاده از روش RF برازش‌شده بر روی پایگاه داده Synthetic

شکل (۴-۳۷) نتایج به‌دست آمده از روش RF در برآورد غلظت رنگدانه با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ و با قرار دادن پایگاه داده Synthetic به‌عنوان داده‌های مدل‌ساز و پایگاه داده SeaBAM به‌عنوان داده‌های آزمون را نشان می‌دهد. در این شکل، مقدار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز (Synthetic) و آزمون (SeaBAM) نشان داده شده است. با توجه به این‌که بهینه‌سازی مدل بر اساس معیارهای R^2 و RMSE صورت گرفته است و با در نظر گرفتن نمودارهای مربوط به این دو معیار، می‌توان گفت که روش RF در برآورد غلظت رنگدانه، نسبت به افزایش مقدار نوفه، تا حد محسوسی پایدار عمل کرده است. البته این پایداری، به‌ازای نوفه‌ی یکنواخت بیشتر می‌باشد. با توجه به شکل، ملاحظه می‌شود که ضریب تعیین در برآوردهای حاصل از روش RF در داده‌های مدل‌ساز، به ازای مقادیر مختلف نوفه (نرمال و یکنواخت)، همواره حدود ۱۰۰٪ می‌باشد. این در حالی است که میزان این معیار در داده‌های آزمون همواره حدود ۸۰٪ است.



شکل (۴-۳۷)- نتایج روش RF برازش شده به روی پایگاه داده Synthetic در تخمین غلظت رنگدانه از متغیرهای ($R_{rs}(\lambda)$ در پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

شکل (۴-۳۸) میزان اهمیت هر یک از متغیرهای $R_{rs}(\lambda)$ پایگاه داده Synthetic را در برآورد غلظت رنگدانه پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه نشان می‌دهد. این بررسی به‌طور جداگانه به ازای نوفه‌های نرمال (قاب چپ) و یکنواخت (قاب راست) صورت گرفته است. با توجه به شکل زیر می‌توان گفت که به‌طور کلی متغیر $R_{rs}(443)$ همواره جزو متغیرهای بااهمیت در میان متغیرهای توضیحی این پایگاه داده است. البته این برتری در داده‌های با ۵٪ نوفه‌ی یکنواخت، بسیار چشمگیر است. به‌طور کلی، متغیرها در این پایگاه داده، از لحاظ اهمیت دارای برتری چندان محسوسی نسبت به یکدیگر نیستند (به جز در داده‌های با ۵٪ نوفه‌ی یکنواخت). همچنین می‌توان متغیر $R_{rs}(490)$ را به‌عنوان کم‌اهمیت‌ترین متغیر، در برآورد غلظت رنگدانه در این پایگاه داده شناسایی کرد. در بین سایر متغیرها، نمی‌توان برتری قابل ملاحظه‌ای را شناسایی نمود.

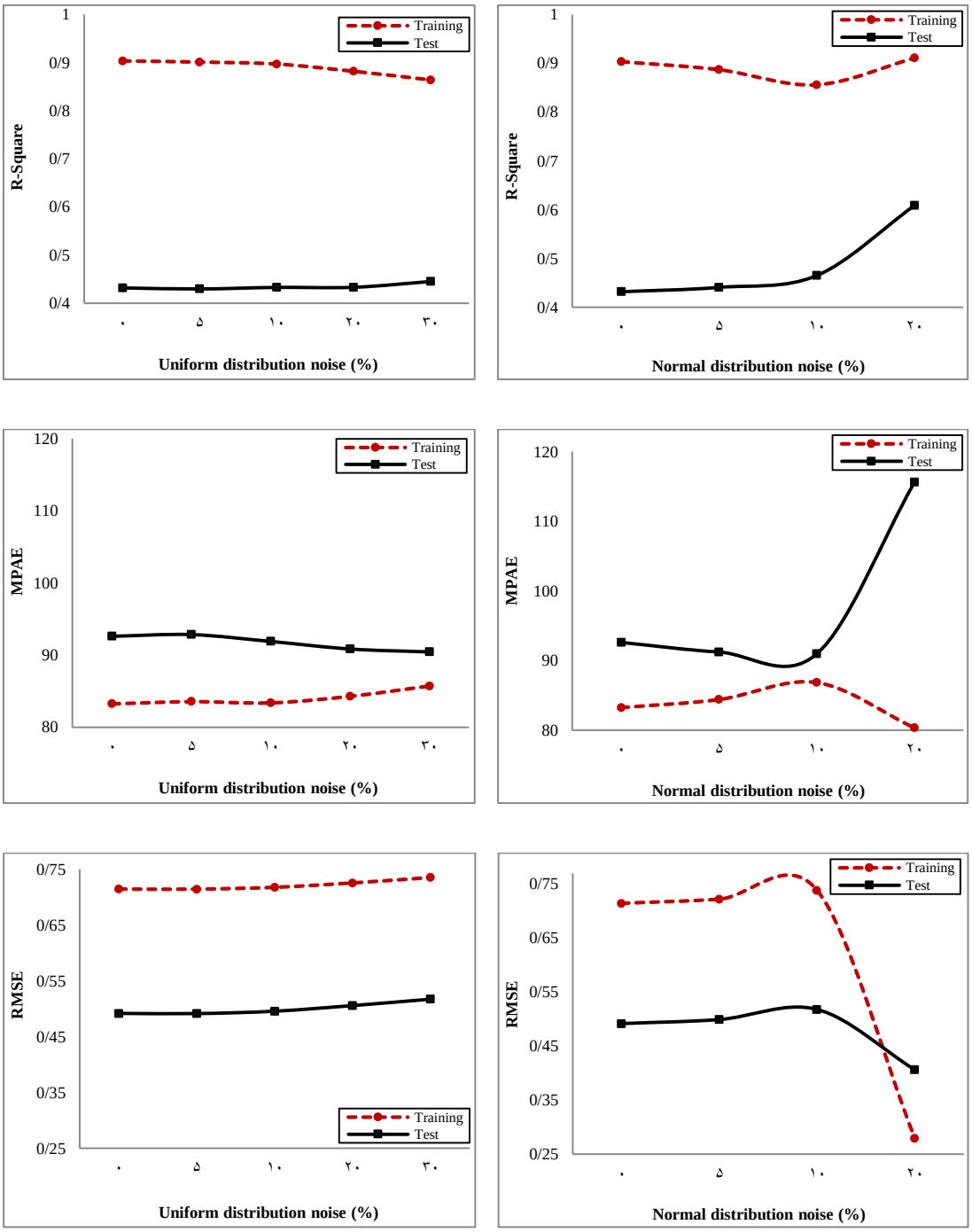


شکل (۴-۳۸)-میزان اهمیت متغیرهای $R_{rs}(\lambda)$ پایگاه داده Synthetic در برآورد غلظت رنگدانه پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب راست) و یکنواخت (قاب چپ)

۴-۴-۲ تخمین غلظت رنگدانه در پایگاه داده SeaBAM با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ با استفاده از روش SVM برازش‌شده برروی پایگاه داده Synthetic

شکل (۴-۳۹) نتایج به‌دست آمده از روش SVM در برآورد غلظت رنگدانه با به‌کارگیری متغیرهای $R_{rs}(\lambda)$ و با قرار دادن پایگاه داده Synthetic به‌عنوان داده‌های مدل‌ساز و پایگاه داده SeaBAM به‌عنوان داده‌های آزمون را نشان می‌دهد. در این شکل، مقدار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز (Synthetic) و آزمون (SeaBAM) نشان داده شده است. با توجه به شکل (۴-۳۹) می‌توان گفت که روش SVM در برآورد غلظت رنگدانه، نسبت به افزایش مقدار نوفه‌ی یکنواخت، بسیار پایدار عمل می‌کند. این در حالی است که با توجه به شکل (۴-۳۹)، این پایداری نسبت به افزایش مقدار نوفه‌ی نرمال وجود ندارد. با توجه به قاب‌های راست این شکل، به‌نظر می‌رسد که روش SVM به‌ازای نوفه‌ی ۲۰٪ نرمال عملکرد مناسبی داشته است. با توجه به اختلاف محسوس بین ضریب تعیین به‌دست آمده در داده‌های مدل‌ساز و داده‌های آزمون، می‌توان گفت که روش SVM برای برآورد غلظت رنگدانه در پایگاه داده SeaBAM با استفاده از اطلاعات موجود در پایگاه داده Synthetic، چندان مناسب عمل نکرده است. شکل زیر حاکی از آن است که دقت روش SVM در داده‌های مدل‌ساز، به ازای مقادیر مختلف نوفه، همواره حدود ۹۰٪ است. این در حالی است که میزان دقت این روش در داده‌های آزمون حدود ۴۰٪ است.

شایان ذکر است در این پایگاه داده، نتایج بهینه با به‌کارگیری هسته‌ی نرمال در روش SVM به‌دست آمده است.

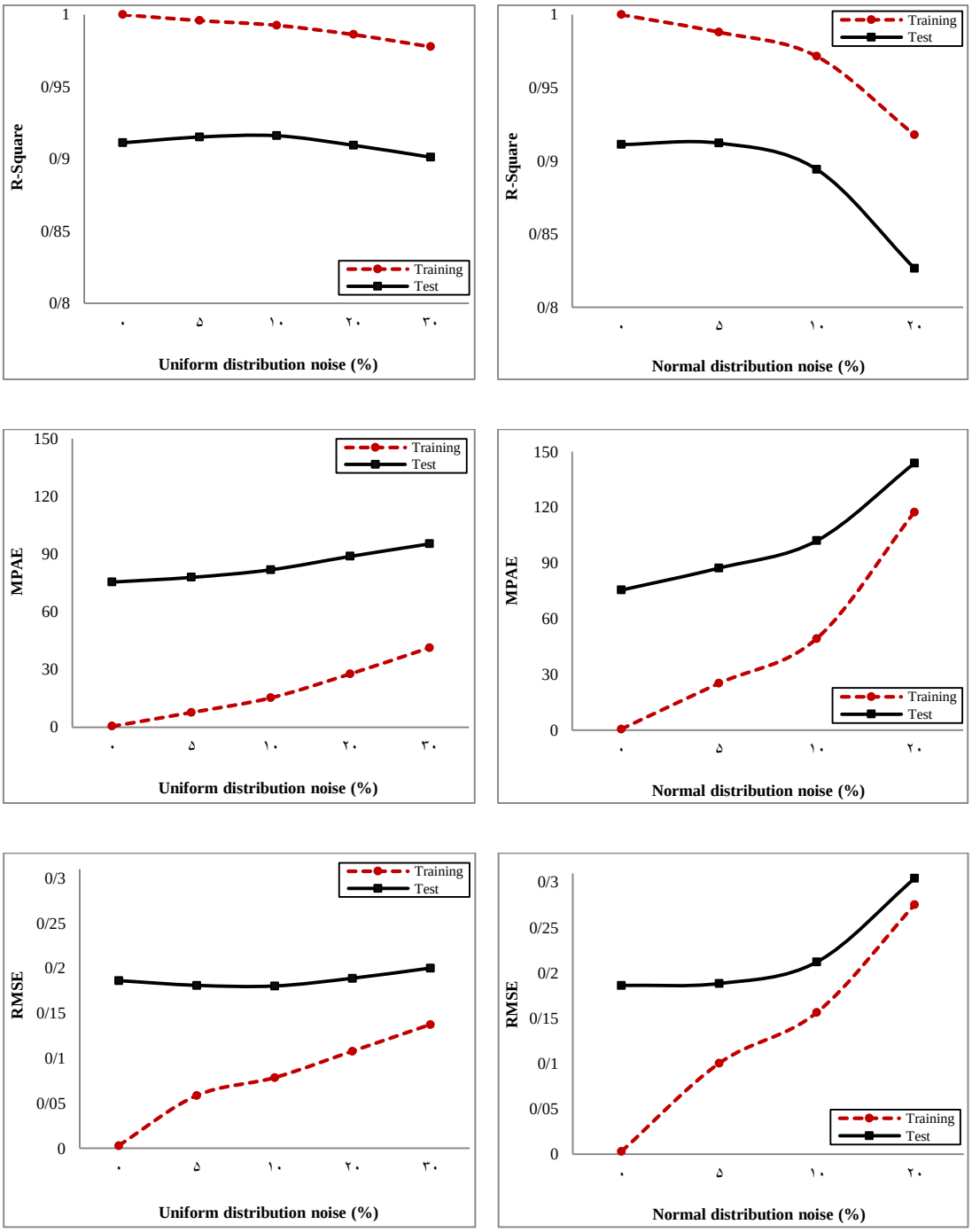


شکل (۴-۳۹)- نتایج روش SVM برازش شده به روی پایگاه داده Synthetic در تخمین غلظت رنگدانه از متغیرهای $R_{rs}(\lambda)$ در پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

۴-۴-۳ تخمین غلظت رنگدانه در پایگاه داده SeaBAM با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ با استفاده از روش RF برازش شده به‌روی پایگاه داده Synthetic

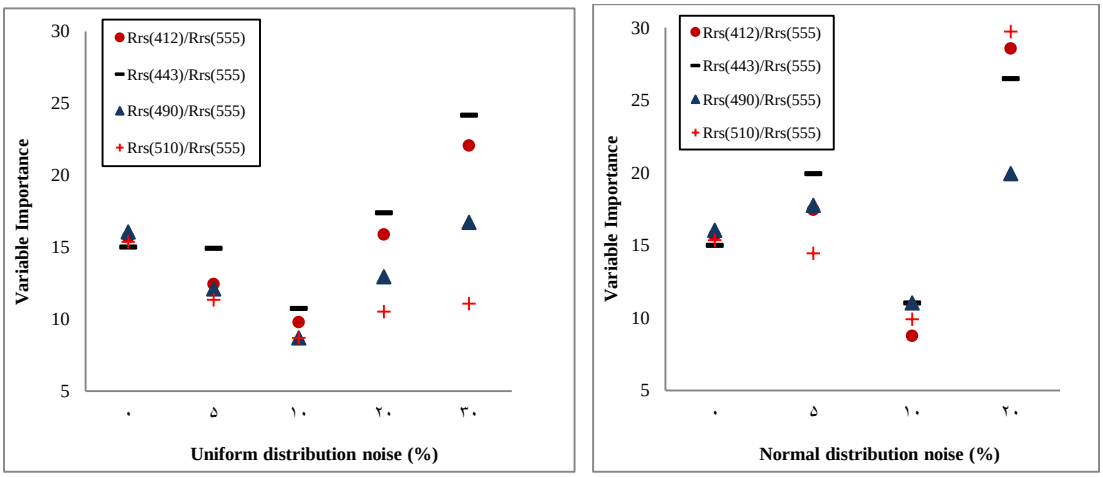
شکل (۴-۴۰) نتایج حاصل از روش RF را در برآورد غلظت رنگدانه بر حسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ و با قرار دادن پایگاه داده Synthetic به‌عنوان داده‌های مدل‌ساز و پایگاه داده SeaBAM به‌عنوان داده‌های آزمون نشان می‌دهد. در این شکل، مقدار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز (Synthetic) و آزمون (SeaBAM) نشان داده شده است. با توجه به شکل، می‌توان گفت که روش RF نسبت به افزایش مقدار نوفه‌ی یکنواخت، تا حد زیادی پایدار است. این پایداری، به‌ازای افزایش مقدار نوفه‌ی نرمال وجود ندارد. همچنین می‌توان اضافه کرد که عملکرد این روش نسبت به افزودن مقدار نوفه‌ی یکنواخت بهتر از افزودن مقدار نوفه‌ی نرمال می‌باشد.

با توجه به دو قاب بالایی شکل (۴-۴۰)، ملاحظه می‌شود که به ازای افزایش مقدار نوفه، میزان دقت برآورد مدل کاهش یافته است. به‌طوری که با افزایش مقدار نوفه‌ی یکنواخت، مقدار R^2 در داده‌های آزمون، از ۹۱٪ به ۹۰٪ کاهش یافته است. این در حالی است که مقدار این معیار در داده‌های آزمون، به ازای افزایش مقدار نوفه‌ی نرمال از ۹۱٪ به ۸۲٪ کاهش می‌یابد.



شکل (۴-۴۰)- نتایج روشRF برازش شده روی پایگاه دادهSynthetic در تخمین غلظت رنگدانه برحسب متغیرهای $R_{rs}(\lambda)/R_{rs}(555)$ در پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

شکل (۴-۴۱) میزان اهمیت هر یک از متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ پایگاه داده Synthetic را در برآورد غلظت رنگدانه پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه نشان می‌دهد. با توجه به این شکل، نمی‌توان برتری محسوس و مطلقی در بین متغیرها شناسایی کرد. به‌طور کلی می‌توان گفت متغیر $R_{rs}(443)/ R_{rs}(555)$ تا حدی از اهمیت بیشتری نسبت به سایر متغیرها برخوردار است. همچنین به‌نظر می‌رسد که متغیر $R_{rs}(510)/ R_{rs}(555)$ در داده‌های به همراه نوفه‌ی نرمال دارای اهمیت بیشتری نسبت به داده‌های به همراه نوفه‌ی یکنواخت است. به‌طوری که اهمیت این متغیر به ازای ۲۰٪ نوفه‌ی نرمال، دارای بالاترین اولویت می‌باشد.



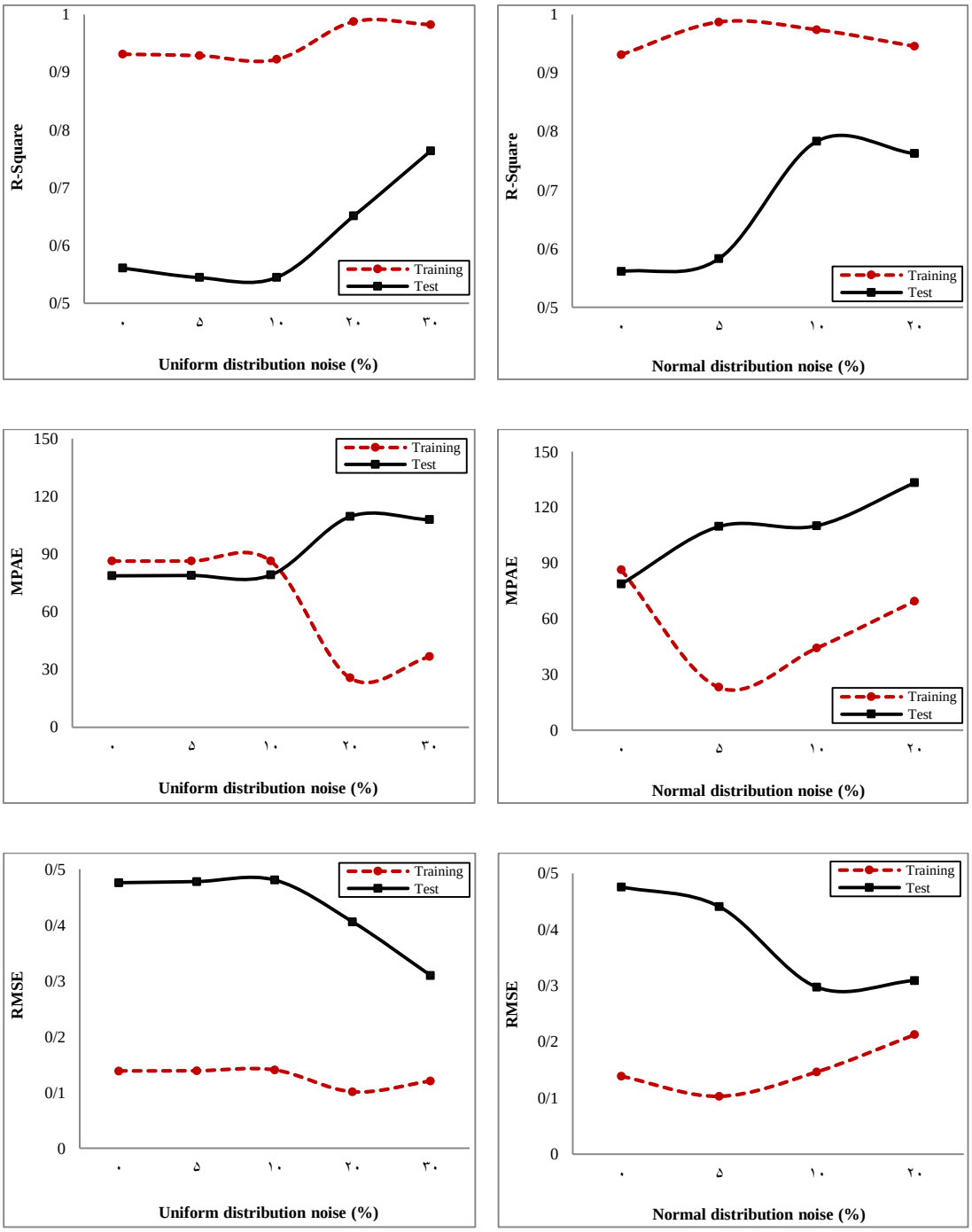
شکل (۴-۴۱)-میزان اهمیت متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ پایگاه داده Synthetic در برآورد غلظت رنگدانه پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب راست) و یکنواخت (قاب چپ)

با توجه به شکل (۴-۴۱)، مشاهده می‌گردد که در حالت داده‌های بدون نوفه، اهمیت هیچ‌یک از متغیرها، دارای برتری قابل ملاحظه‌ای نسبت به سایر متغیرها نیست.

۴-۴-۴ تخمین غلظت رنگدانه در پایگاه داده SeaBAM با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ با استفاده از روش SVM برازش‌شده بر روی پایگاه داده Synthetic

در شکل (۴-۴۲) نتایج به‌دست آمده از روش SVM در برآورد غلظت رنگدانه با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ و با قرار دادن پایگاه داده Synthetic به‌عنوان داده‌های مدل‌ساز و پایگاه داده SeaBAM به‌عنوان داده‌های آزمون نشان داده شده است. در این شکل، مقدار R^2، MPAE و RMSE به ازای مقادیر مختلف نوفه‌های نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ) در داده‌های مدل‌ساز (Synthetic) و آزمون (SeaBAM) نشان داده شده است. ملاحظه می‌شود که با توجه به شکل، نمی‌توان پایداری مناسبی را در رفتار منحنی این دو معیار نسبت به افزایش مقدار نوفه‌ی نرمال شاهد بود. البته به ازای مقادیر کم از نوفه‌ی یکنواخت، تا حدی می‌توان شاهد پایداری در عملکرد این روش بود. به‌طور کلی با توجه به هر ۶ قاب موجود در شکل (۴-۴۲)، مشاهده می‌گردد که روش SVM در برآورد غلظت رنگدانه با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ به‌عنوان متغیرهای توضیحی، و با قرار دادن پایگاه داده Synthetic به‌عنوان داده‌های مدل‌ساز و پایگاه داده SeaBAM به‌عنوان داده‌های آزمون، دارای عملکرد پایدار و با ثباتی نیست.

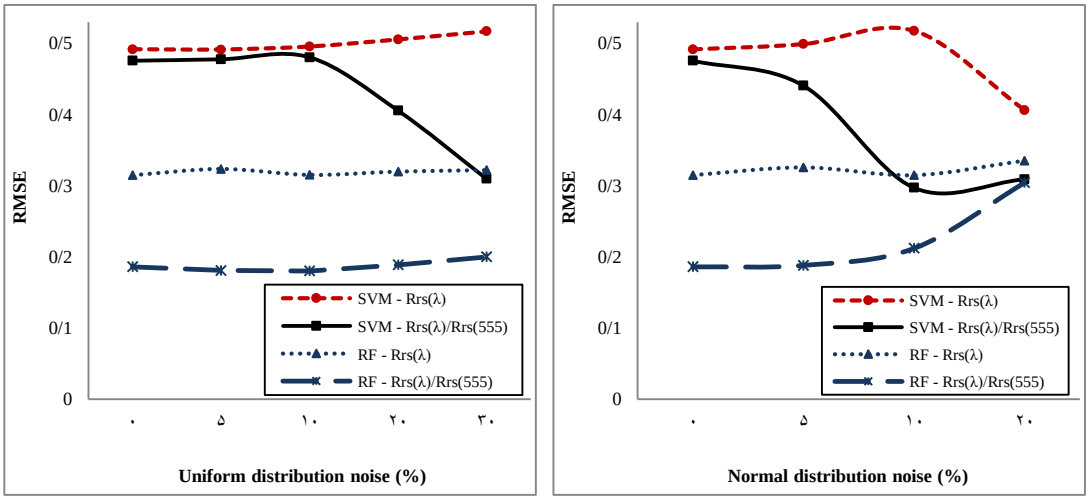
شایان ذکر است در این پایگاه داده، نتایج بهینه با به‌کارگیری هسته‌ی نرمال در روش SVM به‌دست آمده است.



شکل (۴-۴۲)- نتایج روش SVM برازش شده روی پایگاه داده Synthetic در تخمین غلظت رنگدانه برحسب متغیرهای $R_{rs}(\lambda)/R_{rs}(555)$ در پایگاه داده SeaBAM به ازای مقادیر مختلف نوفه‌ی نرمال (قاب‌های راست) و یکنواخت (قاب‌های چپ)

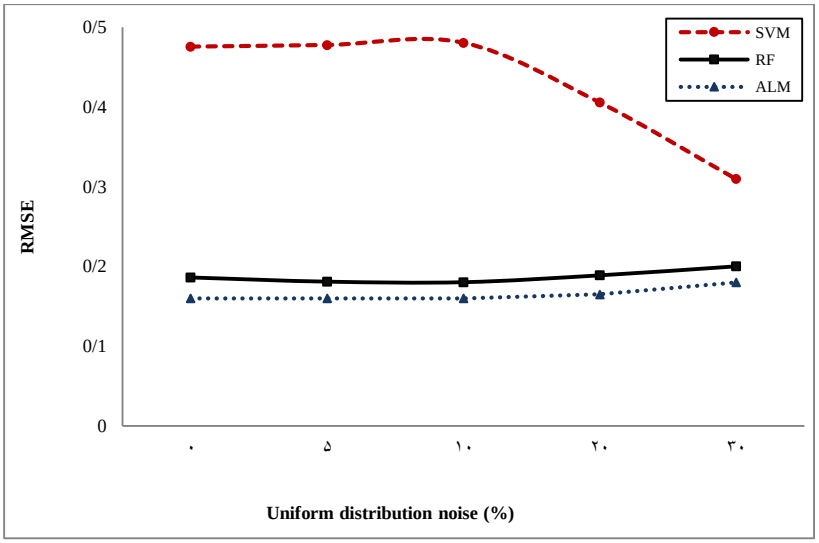
۴-۴-۵ مقایسه‌ی نتایج به‌دست آمده در پایگاه داده SeaBAM با استفاده پایگاه داده Synthetic

شکل (۴-۴۳) روند معیار RMSE را برحسب مقدار نوفه‌ی نرمال (قاب راست) و نوفه‌ی یکنواخت (قاب چپ) نشان می‌دهد. ملاحظه می‌شود که با قرار دادن پایگاه داده Synthetic به عنوان داده‌های مدل‌ساز، به‌طور کلی روش RF در برآورد غلظت رنگدانه‌ی پایگاه داده SeaBAM نسبت به روش SVM دارای خطای کمتری می‌باشد. همچنین مشاهده می‌شود که در هر دو روش RF و SVM، همواره به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ منجر به نتایج بهتری نسبت به استفاده از متغیرهای $R_{rs}(\lambda)$ به‌عنوان متغیرهای توضیحی می‌گردد. ضمن این‌که در برآورد غلظت رنگدانه، روش RF نسبت به افزایش مقدار هر دو نوع نوفه، بسیار پایدار عمل کرده است. در این میان به‌نظر می‌رسد که روش SVM به‌ازای مقادیر زیاد نوفه، دارای عملکردی مشابه روش RF بوده است.



شکل (۴-۴۳)- مقایسه‌ی روش‌های RF و SVM در تخمین غلظت رنگدانه پایگاه داده SeaBAM برحسب متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ و$R_{rs}(\lambda)$ با به‌کارگیری پایگاه داده Synthetic به عنوان داده‌های مدل‌ساز به ازای مقادیر مختلف نوفه‌ی نرمال (قاب راست) و یکنواخت (قاب چپ)

همان‌طور که در فصل اول اشاره شد، پیش از این طاهری شهرآیینی و همکاران (۲۰۰۹) با افزودن نوفه‌ی یکنواخت به این پایگاه داده، و با به‌کارگیری متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ اقدام به تخمین غلظت رنگدانه توسط روش ALM کرده‌اند. شکل (۴-۴۴) عملکرد روش‌های RF ، SVM و ALM در برآورد غلظت رنگدانه‌ی پایگاه داده SeaBAM با استفاده از متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ پایگاه داده Synthetic را نشان می‌دهد. با توجه به شکل (۴-۴۴)، دو روش RF و SVM نسبت به روش ALM دارای خطای بیشتری در برآورد پارامتر مورد نظر در این پایگاه داده هستند. این در حالی است که به‌نظر می‌رسد عملکرد روش RF ، تفاوت چندانی با روش ALM ندارد. همچنین ملاحظه می‌شود که روش SVM به‌ازای مقادیر زیاد نوفه، تا حدی از عملکرد مناسب‌تری برخوردار است. در این میان، بارزترین مزیت روش‌های RF و ALM را می‌توان پایداری زیاد این دو روش، نسبت به افزایش مقدار نوفه دانست. البته نباید فراموش کرد که روش RF دارای قابلیت شناسایی متغیرهای با اهمیت نیز می‌باشد.



شکل (۴-۴۴)- مقایسه‌ی نتایج روش‌های RF ، SVM و ALM در تخمین غلظت رنگدانه‌ی پایگاه داده SeaBAM از متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ در پایگاه داده Synthetic به ازای مقادیر مختلف نوفه‌ی یکنواخت

۴-۵ تخمین غلظت ذرات معلق، کلروفیل و مواد آلی محلول زرد رنگ در پایگاه داده MOMO

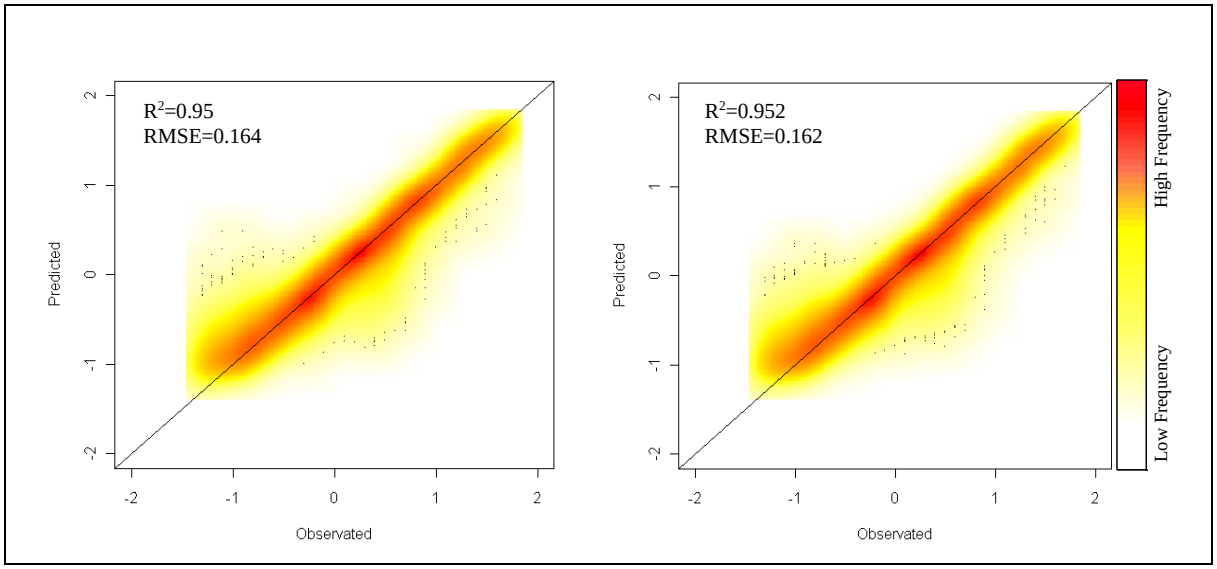
در این بخش، عملکرد دو روش RF و SVM به‌منظور برآورد سه متغیر پاسخ در پایگاه داده MOMO مورد ارزیابی قرار گرفته است. این پایگاه داده که در فصل دوم به‌طور کامل معرفی گردید شامل ۱۸ متغیر توضیحی است. سه متغیر پاسخ این پایگاه داده عبارتند از غلظت ذرات معلق، غلظت کلروفیل و غلظت مواد آلی محلول زرد رنگ.

با توجه به سایر تحقیقات انجام شده بر روی پایگاه داده‌ی MOMO، می‌توان گفت که در برآورد هر یک از متغیرهای پاسخ، معیار مورد توجه محققین، RMSE می‌باشد. لذا کلیه‌ی مراحل بهینه‌سازی پارامترهای مدل در این پایگاه با در نظر گرفتن کمینه‌ی RMSE انجام گرفته است.

۴-۵-۱ تخمین غلظت ذرات معلق در پایگاه داده MOMO

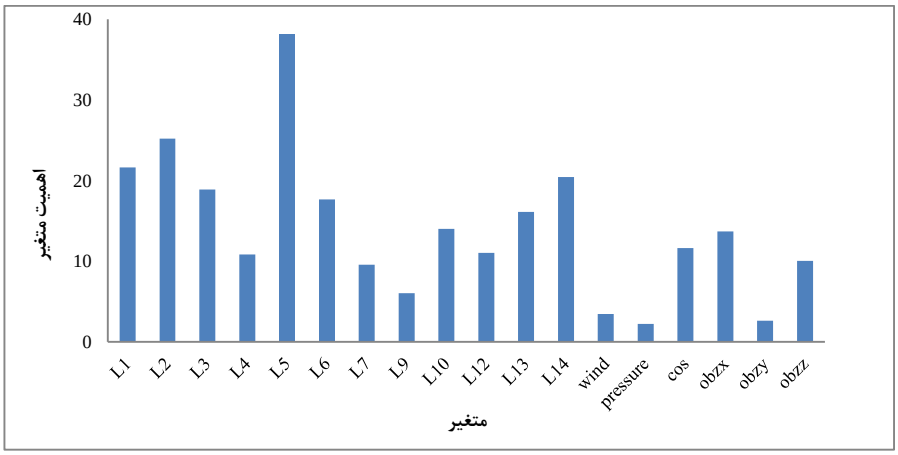
۴-۵-۱-۱ تخمین غلظت ذرات معلق با استفاده از روش RF

شکل (۴-۴۵) نتایج به‌دست آمده از اعمال روش RF در برآورد غلظت ذرات معلق را در این پایگاه داده نشان می‌دهد. با توجه به شکل، مقدار RMSE در داده‌های مدل‌ساز (قاب چپ) ۰/۱۶۴ و در داده‌های آزمون (قاب راست) ۰/۱۶۲ می‌باشد. همچنین با توجه به نتایج حاصل از این روش، مقدار ضریب تعیین در داده‌های مدل‌ساز ۹۵٪ (قاب چپ) و در داده‌های آزمون ۹۵٪ (قاب راست) به‌دست آمده است.



شکل (۴-۴۵)- نمودار لگاریتم مقادیر مشاهده‌شده در مقابل لگاریتم مقادیر پیش‌بینی شده توسط روش RF به‌منظور برآورد غلظت ذرات معلق در پایگاه داده MOMO به ازای داده‌های مدل‌ساز (قاب چپ) و داده‌های آزمون (قاب راست)

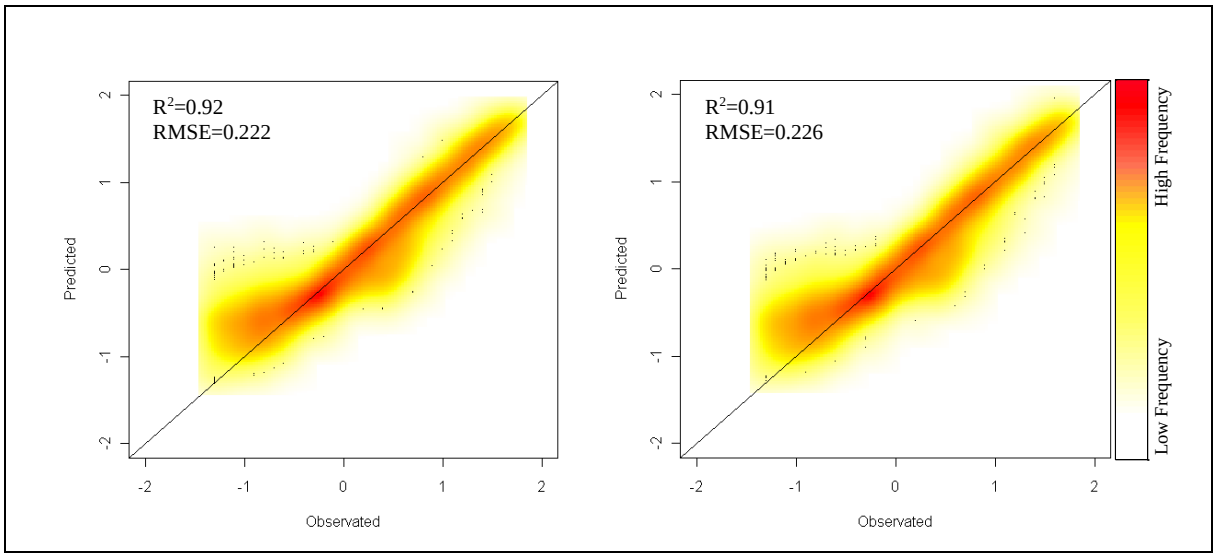
همچنین با استفاده از مدل بهینه‌ی حاصل از روش RF ، میزان اهمیت هر یک از متغیرهای توضیحی در برآورد غلظت ذرات معلق در این پایگاه داده مشخص گردید. شکل (۴-۴۶) میزان اهمیت این متغیرها را نشان می‌دهد. با توجه به شکل، ملاحظه می‌شود که متغیر L_5 دارای بیشترین اهمیت در بین تمامی متغیرها است. همچنین متغیرهای wind، pressure و obzy دارای کم‌ترین اهمیت در بین متغیرهای توضیحی می‌باشند.



شکل (۴-۴۶)- اهمیت متغیرهای توضیحی پایگاه داده MOMO در برآورد غلظت ذرات معلق

۴-۵-۱-۲ تخمین غلظت ذرات معلق در پایگاه داده MOMO با استفاده از روش SVM

در شکل (۴-۴۷) نتایج حاص از روش SVM در برآورد غلظت ذرات معلق در این پایگاه داده به‌نمایش در آمده است. با توجه به شکل، مقدار RMSE در داده‌های مدل‌ساز (قاب چپ) ۰/۲۲۲ و در داده‌های آزمون (قاب راست) ۰/۲۲۶ می‌باشد. همچنین با توجه به شکل، مقدار R^2 در داده‌های مدل‌ساز ۹۲٪ (قاب چپ) و در داده‌های آزمون ۹۱٪ (قاب راست) می‌باشد.



شکل (۴-۴۷)- نمودار لگاریتم مقادیر مشاهده‌شده در مقابل لگاریتم مقادیر پیش‌بینی شده توسط روش SVM در برآورد غلظت ذرات معلق در پایگاه داده MOMO به ازای داده‌های مدل‌ساز (قاب چپ) و داده‌های آزمون (قاب راست)

۴-۵-۱-۳ مقایسه‌ی نتایج به‌دست آمده در برآورد غلظت ذرات معلق از پایگاه داده MOMO

همان‌طور که در فصل اول اشاره شد، پیش از این، شرودر (۲۰۰۵) از روش ANN به‌منظور تخمین برخی از پارامترهای این پایگاه داده استفاده کرده است. مقایسه‌ی عملکرد سه روش RF ، SVM و ANN در برآورد غلظت ذرات معلق در پایگاه داده MOMO در جدول (۴-۲) نشان داده شده است. با توجه به جدول (۴-۲)، هر دو روش RF و SVM دارای خطای کمتری نسبت به روش ANN در برآورد پارامتر

مورد نظر می‌باشند. این برتری نه‌تنها در داده‌های آزمون، بلکه در داده‌های مدل‌ساز نیز مشهود است. در میان این سه روش، برتری روش RF نسبت به دو روش دیگر، کاملاً چشمگیر می‌باشد.

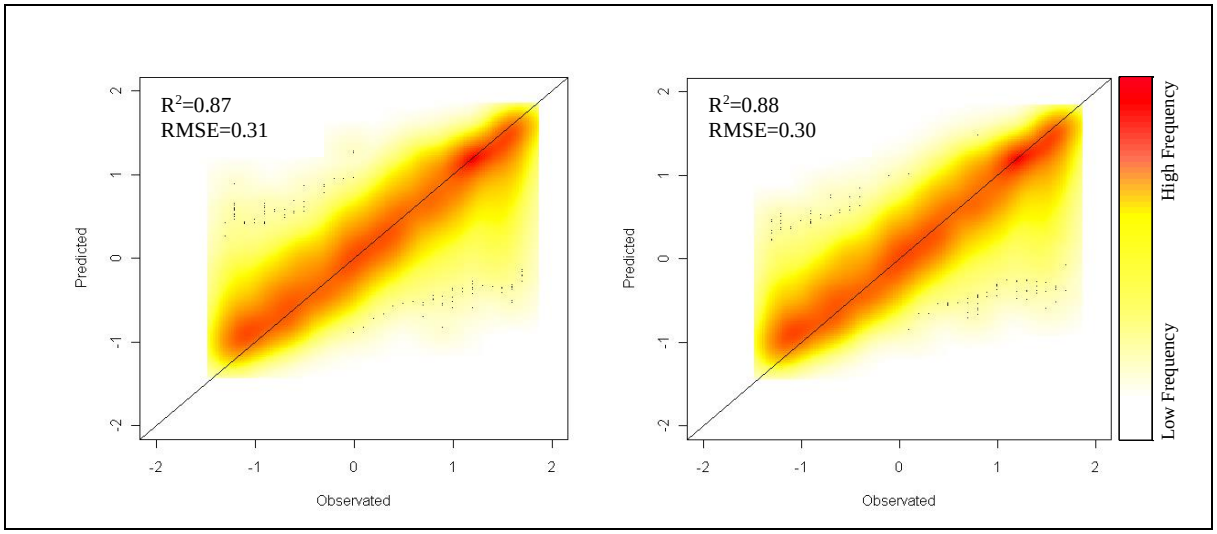
جدول (۴-۲)- عملکرد روش‌های RF، SVM و ANN در برآورد غلظت ذرات معلق پایگاه داده MOMO

روش	RMSE	
	داده‌های مدل‌ساز	داده‌های آزمون
RF	۰/۱۶۴	۰/۱۶۲
SVM	۰/۲۲۲	۰/۲۲۶
ANN	۰/۲۳۸	۰/۲۳۶

۴-۵-۲ تخمین غلظت کلروفیل در پایگاه داده MOMO

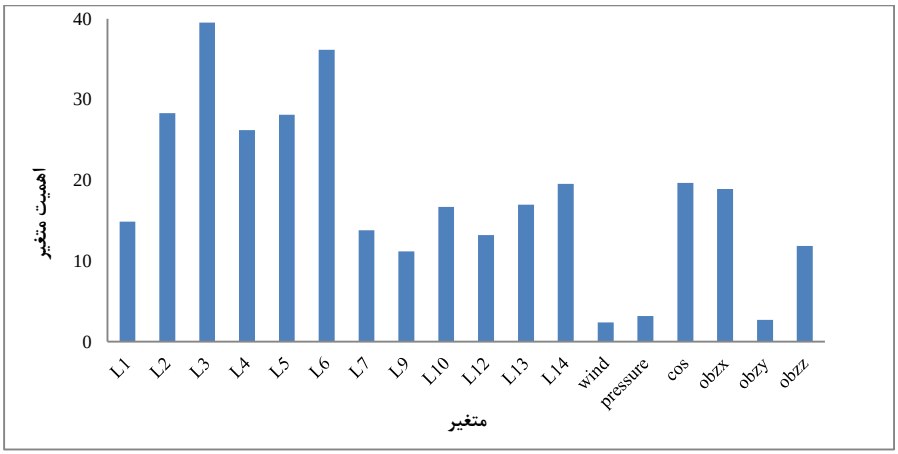
۴-۵-۲-۱ تخمین غلظت کلروفیل با استفاده از روش RF

در شکل (۴-۴۸) نتایج به‌دست آمده از اعمال روش RF در برآورد غلظت کلروفیل در پایگاه داده MOMO به‌نمایش در آمده است. با توجه به این شکل، مقدار RMSE در داده‌های مدل‌ساز (قاب چپ) ۰/۳۱ و در داده‌های آزمون (قاب راست) ۰/۳۰ می‌باشد. همچنین با توجه به شکل، مقدار R^2 در داده‌های مدل‌ساز ۸۷٪ (قاب چپ) و در داده‌های آزمون ۸۸٪ (قاب راست) می‌باشد.



شکل (۴-۴۸)- نمودار لگاریتم مقادیر مشاهده‌شده در مقابل لگاریتم مقادیر پیش‌بینی شده توسط روش RF به‌منظور برآورد غلظت کلروفیل در پایگاه داده MOMO به ازای داده‌های مدل‌ساز (قاب چپ) و داده‌های آزمون (قاب راست)

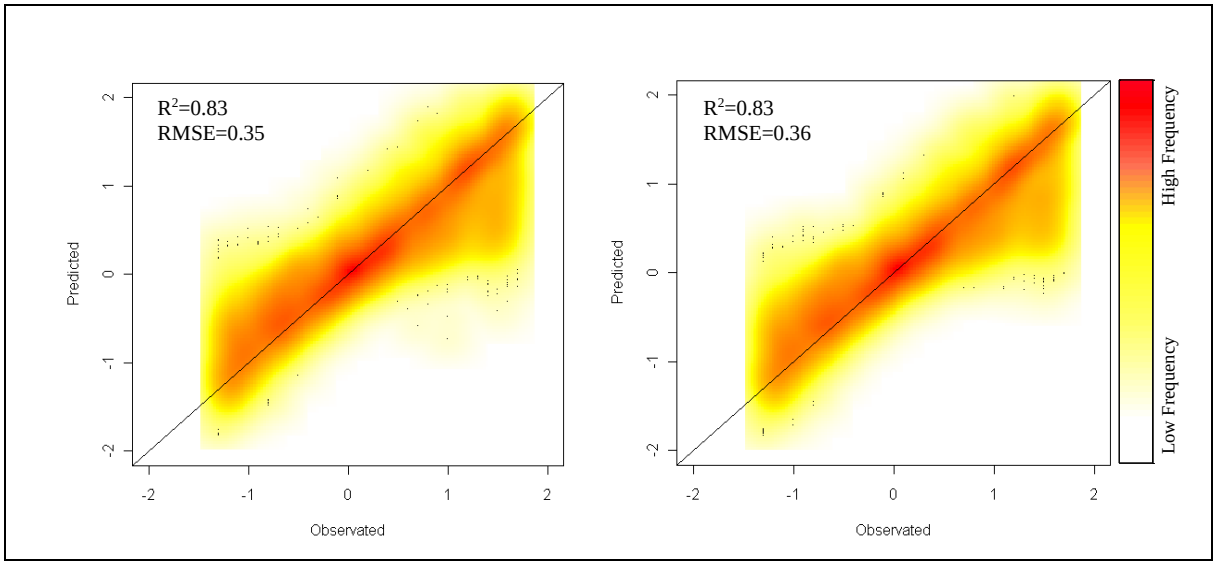
با توجه به قابلیت شناسایی اهمیت هر متغیر توسط روش RF ، میزان اهمیت هر یک از متغیرهای توضیحی در برآورد غلظت کلروفیل در این پایگاه داده مشخص گردید. شکل (۴-۴۹) میزان اهمیت هریک از متغیرهای توضیحی را به تصویر کشیده است. ملاحظه می‌شود که متغیرهای L_3 و L_6 دارای بیشترین اهمیت می‌باشند. همچنین متغیرهای wind، pressure و obzy دارای کم‌ترین اهمیت در بین متغیرهای توضیحی می‌باشند. (توجه شود که در برآورد غلظت ذرات معلق نیز این سه متغیر دارای کم-ترین اهمیت بودند.)



شکل(۴-۴۹)- اهمیت متغیرهای توضیحی پایگاه داده MOMO در برآورد غلظت کلروفیل

۴-۵-۲-۲ تخمین غلظت کلروفیل در پایگاه داده MOMO با استفاده از روش SVM

شکل (۴-۵۰) نتایج به‌دست آمده از اعمال روش SVM در برآورد غلظت کلروفیل در پایگاه داده MOMO را به‌تصویر در آورده است. با توجه به این شکل، مقدار RMSE در داده‌های مدل‌ساز (قاب چپ) ۰/۳۵ و در داده‌های آزمون (قاب راست) ۰/۳۶ می‌باشد. همچنین با توجه به شکل، مقدار R^2 در داده‌های مدل‌ساز ۸۳٪ (قاب چپ) و در داده‌های آزمون ۸۳٪ (قاب راست) می‌باشد.



شکل (۴-۵۰)- نمودار لگاریتم مقادیر مشاهده‌شده در مقابل لگاریتم مقادیر پیش‌بینی شده توسط روش SVM به‌منظور برآورد غلظت کلروفیل در پایگاه داده MOMO به ازای داده‌های مدل‌ساز (قاب چپ) و داده‌های آزمون (قاب راست)

۴-۵-۲-۳ مقایسه‌ی نتایج به‌دست آمده در برآورد غلظت کلروفیل از پایگاه داده MOMO

به‌منظور ارزیابی عملکرد دو روش RF و SVM در برآورد غلظت کلروفیل در پایگاه داده MOMO، نتایج این دو روش با نتایج روش ANN (شرودر، ۲۰۰۵) مقایسه گردیده‌است. جدول (۴-۳) نتایج این سه روش را در برآورد پارامتر مورد نظر نشان می‌دهد. با توجه به این نتایج، روش RF دارای بهترین عملکرد در برآورد پارامتر مورد نظر می‌باشد. همچنین در میان این سه روش، روش SVM دارای بیشترین خطای برآورد است.

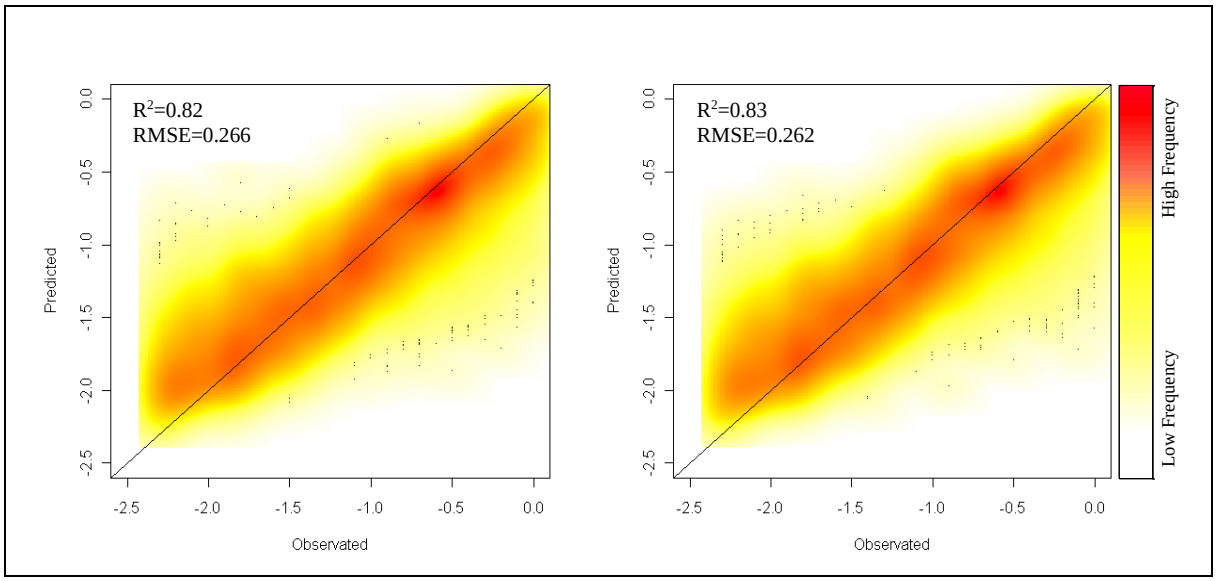
جدول (۴-۳)- عملکرد روش‌های RF ، SVM و ANN در برآورد غلظت کلروفیل پایگاه داده MOMO

روش	RMSE	
	داده‌های مدل‌ساز	داده‌های آزمون
RF	۰/۳۱	۰/۳۰
SVM	۰/۳۵	۰/۳۶
ANN	۰/۳۲۶	۰/۳۲۵

۴-۵-۳ تخمین غلظت مواد آلی محلول زرد رنگ در پایگاه داده MOMO

۴-۵-۳-۱ تخمین غلظت مواد آلی محلول زرد رنگ با استفاده از روش RF

در شکل (۴-۵۱) نتایج به‌دست آمده از اعمال روش RF در برآورد غلظت مواد آلی محلول زرد رنگ در پایگاه داده MOMO را به‌نمایش در آمده است. با توجه به این شکل، مقدار RMSE در داده‌های مدل‌ساز (قاب چپ) ۰/۲۶۶ و در داده‌های آزمون (قاب راست) ۰/۲۶۲ می‌باشد. همچنین با توجه به شکل، مقدار R^2 در داده‌های مدل‌ساز ۸۲٪ (قاب چپ) و در داده‌های آزمون ۸۳٪ (قاب راست) می‌باشد.

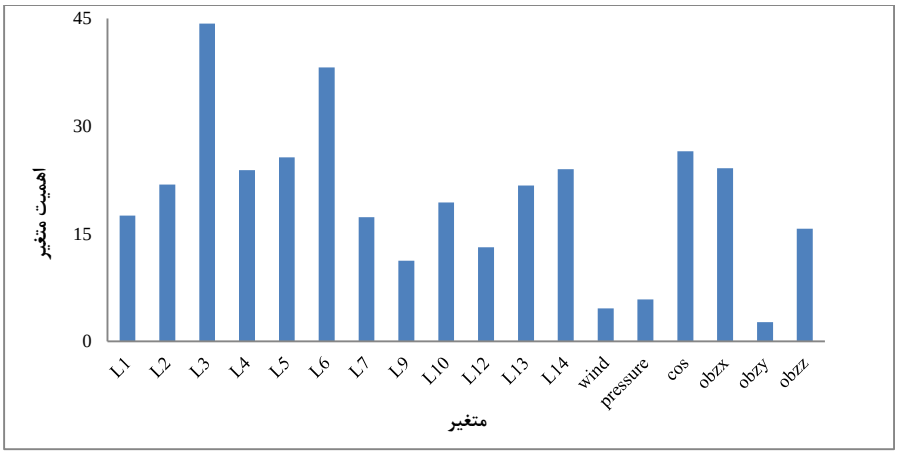


شکل (۴-۵۱)- نمودار لگاریتم مقادیر مشاهده‌شده در مقابل لگاریتم مقادیر پیش‌بینی شده توسط روش RF به‌منظور برآورد غلظت مواد آلی محلول زرد رنگ در پایگاه داده MOMO به ازای داده‌های مدل‌ساز (قاب چپ) و داده‌های آزمون (قاب راست)

در ادامه، با استفاده از روش RF، میزان اهمیت هر یک از متغیرهای توضیحی در برآورد غلظت مواد آلی محلول زرد رنگ در این پایگاه داده مشخص گردید. شکل (۴-۵۲) میزان اهمیت هر متغیر را به تصویر کشیده است. با توجه به شکل (۴-۵۲)، ملاحظه می‌گردد که متغیرهای L_3 و L_6 از بیشترین

اهمیت برخوردارند. همچنین متغیرهای wind، pressure و obzy دارای کم‌ترین اهمیت در بین متغیرهای توضیحی می‌باشند.

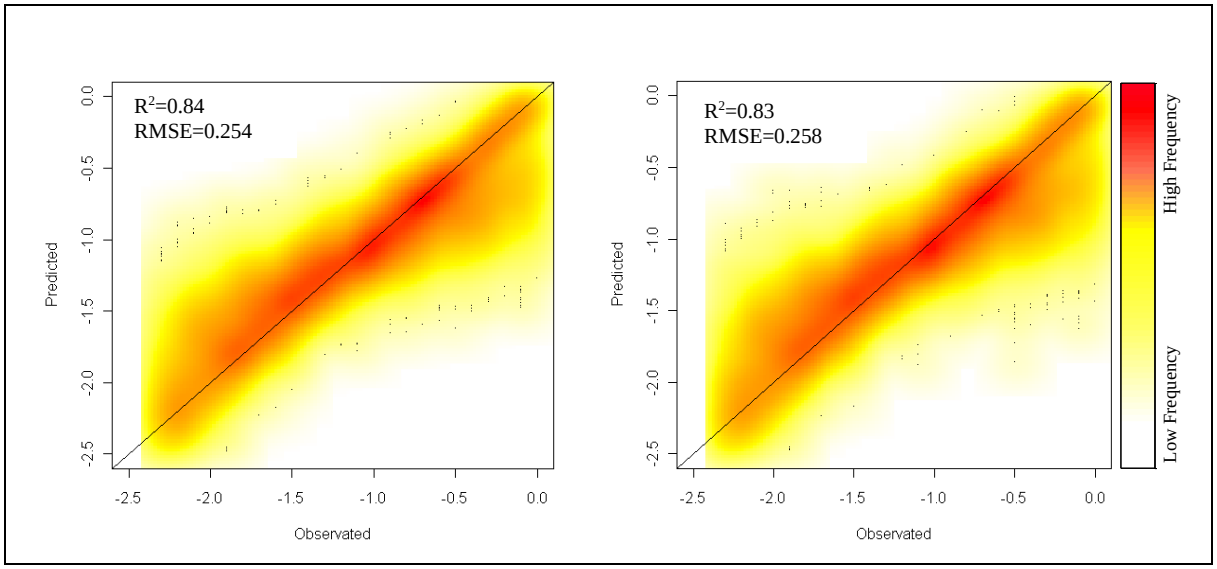
توجه شود که در برآورد هر یک از سه متغیر پاسخ (غلظت ذرات معلق، غلظت کلروفیل و غلظت مواد آلی محلول زرد رنگ)، سه متغیر wind، pressure و obzy به‌عنوان کم اهمیت‌ترین متغیرها در در بین تمامی متغیرهای توضیحی شناخته شدند.



شکل (۴-۵۲)- اهمیت متغیرهای توضیحی پایگاه داده MOMO در برآورد غلظت مواد آلی محلول زرد رنگ

۴-۵-۳-۲ تخمین غلظت مواد آلی محلول زرد رنگ در پایگاه داده MOMO با استفاده از روش SVM

شکل (۴-۵۳) نتایج به‌دست آمده از به‌کارگیری روش SVM در برآورد غلظت مواد آلی محلول زرد رنگ در پایگاه داده MOMO را نشان می‌دهد. با توجه به این شکل، مقدار RMSE در داده‌های مدل‌ساز (قاب چپ) ۰/۲۵۴ و در داده‌های آزمون (قاب راست) ۰/۲۵۸ می‌باشد. همچنین با توجه به شکل، مقدار R^2 در داده‌های مدل‌ساز ۸۴٪ (قاب چپ) و در داده‌های آزمون ۸۳٪ (قاب راست) می‌باشد.



شکل (۴-۵۳)- نمودار لگاریتم مقادیر مشاهده‌شده در مقابل لگاریتم مقادیر پیش‌بینی شده توسط روش SVM به‌منظور برآورد غلظت مواد آلی محلول زرد رنگ در پایگاه داده MOMO به ازای داده‌های مدل‌ساز (قاب چپ) و داده‌های آزمون (قاب راست)

۴-۵-۳-۳ مقایسه‌ی نتایج به‌دست آمده در برآورد غلظت مواد آلی محلول زرد رنگ در پایگاه داده MOMO

به‌منظور ارزیابی عملکرد دو روش RF و SVM در برآورد غلظت مواد آلی محلول زرد رنگ در پایگـاه داده MOMO، نتایج این دو روش با نتایج روش ANN (شرودر، ۲۰۰۵) مقایسه گردیده است. جـدول (۴-۴) نتایج این سه روش را در برآورد پارامتر مورد نظر نشان می‌دهد. با توجه به این نتـایج، روش SVM دارای بهترین عملکرد در برآورد پارامتر مورد نظر در داده‌های آزمون می‌باشد. روش RF در رتبه‌ی دوم عملکـرد قرار می‌گیرد و روش ANN دارای بیشترین خطای برآورد در این ارزیابی است.

جدول (۴-۴)- عملکرد روش‌های RF ، SVM و ANN در برآورد غلظت مواد آلی محلول زرد رنگ پایگاه داده MOMO

روش	RMSE	
	داده‌های مدل‌ساز	داده‌های آزمون
RF	۰/۲۶۶	۰/۲۶۲
SVM	۰/۲۵۴	۰/۲۵۸
ANN	۰/۲۶۵	۰/۲۶۶

۶-۴ هزینه‌ی محاسبات

یکی از مسایل حایز اهمیت در به‌کارگیری روش‌های یادگیری ماشین، بحث زمان مورد نیاز محاسبات است که به‌عنوان هزینه‌ی محاسبات مطرح می‌شود. بدیهی است که با افزایش تعداد داده‌های مدل‌ساز، این هزینه افزایش می‌یابد. با توجه به حجم پایگاه‌های داده‌ی به‌کار گرفته شده در این پایان‌نامه، و با در نظرگرفتن این‌که پارامترهای مدل در روش RF، گسسته و در روش SVM، پیوسته می‌باشند، روش RF از لحاظ هزینه‌ی زمان محاسبات، مناسب‌تر ارزیابی می‌شود.

فصل پنجم

نتیجه‌گیری و پیشنهادات

۵-۱ بحث و نتیجه‌گیری

بنا بر نتایج به‌دست آمده از اجرای دو روش RF و SVM در برآورد پارامترهای کیفی آب از داده‌های تشعشع طیفی، می‌توان گفت که عملکرد این دو روش بر روی پایگاه‌های مختلف داده‌های تشعشع طیفی، در مقایسه با سایر روش‌ها، یکسان و یکنواخت نمی‌باشد. به‌طور کلی می‌توان عملکرد این دو روش را در پایگاه داده‌های اعمال‌شده به‌صورت زیر بیان کرد.

- در پایگاه داده NOMAD، با به‌کارگیری هر دو روش RF و SVM در برآورد غلظت کلروفیل-a، استفاده از متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ منجر به نتایج بهتری نسبت به استفاده از متغیرهای $R_{rs}(\lambda)$ به‌عنوان متغیرهای توضیحی می‌گردد. همچنین در بین سه روش RF، SVM و ALM، روش SVM دارای کمترین مقدار خطای MPAE می‌باشد و پس از آن روش RF در رتبه‌ی دوم از لحاظ عملکرد قرار دارد. ضمن این‌که روش SVM بیشترین پایداری را نسبت به افزایش مقدار نوفه نشان می‌دهد.
- در پایگاه داده SeaBAM، با به‌کارگیری هر دو روش RF و SVM در برآورد غلظت رنگدانه، استفاده از متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ منجر به نتایج بهتری نسبت به استفاده از متغیرهای $R_{rs}(\lambda)$ به‌عنوان متغیرهای توضیحی می‌گردد. به‌طور کلی عملکرد این دو روش در برآورد غلظت رنگدانه، از لحاظ مقدار خطای برآورد (RMSE)، نسبت به روش‌های ALM، ANN، OC4 و CCO بهتر می‌باشد. به‌طوری که علاوه بر پایداری دو روش RF و SVM نسبت به افزایش مقدار نوفه، هر دو روش به‌ازای مقدار نوفه‌ی ۳۰٪ یکنواخت، منجر به RMSE=0.17 می‌شوند.
- در پایگاه داده Synthetic، با به‌کارگیری هر دو روش RF و SVM در برآورد غلظت رنگدانه پایگاه داده SeaBAM، استفاده از متغیرهای $R_{rs}(\lambda)/ R_{rs}(555)$ منجر به نتایج بهتری نسبت به استفاده از متغیرهای $R_{rs}(\lambda)$ به‌عنوان متغیرهای توضیحی می‌گردد. در این پایگاه داده، به‌کارگیری روش SVM به‌ازای کلیه‌ی مقادیر نوفه، منجر به بیشترین خطای RMSE در برآورد می‌شود. این در حالی است

که مقایسه‌ی دو روش RF و ALM حکم به برتری روش ALM می‌دهد. هرچند که این برتری به‌ازای کلیه‌ی مقادیر نوفه نیز وجود دارد، ولی اختلاف خطای برآورد این دو روش چندان زیاد نیست. به‌طوری که به‌ازای نوفه‌ی ۳۰٪ یکنواخت مقدار خطای RMSE در دو روش RF و ALM به‌ترتیب ۰/۲ و ۰/۱۸ می‌باشد. همچنین بایستی یادآور شد که در برآورد غلظت رنگدانه، هر دو روش RF و ALM نسبت به افزایش مقدار نوفه بسیار پایدار عمل می‌کنند.

- در پایگاه داده MOMO، با توجه به وجود سه متغیر پاسخ، مقایسه‌ی عملکرد سه روش RF، SVM و ANN به‌صورت جدول (۵-۱) خلاصه شده است. در این جدول به ازای هر یک از متغیرهای پاسخ، با توجه به مقدار RMSE در داده‌های آزمون، عملکرد هر سه روش مورد مقایسه قرار گرفته‌اند.

جدول (۵-۱)- مقایسه‌ی RMSE داده‌های آزمون حاصل از روش‌های RF، SVM و ANN در برآورد پارامترهای کیفی پایگاه داده MOMO

متغیر پاسخ	مقایسه‌ی **RMSE** حاصل از سه روش در داده‌های آزمون
غلظت ذرات معلق	RF<SVM<ANN
غلظت کلروفیل	RF<ANN<SVM
غلظت مواد آلی محلول زرد رنگ	SVM<RF<ANN

۵-۲ پیشنهادات

به‌منظور انجام تحقیقات و پژوهش‌های آتی در زمینه‌ی برآورد پارامترهای کیفی آب از داده‌های تشعشع طیفی به‌وسیله‌ی روش‌های یادگیری ماشین، می‌توان از راهکارهای ذیل بهره گرفت.

- معرفی و پیاده‌سازی سایر روش‌های یادگیری ماشین از قبیل روش اسپلاین[۱]، روش رگرسیون تطبیقی چندگانه اسپلاین‌ها[۲] (MARS) و روش هسته، در برآورد پارامترهای کیفی آب از داده‌های تشعشع طیفی.
- استفاده از سایر روش‌های بهینه‌سازی پارامترهای مدل از قبیل الگوریتم ژنتیک[۳]، روش نیوتن[۴]، روش گاوس-نیوتن[۵] و روش جستجوی ساده[۶]، جهت کاهش زمان محاسبات.
- حذف متغیرهای کم‌اهمیت شناخته‌شده توسط روش RF و پیاده‌سازی مجدد روش‌های برآوردیابی مختلف با متغیرهای باقیمانده.
- به‌کارگیری روش‌های RF و SVM به‌منظور رده‌بندی تصاویر ماهواره‌ای.

[1] Spline
[2] Multivariate Adaptive Regression Spline
[3] Genetic Algorithm
[4] Newton
[5] Gauss-Newton
[6] Simplex Search

پیوست

پیوست الف) سنجش از دور و داده‌های تشعشع طیفی

الف-۱ مقدمه‌ای بر سنجش از دور

از دیرباز روش‌های متفاوتی برای جمع‌آوری داده‌ها در علوم مختلف وجود داشته است که یکی از این روش‌ها سنجش از دور می‌باشد. سنجش از دور روشی برای جمع‌آوری داده محسوب می‌شود که در این روش، تماس مستقیم فیزیکی با اشیای مورد اندازه‌گیری به حداقل می‌رسد. در این روش، جمع‌آوری داده بر عهده‌ی سنجنده است در حالی‌که در روش‌های زمینی که معمولاً در تماس مستقیم یا با فاصله‌ی کم از اشیا انجام می‌شود، عامل انسانی وظیفه‌ی برداشت داده‌ها را بر عهده دارد. اولین کاربردهای سنجش از دور را می‌توان در اوایل قرن بیستم میلادی جست، جایی که در جنگ جهانی اول، عکس‌های هوایی از مناطق استراتژیک جنگی، کمک شایانی به پیشبرد جنگ برای ملل مختلف می‌کرد. در سال ۱۹۷۲ میلادی با پرتاب اولین ماهواره‌ی تحقیقاتی NASA به فضا، گام بلندی در پیشرفت علوم فضایی برداشته شد. پرتاب این ماهواره‌ی تحقیقاتی، آغازگر مرحله‌ی جدیدی از سنجش از دور بود و پس از آن با پیشرفت علوم فضایی، کشورها و سازمان‌های مختلفی اقدام به پرتاب ماهواره‌های تحقیقاتی به فضا کردند تا بتوانند از دستاوردهای این فن‌آوری بهره گیرند. در ادامه برای شناخت بهتر سنجش از دور و نحوه‌ی ثبت تصاویر ماهواره‌ای به معرفی برخی از مفاهیم و تعاریف اولیه در سنجش از دور پرداخته خواهد شد.

الف-۲ انرژی الکترومغناطیس

اساس سیستم‌های سنجش از دور بر اندازه‌گیری نوعی از انرژی به نام انرژی الکترومغناطیس است. انرژی الکترومغناطیسی که از طرف اشیا به سمت سنجنده حرکت می‌کند، توسط سنجنده دریافت، اندازه‌گیری و ثبت می‌شود. بنابراین دانستن اطلاعاتی در این مورد برای فهم بهتر تصاویر و پردازش‌های انجام‌شده روی آن‌ها بسیار ضروری است.

شناخته‌ترین نوع انرژی الکترومغناطیس همان نور است که برای ساکنان زمین عمده‌ترین منبع تولید آن خورشید می‌باشد. خورشید انرژی الکترومغناطیس را در طول موج‌های مختلف تولید کرده و به اطراف گسیل می‌دارد. زمین نیز در معرض تابش خورشید، این انرژی حیات‌بخش را دریافت کرده و بخشی از آن را منعکس و بخشی دیگر را جذب می‌کند.

تاکنون دو نظریه‌ی عمده در رابطه با انرژی الکترومغناطیس ارایه گردیده است که هر کدام بخشی از خصوصیات این انرژی را بیان می‌نمایند: نظریه‌ی موجی بودن انرژی الکترومغناطیس[۱] و نظریه‌ی ذره‌ای بودن[۲]. نظریه‌ی موجی بودن به جداسازی انواع انرژی‌های الکترومغناطیس در طول موج‌های مختلف (مثل مادون قرمز[۳] و مایکروویو[۴]) می‌پردازد. بر طبق این نظریه، انرژی الکترومغناطیس به شکل امواج سینوسی در فضا حرکت می‌کند. در حالی‌که نظریه‌ی ذره‌ای بودن به واکنش‌های میان انرژی الکترومغناطیس و سطح زمین یا اتمسفر می‌پردازد. طرفداران این دو نظریه، سال‌ها در جدال با یکدیگر، مسایل بسیاری در رابطه با انرژی الکترومغناطیس را کشف نمودند تا اینکه دانشمند بزرگ آلمانی، آلبرت انیشتین[۵] ، عنوان نمود که در سطوح ذرات بنیادین می‌توان هر دو رفتار را برای این نوع انرژی درست دانست.

یکی از مهمترین خصوصیات یک موج الکترومغناطیس، طول موج[۶] آن است. به فاصله‌ی میان دو نقطه‌ی یکسان (تکراری) موج، طول موج می‌گویند. مقدار انرژی هر موج، بستگی به طول آن موج دارد، به‌طوری‌که امواج با طول موج کوچکتر، انرژی بیشتری دارند و بالعکس. در بین امواج الکترومغناطیس،

[1] Wave Theory
[2] Particle Theory
[3] Infra-Red
[4] Microwave
[5] Albert Einstein(1879-1955)
[6] Wavelength

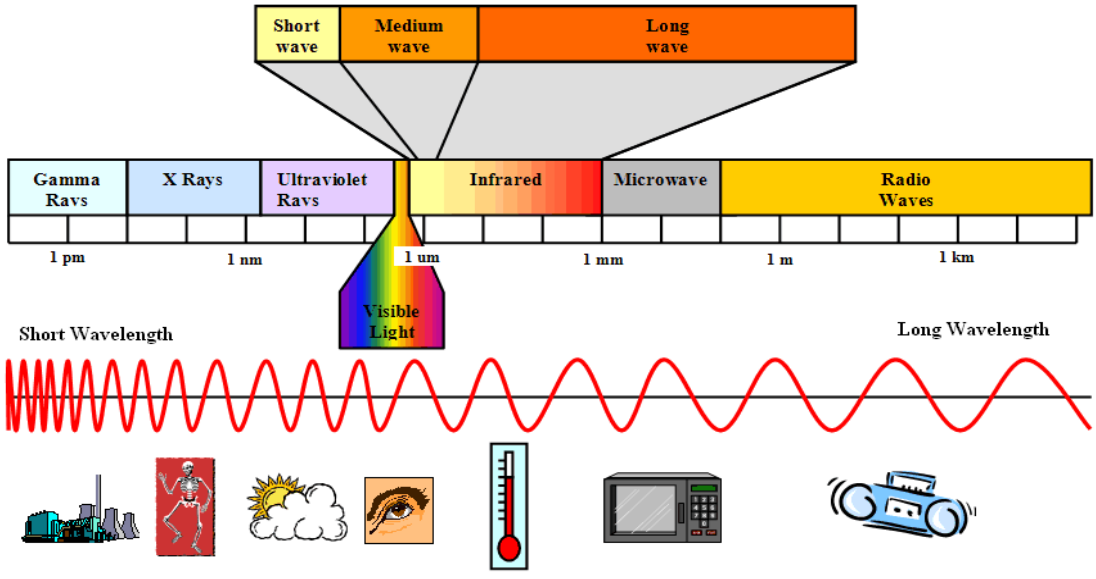
امواج گاما[1] دارای بیشترین انرژی (کمترین طول موج) و امواج رادیویی دارای کمترین انرژی (بیشترین طول موج) می‌باشند.

الف-۳ طیف الکترومغناطیس

به مجموعه‌ی کل طول موج‌های امواج الکترومغناطیس که در کنار یکدیگر و به‌ترتیب خاصی قرار گرفته‌اند طیف الکترومغناطیس[2] گویند. این طیف از اشعه‌ی گاما شروع شده و به امواج رادیویی ختم می‌شود. سنجنده‌های سنجش از دور در بخش‌های مختلفی از این طیف عمل می‌کنند. در دنیای سنجش از دور دامنه‌های مختلف طیف الکترومغناطیس دارای اسامی خاصی هستند که در متون فنی و علمی از آنها به وفور استفاده می‌شود. آشناترین بخش طیف الکترومغناطیس، بخش مریی است که طول موج‌های ۰/۴ تا ۰/۷ میکرومتر را پوشش می‌دهد. چشم ما قادر به تشخیص این امواج الکترومغناطیس می‌باشد و به همین خاطر رنگ‌های مختلف در همین طول موج‌ها تعریف می‌شود. در بین طول موج‌های بخش مریی، کوچکترین طول موج مربوط به رنگ آبی و بزرگترین طول موج مربوط به رنگ قرمز است. طول موج‌های کوچکتر از ۰/۴ میکرومتر به سه دسته‌ی عمده تقسیم می‌شوند که عبارتند از اشعه‌های گاما، ایکس[3] و فرابنفش[4]. این امواج به علت طول موج کوچک، توسط اتمسفر جذب می‌شوند. به همین دلیل این امواج در سنجش از دور مورد استفاده قرار نمی‌گیرند. بخش مادون قرمز از انتهای بخش مریی (طول موج ۰/۷ میکرومتر) شروع شده و به طول موج‌های حدود ۱ میلی‌متر ختم می‌گردد. بخش مادون قرمز طیف به بخش‌های متعددی نام‌گذاری شده است. طول موج‌های بین ۰/۷ تا ۳ میکرومتر مادون قرمز نزدیک[5]، طول موج‌های بین ۳ تا ۳۰ میکرومتر مادون قرمز میانی[6] و دامنه‌ی ۳۰ میکرومتر تا ۱ میلی‌متر را مادون قرمز

[1] Gamma-Rays
[2] Electromagnetic Spectrum
[3] X-Rays
[4] Ultra-Violet
[5] Near Infra-Red
[6] Middle Infra-Red

دور[1] می‌گویند. همچنین در برخی از مقالات به دامنه‌ی ۱ تا ۲/۵ میکرومتر مادون قرمز کوتاه[2] گفته می‌شود. یکی از مهم‌ترین بخش‌های مادون قرمز، مادون قرمز حرارتی[3] است که به آن دسته از طول موج‌هایی اطلاق می‌گردد که در اثر حرارت اجسام تولید و تابیده می‌شود. البته قسمت اعظمی از این انرژی‌های تابشی توسط اتمسفر جذب می‌گردد و تنها بازه‌ای در دامنه‌ی ۸ تا ۱۴ میکرومتر وجود دارد که جذب اتمسفری در آن پایین است. طول موج‌های حرارتی سنجنده‌ها در همین بازه قرار می‌گیرند. بخش امواج مایکروویو، دامنه‌ی ۱ تا ۱۰۰۰ میلی‌متر را شامل می‌شود. این دسته از امواج به علت بزرگی طول موجشان، قابلیت نفوذ در ابرها و شرایط اتمسفری نامناسب (مانند بارندگی) را دارند. به همین خاطر در سیستم‌های تصویربردار راداری از آن‌ها استفاده می‌شود. در شکل (الف-۱)، ترتیب امواج مختلف در طیف الکترومغناطیس به همراه طول موجشان نشان داده شده است.



شکل (الف-۱)- نمایش طیف الکترومغناطیس

[1] Far Infra-Red
[2] Short Wave Infra-Red
[3] Thermal Infra-Red

دامنه‌های مذکور علاوه بر این‌که گاهی با یکدیگر هم‌پوشانی دارند، در بعضی از کتاب‌ها و مراجع به شکل دیگری نامگذاری شده‌اند. بنابراین همیشه در برخورد با اینچنین واژه‌ها و اصطلاحاتی باید به تعاریف اولیه‌ی نویسنده‌ی اثر توجه داشت. البته قالب کلی این نامگذاری‌ها مطابق بالاست و تفاوت‌های جزیی در بخش مادون‌قرمز، باعث اختلاف میان تعاریف می‌شود.

هر بخش باریک پیوسته از طیف الکترومغناطیس را باند[1] می‌نامند. معمولاً با توجه به اینکه هر باند در کدام یک از بخش‌های طیف قرار می‌گیرد، آن باند را نامگذاری می‌کنند. هر سنجنده می‌تواند از یک تا صدها باند داشته باشد که شناخت باندهای یک سنجنده در تهیه‌ی تصاویر مناسب با هر پروژه اهمیت بسزایی دارد.

الف-۴ سنجنده

سنجنده دستگاهی است که انرژی الکترومغناطیس را دریافت کرده و انرژی دریافتی را پس از اعمال یک سری تبدیلات به صورتی قابل بازیافت (به‌طور عددی یا آنالوگ) ثبت می‌نماید. شناخت خصوصیات سنجنده‌ها، در انتخاب هرچه بهتر تصاویر مورد نیاز یک پروژه کمک می‌کند. در واقع با داشتن شناخت دقیق از خصوصیات سنجنده‌ها می‌توان به راحتی تصمیم گرفت که کدام سنجنده، داده‌های مناسب‌تری برای پروژه‌ی مورد نظر تولید می‌کند. در هر سنجنده، خصوصیاتی از قبیل قدرت تفکیک مکانی[2]، قدرت تفکیک طیفی[3]، قدرت تفکیک رادیومتریکی[4] و قدرت تفکیک زمانی[5] تاثیر بسزایی به روی تصاویر ثبت شده و در نهایت بر روی داده‌های ثبت‌شده می‌گذارد.

[1] Band
[2] Spatial Resolution
[3] Spectral Resolution
[4] Radiometric Resolution
[5] Temporal Resolution

سنجنده‌ها در دنیای سنجش از دور از دیدگاه‌های مختلف دسته‌بندی می‌شوند. شناخت این که هر سنجنده در کدام دسته قرار می‌گیرد می‌تواند به استفاده‌ی بهتر از تصاویر ماهواره‌ای موردنظر کمک کند. مثلاً سنجنده‌ها را می‌توان از لحاظ منبع انرژی یا از لحاظ طیفی تقسیم‌بندی کرد.

الف-۴-۱ انواع سنجنده‌ها از لحاظ منبع انرژی

هر شیء‌ای که دمایی بالاتر از صفر مطلق (−273 °C) دارد، از خود انرژی الکترومغناطیس ساطع می‌کند که در طول موج‌های مختلف مقدار آن متفاوت خواهد بود. خورشید یکی از بزرگترین منابع انرژی است که بشر با آن در ارتباط است. خورشید منبع گرما و نور برای کره‌ی زمین است که در فاصله‌ی ۱۵۰ میلیون کیلومتری زمین قرار دارد. از مقدار انرژی رسیده به زمین تقریباً ۳۵ درصد دوباره منعکس‌شده، ۱۷ درصد توسط اتمسفر و ۴۷ درصد هم توسط زمین و اشیای روی آن جذب می‌شود.

همان‌طوری که پیش از این اشاره شد، سنجنده‌ها دستگاه دریافت، اندازه‌گیری و ثبت انرژی الکترومغناطیس می‌باشند. سنجنده‌هایی که انرژی انعکاسی خورشید یا انرژی انتشاری اشیا را دریافت می‌کنند، سنجنده‌های غیرفعال[1] نامیده می‌شوند. در واقع این سنجنده‌ها منبع تولید انرژی نیستند و تنها انرژی منعکس شده از سطح زمین و اشیا را دریافت می‌کنند. به همین دلیل این سنجنده‌ها وابستگی زیادی به شرایط جوی دارند و تصاویر آن‌ها متاثر از اثرات اتمسفری و به‌ویژه ابرها می‌باشند. در مقابل این نوع سنجنده‌ها، سنجنده‌های فعال[2] قرار دارند که انرژی الکترومغناطیس را خود تولید کرده و انعکاس آن را دریافت و ثبت می‌کنند. طول موج‌های بلندی که سنجنده‌های فعال به‌کار می‌برند، باعث می‌شود تا وابستگی آن‌ها به عوامل جوی به پایین‌ترین حد کاهش یابد و بنابراین از این جهت برتری کاملی بر تصاویر سنجنده‌های غیرفعال دارند.

[1] De Active Sensors
[2] Active Sensors

الف-۴-۲ انواع سنجنده‌ها از لحاظ طیفی

سنجنده‌ها از لحاظ طیفی به دسته‌های مختلفی تقسیم می‌شوند. آن‌چه که در این تقسیم‌بندی مبنا قرار می‌گیرد، معمولاً تعداد باندهای سنجنده است. اولین دسته، سنجنده‌های تک‌باندی می‌باشند که با نام سنجنده‌های پانکروماتیک[1] معروف هستند. تصاویر این سنجنده‌ها معمولاً یک دامنه‌ی وسیع طیفی از بخش مریی تا مادون‌قرمز نزدیک را پوشش می‌دهند و به علت همین عرض باند وسیع برای اخذ انرژی الکترومغناطیس مشکلی نداشته و معمولاً از قدرت تفکیک مکانی بالایی برخوردارند. سنجنده‌های با تعداد باند کم را سنجنده‌های چندطیفی[2] می‌نامند. تاکنون اعداد مختلفی برای حداکثر تعداد باند سنجنده‌های چندطیفی ارایه شده است، اما حداکثر ۱۰ یا ۱۵ باند، مقبول‌ترین اعدادی است که تا به‌حال مورد استفاده قرار گرفته است. در واقع هر سنجنده‌ای که بیش از یک باند داشته باشد، چندطیفی خواهد بود. سنجنده-های چندطیفی، پرکاربردترین سنجنده‌ها در دنیای سنجش از دور می‌باشند. با افزایش تعداد باندهای سنجنده‌ها، دیگر از حیطه‌ی سنجنده‌های چندطیفی خارج شده و سنجنده‌ها وارد دسته‌ی جدیدی با نام سنجنده‌های فراطیفی[3] می‌شوند. این‌که مرز میان این دو مجموعه (سنجنده‌های چندطیفی و فراطیفی) کجاست هنوز دقیقاً مشخص نیست اما به‌طور تقریبی می‌توان سنجنده‌ای با بیش از ۳۰ باند را به عنوان سنجنده‌ای فراطیفی در نظر گرفت. سنجنده‌های فراطیفی قادرند تا ده‌ها باند طیفی را از یک منطقه جمع‌آوری نموده و به این ترتیب اطلاعات طیفی جامعی درباره‌ی اشیا و پوشش زمین در اختیار قرار دهند. قدرت تمایز اشیا در تصاویر این‌گونه سنجنده‌ها فوق‌العاده بالاست و به همین خاطر معمولاً در کارهای حساس از آن‌ها استفاده می‌شود.

[1] Panchromatic
[2] Multi Spectral
[3] Hyper Spectral

الف-۵ اتمسفر و نقش آن در سنجش از دور

به مجموعه‌ای از لایه‌های گازهای مختلف که اطراف سیاره‌ی زمین را احاطه کرده‌اند و به وسیله‌ی نیروی جاذبه‌ی زمین نگه داشته شده‌اند، اتمسفر می‌گویند. امواج الکترومغناطیس در مسیر خود از منبع انرژی به سمت اشیا از اتمسفر می‌گذرند. امواج تحت تاثیر مولکول‌ها و ذرات معلق در اتمسفر قرار گرفته و دچار تغییراتی می‌شوند. به‌طور کلی اتمسفر از دو طریق بر روی امواج الکترومغناطیس اثر می‌گذارند: پراکنش[1] و جذب[2]. پراکنش باعث انحراف موج از مسیر اصلی آن می‌شود و جذب انرژی باعث تغییر انرژی درونی مولکول‌های اتمسفر خواهد شد. تاثیر این دو نوع تعامل اتمسفر و انرژی الکترومغناطیس، در برخی از طول موج‌ها بسیار شدید است. بنابراین باید کاملاً از چگونگی مکانیسم انتقال امواج در اتمسفر آگاهی داشت. این اطلاعات در طراحی سنجنده‌ها و همچنین انجام تصحیحات بر روی تصاویر ثبت شده بسیار مهم و کارگشا هستند.

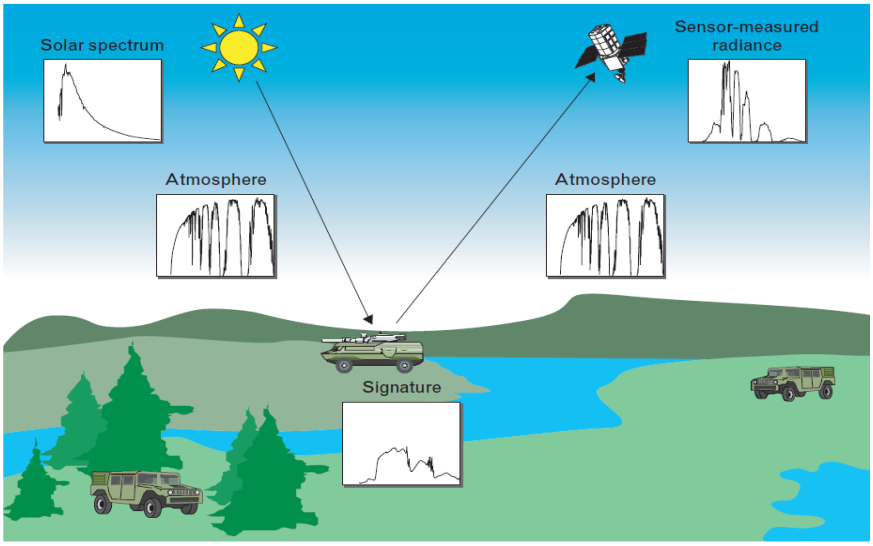
توجه کنید که در سنجش از دور، همه‌ی بخش‌های طیف الکترومغناطیس مورد استفاده قرار نمی-گیرد و مهم‌ترین علل این مساله عبارتند از: جذب و پراکنش شدید اتمسفری در برخی از طول موج‌ها، اهمیت نوع داده‌های جمع‌آوری‌شده و ملاحظات فنی. مجموعه‌ی این عوامل باعث شده است تا برخی از بخش‌های طیف الکترومغناطیس یا به‌طور کلی مورد استفاده قرار نگیرد یا به‌ندرت از آن‌ها استفاده شود. به عنوان مثال باند آبی به علت پراکنش زیاد در بسیاری از سنجنده‌های چندطیفی وجود ندارد یا باندهای مادون قرمز نزدیک، به علت کاربرد زیاد تقریباً در تمامی سنجنده‌های چندطیفی تعبیه می‌گردند.

[1] Scattering
[2] Absorption

الف-۶ ایده‌ی اصلی سنجش از دور در تصاویر تشعشع طیفی

در سنجش از دور، مقدار انرژی بازتاب‌یافته از هر جسم در طول موج موردنظر، از اهمیت بسزایی برخوردار است. اگر واکنش امواج در برخورد انرژی با یک جسم به صورت نمودار نشان داده شود به آن نمودار، در اصطلاح، منحنی بازتاب طیفی آن جسم گویند. ترکیب و ویژگی منحنی بازتاب طیفی یک جسم، می‌تواند در شناسایی ویژگی‌های طیفی آن جسم، کمک شایانی کند. دستاورد مزبور در انتخاب محدوده‌های طیفی لازم برای تهیه‌ی اطلاعات سنجش از دور با هدف (کاربرد ویژه) اهمیت فراوانی دارد. مثلا در صورتی که هدف سنجش از دور شناسایی درختان برگ‌سوزنی از درختان پهن‌برگ یک ناحیه‌ی جنگلی باشد با توجه به منحنی طیفی آن‌ها روشن می‌شود که سنجنده‌ی لازم باید در محدوده‌ی طیفی مادون قرمز حساس باشد.

ایده‌ی اصلی تصویربرداری طیفی از این واقعیت ناشی می‌شود که مقدار بازتاب طیفی هر جسم در طول موج‌های مختلف، متفاوت است. در واقع سنجنده‌ها مقدار بازتاب طیفی اجسام را در یک منطقه اندازه‌گیری می‌کنند. همان‌طور که گفته شد در یک سیستم سنجش از دور غیرفعال، منبع اصلی روشنایی خورشید است. توزیع انرژی انتشار یافته از خورشید به عنوان تابعی از طول موج در طیف الکترومغناطیس، به عنوان طیف خورشیدی شناخته می‌شود. انرژی خورشیدی از میان اتمسفر عبور کرده و شدت و توزیع طیفی آن به‌دلیل تاثیر اتمسفر تغییر می‌کند. پس از برخورد انرژی با سطح اجسام، سه واکنش عبور، جذب یا بازتاب توسط جسم صورت می‌گیرد. انرژی بازتابی و تابش‌شده توسط اجسام از میان اتمسفر عبور می‌کنند، که این عبور نیز تغییراتی را در شدت و طیف انرژی ناشی می‌شود. در نهایت، انرژی توسط سنجنده اندازه‌گیری و تبدیل به مقدار عددی برای پردازش‌های بعدی می‌گردد. در شکل (الف-۲)، نمایی از مسیر انرژی خورشید به سطح اشیا و بازتاب آن به سمت سنجنده به نمایش درآمده است.



شکل (الف-۲)- سیستم تصویربرداری طیفی و تأثیرات اتمسفر

حال با توجه به اثری که اتمسفر بر روی تصاویر ثبت‌شده می‌گذارد، لازم است هنگام پردازش و تفسیر تصاویر، اثر اتمسفر نیز مدنظر قرار گیرد. معمولاً محققین به علت اثرات منبع روشنایی، اتمسفر و پاسخ طیفی سنجنده، بازتاب طیفی خام به‌دست آمده از سنجنده‌ها را مستقیماً با طیف خام جمع‌آوری‌شده در زمان‌ها و مکان‌های دیگر مقایسه نمی‌کنند و برای رفع این مشکل از طیف بازتاب[1] استفاده می‌کنند.

$$\text{طیف بازتاب در طول موج}\lambda = \frac{\text{مقدار انرژی بازتابی}^{2}\text{ در طول موج}\lambda}{\text{مقدار انرژی برخوردی}^{3}\text{ در طول موج}\lambda} \qquad \text{(الف-۱)}$$

طیف بازتاب، برابر کسری از انرژی (نوعاً انرژی خورشید) است، که توسط یک ماده به صورت تابعی از طول موج، بازتاب داده شده است. بنابراین، اثر منابع روشنایی و اتمسفری که موج در آن انتشار یافته است، تا حد بالایی از بین می‌رود. تخمین طیف بازتاب از طیف بازتاب طیفی مشاهده شده توسط سنجنده که توسط تصحیح اتمسفری انجام می‌شود، یکی از مراحل مهم در بسیاری از کاربردهای استخراج اطلاعات از تصاویر طیفی است.

[1] Reflectance Spectrum
[2] Reflected Radiation
[3] Incident Radiation

الف-۷ مزایای سنجش از دور

امروزه مزایایی که کار با داده‌های سنجش از دور در اختیار کاربران قرار می‌دهد، باعث گردیده توجه بسیاری از کارشناسان را به خود جلب نماید. بطوری‌که می‌توان شاهد گسترش استفاده از این فن‌آوری نوظهور در دنیای امروز بود. روش‌های سنجش از دور در مقایسه با روش‌های دیگر تولید اطلاعات مانند نقشه‌برداری زمینی و آمارگیری محلی، از مزایای بسیاری برخوردار هستند. سنجش از دور علاوه بر این‌که مشکل دسترسی به محل و حضور فیزیکی در آن را که لازمه‌ی روش‌های زمینی و سنتی است مرتفع ساخته و آن را به حداقل رسانده است، با ایجاد پوشش مناسبی از منطقه‌ی مورد مطالعه، امکان دید کلی از آن را فراهم می‌سازد. با توجه به سطحی که یک تصویر ماهواره‌ای پوشش می‌دهد، در کل هزینه‌ی انجام کار پایین آمده و از لحاظ اقتصادی نیز مقرون به صرفه است، چرا که استفاده از این فن‌آوری به نیروی انسانی کم (ولی متخصص) و عملیات زمینی بسیار محدود نیاز دارد.

امروزه، اکثر داده‌ها در سنجش از دور به صورت رقومی بوده و همین مساله باعث می‌شود تا از فن-آوری موجود حداکثر استفاده برده شود. این مساله همچنین در ایجاد ارتباط ساده و آسان بین سیستم-های سنجش از دور و سیستم‌های اطلاعات جغرافیایی[1] (GIS) نیز بسیار موثر است، که در نهایت این ویژگی موجب صرفه‌جویی در زمان و هزینه و همچنین بالا بردن دقت نتایج می‌گردد.

یکی دیگر از مزایای سنجش از دور، وجود انواع متنوعی از تصاویر ماهواره‌ای با خصوصیات مکانی و طیفی مختلف است، که به محققان اجازه می‌دهد با استفاده از این قابلیت، مجموعه اطلاعات جامع‌تری را در زمان کوتاه‌تری نسبت به روش‌های مرسوم کسب کنند. وجود سنجنده‌های متعدد که تصویربرداری‌های متنوعی را در بخش‌های مختلف طیف الکترومغناطیس انجام می‌دهند، علاوه بر تنوع داده باعث گردیده است تا برای یک مکان مشخص، تصاویر متعددی در دسترس باشد و بتوان تحلیل‌های چندزمانه را نیز

[1] Geographic Information Systems

انجام داد. وجود این قابلیت‌ها در این فن‌آوری باعث شده است تا سنجش از دور از انحصار سیستم‌های نظامی و جاسوسی خارج شده و در خدمت مقاصد کاربردی و تحقیقاتی قرار گیرد.

الف-۸ کاربردهای سنجش از دور

سنجش از دور گرچه در ابتدا به عنوان یک موضوع صرفاً علمی و تحقیقاتی شناخته می‌شد اما امروزه کاملاً در اختیار شاخه‌های مختلف علوم قرار گرفته و کاربردهای گوناگونی پیدا کرده است. وجود انواع زمینه‌های کاربردی و علمی برای سنجش از دور، این فن‌آوری را در دنیای پیشرفته‌ی امروزی به عنوان یکی از مهمترین منابع جمع‌آوری داده مطرح نموده است. قابلیت‌های منحصربه‌فرد سنجش از دور باعث شده است تا بتوان این فن‌آوری را در زمینه‌های مختلفی به‌کار گرفت. نمونه‌هایی از کاربردهای تصاویر ماهواره‌ای عبارتند از: برآورد سطح یک محصول خاص، تخمین تولید و وسعت اراضی کشاورزی، تهیه‌ی نقشه‌های زمین‌شناسی، تهیه‌ی نقشه‌های خاک‌شناسی، تهیه و تنظیم شبکه‌های جاده‌ای، نظارت و کنترل توسعه‌ی شهری، مطالعه و بررسی آلودگی آب، ارزیابی ذخایر آبی، اقیانوس‌شناسی، مطالعه‌ی یخچال‌های طبیعی، بررسی خشکسالی، جنگ، کاربردهای نظامی-امنیتی، برآورد خسارت‌های ناشی از زلزله، سیل و آتش‌سوزی و غیره.

با وجود جنبه‌های فراوان کاربردی-تحقیقاتی سنجش از دور، بهتر است این نکته در نظر گرفته شود که تنها استفاده‌ی صحیح و در حد قابلیت‌های داده‌های در دسترس، می‌تواند موفقیت استفاده از تصاویر سنجش از دور را تضمین نماید. توقع‌های بالا و تصورهای نادرست از داده‌های سنجش از دور، همیشه باعث گردیده تا بسیاری از محققان و مدیران در استفاده از این فن‌آوری احتیاط کنند و شاید از مزایای آن بهره نبرند. (ونگ، ۲۰۱۰)

پیوست ب) شرایط *KKT*

مساله‌ی بهینه‌سازی (ب-۱) را در نظر بگیرید.

$$\text{minimize} \quad J(\boldsymbol{\theta}), \qquad \text{(ب-۱)}$$

$$Subject\ to \quad f_i(\boldsymbol{\theta}) \geq 0, \quad i = 1, \dots, m.$$

اگر جواب بهینه‌ی این مساله، به‌ازای $\boldsymbol{\theta}^*$ رخ دهد، آن‌گاه مجموعه‌ای از شرایط لازم تحت عنوان شرایط *KKT* وجود دارند که بایستی $\boldsymbol{\theta}^*$ در این شرایط صدق کند. این شرایط عبارتند از

$$\frac{\partial}{\partial \boldsymbol{\theta}} L\left(\boldsymbol{\theta}^*, \boldsymbol{\alpha}\right) = 0, \qquad \text{(ب-۲)}$$

$$\alpha_i \geq 0, \quad i = 1, \dots, m,$$

$$\alpha_i f_i(\boldsymbol{\theta}^*) = 0, \quad i = 1, \dots, m,$$

که در آن، α_i ضرایب لاگرانژ و L معادله‌ی لاگرانژ مربوط به این مساله‌ی بهینه‌سازی است (تیودوریدیس و کوترومباس، ۲۰۰۶؛ بازارا[1] و همکاران، ۲۰۱۰).

[1] Bazaraa M.S.

پیوست پ) مساله‌ی دوگان کمینه-بیشینه

فرض کنید دو بازیکن به نام‌های X و Y اقدام به یک بازی دو نفره می‌کنند. در این بازی، x و y به‌ترتیب، استراتژی بازیکن X و Y می‌باشد. با توجه به نتیجه‌ی بازی، بازیکن X، مقدار $F(x,y)$ به بازیکن Y پرداخت می‌کند (که اگر این مقدار منفی باشد، آن‌گاه بازیکن X برده است). تابع $F(x,y)$ را تابع بازده می‌نامند. واکنش دو بازیکن در این بازی به‌صورت زیر خواهد بود.

بازیکن X: اگر Y بداند که من قصد انتخاب استراتژی x را دارم، آن‌گاه او با انتخاب استراتژی y، تابع بازده را بیشینه می‌کند.

$$F^*(x) = max_y F(x,y),$$

بنابراین برای کمینه کردن سود Y، من بایستی استراتژی x را انتخاب کنم تا $F^*(x)$ کمینه گردد.

$$min_x\ F^*(x) = min_x\ max_y F(x,y),$$

بازیکن Y: اگر X بداند که من قصد انتخاب استراتژی y را دارم، آن‌گاه او با انتخاب استراتژی x، تابع بازده را کمینه می‌کند.

$$F_*(y) = min_x F(x,y),$$

بنابراین برای بیشینه کردن زیان X، من بایستی استراتژی y را انتخاب کنم تا $F_*(y)$ بیشینه گردد.

$$max_y\ F_*(y) = max_y\ min_x F(x,y).$$

به ازای هر x و y داریم

$$F_*(y) \equiv\ min_x F(x,y) \le F(x,y) \le\ max_y F(x,y) \equiv F^*(x).$$

بنابراین

$$max_y\ min_x F(x,y) \le min_x\ max_y F(x,y).$$

زوج (x_*, y_*) به ازای هر x و y در شرط

$$F(x_*, y) \le F(x_*, y_*) \le F(x, y_*)$$

صدق می‌کند، اگر و تنها اگر

$$max_y\ min_x F(x,y) = min_x\ max_y F(x,y) = F(x_*, y_*).$$

این مساله‌ی بهینه‌سازی که در آن کمینه‌ی $F^*(x)$ و بیشینه‌ی $F_*(y)$ به‌طور هم‌زمان رخ می‌دهند، به مساله‌ی دوگان کمینه-بیشینه معروف است و اولین بار توسط ون نیمن[1] (۱۹۲۸) اثبات گردیده است (تیودوریدیس و کوترومباس، ۲۰۰۶).

[1] Von Neumann

فهرست منابع

بیژن‌زاده، م.ح.، فرهمندراد، ش.، چاپچی، م.، خاکساری، ا.، صحت‌خواه، م.، احمدی آملی، خ.، فلکی، م.ر.، خسروپور، ف.، (۱۳۸۹)، "**جبر خطی**"، انتشارات پیام‌نور، تهران.

جانفدا، م. (۱۳۹۰)، پایان‌نامه ارشد: "بررسی روش‌های واریانس-مبنا در تحلیل حساسیت مدل‌های تعیینی"، دانشکده ریاضی، دانشگاه صنعتی شاهرود.

Apolloni, B., Malchiodi, D. and Valerio, L., (2010). Relevance regression learning with support vector machines, *Nonlinear Analysis*, **73** , 2855–2867.

Bazaraa, M.S., Jarvis, J.J. and Sherali, H.D., (2010). *Linera Programming and Network Flows*, 4th ed., Wiley, New York.

Breiman, L., (2001). Random forests, *Machine Learning*, **45**, 5–32.

Buckinx, W., Verstraeten, G. and Van den Poel, D., (2007). Predicting customer loyalty using the internal transactional database, *Expert Systems with Applications*, **32**, 125–134.

Cannizzaro, J.P. and Carder, K.L., (2006). Estimating chlorophyll-a concentrations from remote-sensing reflectance in optically shallow waters, *Remote Sensing of Environment*, **101**, 13–24.

Darecki, M., Kaczmarek, S. and Olszewski, J., (2005). SeaWiFS ocean colour chlorophyll algorithms for the southern Baltic Sea. *International Journal of Remote Sensing*, **26**, 247–260.

Durand, D., Bijaouni, J. and Cauneau, F., (2000). Optical remote sensing of shallow–water environmental parameters: A feasibility study, *Remote Sensing of Environment*, **73**, 152–161.

Fell, F. and Fischer, J., (2001). Numerical simulation of the light field in the atmosphere–ocean system using the matrix–operator method. *Journal of Quantitative Spectroscopy and Radiative Transfer*, **69**, 351–388.

Fischer, J. and Kronfeld, U., (1990). Sun-stimulated chlorophyll fluorescence, 1: Influence of oceanic properties, *International Journal of Remote Sensing*, **11**, 2125–2147.

Genuer, R., Poggi, J.M. and Tuleau-Malot, C., (2010). Variable selection using random forests, *Pattern Recognition*, **31**, 2225–2236.

Gordon, H.R., Clark, D.K., Brown, J.W., Brown, O.B., Evans, R.H. and Broenkow, W.W., (1983). Phytoplankton pigment concentrations in the Middle Atlantic Bight: comparison of ship determinations and CZCS estimates, *Applied Optics*, **22**, 20–36.

Goudarzi, N. and Shahsavani, D., (2012). Application of a random forests (RF) method as a new approach for variable selection and modelling in a QSRR study to predict the relative retention time of some polybrominated diphenylethers (PBDEs), *Analytical Methods*, **4**, 3733–3738.

Gunn, S.R., (1998). Support vector machines for classification and regression, *Technical Report*, University of Southampton. Available online at: http://users.ecs.soton.ac.uk/srg/publications/pdf/svm.pdf.

Hamel, L., (2009). *Knowledge Discovery with Support Vector Machines*, Wiley, New York.

Hancock, T., Put, R., Coomans, D., Heyden, Y.V. and Everingham, Y., (2005). A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, *Chemometrics and Intelligent Laboratory Systems*, **76**, 185–196.

Hastie, T., Tibshirani, R. and Friedman, J., (2009). *The Elements of Statistical Learning: Data Mining, Inferences, and Prediction*, 2nd ed., Springer, New York.

Heim, B., Oberhaensli, H., Fietz, S. and Kaufmann, H., (2005). Variation in Lake Baikal's phytoplankton distribution and fluvial input assessed by SeaWiFS satellite data, *Global and Planetary Change*, **46**, 9–27.

Helgee, E.A., (2010). PhD. thesis, Improving Drug Discovery Decision Making Using Machine Learning and Graph Theory in QSAR Modeling, Department of Chemistry, Gothenburg University.

Ho, T.K., (1998). The random subspace method for constructing decision forests. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **20(8)**, 832–844.

Hooke, R. and Jeeves, T.A., (1961). solution of numerical and statistical problems, *J. Assoc. Comput.*, 212–229.

Iluz, D., Yacobi, Y.Z. and Gitelson, A., (2003). Adaption of an algorithm for chlorophyll-a estimation by optical data in oligotrophic Gulf of Eliat, *International Journal of Remote Sensing*, **24**, 1157–1163.

Kocev, D., Dzeroski, S., White, M.D., Newell, G.R. and Griffioen, P., (2009). Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition, *Ecological Modeling*, **220**, 1159–1168.

Kumar, R., Sharma, J.D. and Chanda B., (2012). Writer-independent off-line signature verification using surroundedness feature, *Pattern Recognition Letters*, **33**, 301–308.

Kuncheva, L.I., Del Rio Vilas, V.J. and Rodriguez, J.J., (2007). Diagnosing scrapie in sheep: a classification experiment, *Computers in Biology and Medicine*, **37**, 1194–1202.

Li, W., Liu, L. and Gong, W., (2011). Multi-objective uniform design as a SVM model selection tool for face recognition, *Expert Systems with Applications*, **38**, 6689–6695.

Maritorena, S., O'Reilly, J. and Schieber, B.D., (2002). SeaBAM Evaluation Dataset, Available online at: http://seabass.gsfc.nasa.gov/seabam/pub/maritorena_oreilly_schieber/.

Mitchel, B.G. and Kuhra, M., (1998). *Algorithms for SeaWiFS developed with the CalCOFI dataset*. CalCOFI Report 39, California Cooperative Oceanic Fisheries Investigations Report, Lajolla, California.

Nelles, O., (2001). *Nonlinear System Identification*, Springer, Berlin.

O'Reilly, J. and Maritorena, S., (2002). SeaBAM algorithm evaluation. Available online at: http://seabass.gsfc.nasa.gov/seabam/pub/oreilly_maritorena/.

O'Reilly, J., Maritorena, S., Mitchell, B., Siegel, D., Carder, K. Garver, S., Kahru, M. and McClain, C., (1998). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research*, **103**, 24937–24953.

Pal, M., Singh, N.K. and Tiwari, N.K., (2011). Support vector regression based modeling of pier scour using field data, *Engineering Applications of Artificial Intelligence*, **24**, 911–916.

Pal, M., Singh, N.K. and Tiwari, N.K., (2011). Support vector regression based modeling of pier scour using field data, *Engineering Applications of Artificial Intelligence*, **24**, 911–916.

Scholkopf, B. and Smola, A.J., (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Massachusetts.

Schroeder, T., (2005). PhD. thesis, Fernerkundung von Wasserinhaltsstoffen in Küstengewässern mit MERIS unter Anwendung expliziter und impliziter Atmosphärenkorrekturverfahren, Freie Universität Berlin, Berlin, Germany.

Slabbinck, B., De Baets, B., Dawyndt, P. and De Vos, P., (2009). Towards large-scale FAME-based bacterial species identification using machine learning techniques, *Systematic and Applied Microbiology*, **32**, 163–176.

Steinberg, D., Golovnya, M. and Cardell, N.S., (2004). A brief overview to random Forests, Available at: http://www.salford-systems.com.

Storlie, C.B., Swiler, L.P., Helton, J.C. and Sallaberry, C.J., (2009). Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models, *Reliability Engineering and System Safety*, **94**, 1735–1763.

Su, F.C., HO, C.R., Zheng, Q., Kuo, N.J. and Chen, C.T., (2006). Satellite chlorophyll retrievals with a bipartite artificial neural network model. *International Journal of Remote Sensing*, **27**, 1563–1579.

Sugeno, M. and Yasukawa, T., (1993). A fuzzy-logic based approach to qualitative modelling. *IEEE Transactions on Fuzzy Systems*, **1**, 7–31.

Taheri Shahraiyni, H., Bagheri Shouraki, S., Fell, F., Schaale, M., Fischer, J., Tavakoli, A., Preusker, R., Tajrishy, M., Vatandoust, M. and Khodaparast, H., (2009). Application of the active learning method to the retrieval of pigment from spectral remote sensing reflectance data. *International Journal of Remote Sensing*, **30**, 1045–1065.

Theodoridis, S. and Koutroumbas, K., (2006). *Pattern Recognition*, 3rd ed., Elsevier, USA.

Vapnik, V., (1998). *Statistical Learning Theory*, Wiley, New York.

Vapnik, V., (1999). *The Nature of Statistical Learning Theory*, 2nd ed., Springer, Berlin.

Verikas, A., Gelzinis, A. and Bacauskiene, M., (2011). Mining data with random forests: a survey and results of new tests, *Pattern Recognition*, **44**, 330–349 .

Weng, Q., (2010). *Remote Sensing and GIS Integration: Theories, Methods, and Applications*, The McGraw-Hill Companies, USA.

Werdell, P.J. and Bailey, S.W., (2005). An improved in situ bio-optical dataset for ocean colour algorithm development and satellite data product validation. *Remote Sensing of Environment*, **98**, 122–140.

Whitrow, C., Hand, D.J., Juszczak, P., Weston, D. and Adams, N.M., (2009). Transaction aggregation as a strategy for credit card fraud detection, *Data Mining and Knowledge Discovery*, **18**, 30–55.

Witten, I.H., Frank, E. and Hall, M.A., (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Elsevier, USA.

Xie, Y., Li, X., Ngai, E.W.T. and Ying, W., (2009). Customer churn prediction using improved balanced random forests, *Expert Systems with Applications*, **36**, 5445–5449.

Yan, W., (2006). *Application of random forest to aircraft engine fault diagnosis*, IMACS, 468–475, Beijing, PR China.

Zhan, H., Shi, P. and Chen, C., (2003). Retrieval of oceanic chlorophyll concentration using support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 2947–2951.

Zhang, T., (2003). PhD. thesis, Retrieval of oceanic constituents with artificial neural network based on radiative transfer simulation techniques, Freie Universitat Berlin.

Zhang, T., Fell, F., Liu, Z.S., Preusker, R., Fischer, J. and He, M.X., (2003). Evaluating the performance of artificial neural network techniques for pigment retrieval from ocean colour in case I waters. *Journal of Geophysical Research*, **108**, 3286–3297.

Abstract

Water bodies are known as an effective factor on the human environment and on the other living organism. Hence, investigating the quality of water bodies is one of the most important issues in the environmental researches. For this aim, awareness of water quality parameters is necessary and inevitable. Due to some difficulties in measuring these parameters in different regions, and nowadays due to the effects of various parameters, researches try to estimate these parameters by machine learning and advanced statistical methods. According to the development of space science, researchers have used spectral radiance data. Considering the measurement error of these data, and the impact of atmosphere on the spectral radiance data, noise has always been an inseparable part of this type of data. So, finding a method for appropriate modeling under noise conditions is necessary to to convert the spectral radiance data to water quality data. In this thesis, not only the Random Forest (RF) and Support Vector Machine (SVM) are introduced, but also the performance of these methods for estimation of water quality parameters from spectral radiance data is evaluated. According to the results, utilizing of RF and SVM methods for modeling of chlorophy and pigment using NOMAD and SeaBAM data, respectively and Rrs(λ)/Rrs(555) as input variables lead to appropriate results. In the estimation of chlorophyll-a concentration using NOMAD database, SVM method lead to minimum error of MPAE among three methods (RF, SVM and Active Learning Method (ALM)) under different noise levels. In the estimation of pigment concentration using SeaBAM database, the RF and SVM methods lead to smaller RMSE than ALM, Artificial Neural Networks (ANN) and some of the empirical algorithms. In general, for MOMO database RF and SVM methods presented higher accuracy than ANN method for the estimation of water quality parameters. In general, according to the accuracy of SVM and RF and their calculation costs, it can be implied that the performance of RF is better than SVM Method in this study.

Key words: Machine Learning, Random Forests, Support Vector Machine, Water Quality Parameters, Noise, Spectral Radiance Data.

Shahrood University of Technology

Faculty of Mathematics

M.S. Thesis

Application of different machine learning methods to extract of water quality parameters from spectral radiance data

By: Masoud Afshari

Supervisors:

Dr. D. Shahsavani

Dr. H. Taheri Shahraiyni

February 2013