
بسم الله الرحمن الرحيم

دانشگاه صنعتی شاهرود

دانشکده ریاضی

گروه ریاضی کاربردی

پایان نامه جهت اخذ درجه کارشناسی ارشد آمار ریاضی

مقایسه روش‌های یادگیری ماشین با روش‌های رده‌بندی متداول جهت رده‌بندی تصاویر ماهواره‌ای ابرطیفی

دانشجو: فائضه صفری

اساتید راهنما:

دکتر داود شاهسونی

دکتر حمید طاهری شهرآیینی

استاد مشاور:

دکتر برات مجردی

بهمن ۹۱

تقدیم به گوهرهای گرانمایه زندگیم،

پدر و مادرم

آن دو فرشته‌ای که از خواسته‌هایشان گذشتند، سختی‌ها را به جان خریدند و خود را سپر بلای مشکلات و

ناملایمات کردند تا من به جایگاهی که اکنون در آن ایستاده‌ام، برسم.

تشکر و قدردانی

با نام آن که خلق کرد، هستی را از رحمت بی انتهای خویش تا انسان در قدوم خلقتش به پهنای گسترده ای به بزرگی جهان خیمه زند و به او عقل داد تا عظمت خلقت را دریابد.

اکنون که به یاری خداوند متعال این دوره تحصیلیم را به پایان رسانده ام، هر چند واژه ها را یارای آن نیست که لطف و محبت آنانی را در تمام دوران زندگیم جرعه نوش دریای مهر و محبتشان بوده ام به تصویر بکشم، اما به رسم ادب و احترام بوسه بر دستانشان زده و بر خود واجب می دانم زحمات بی شائبه پدر و مادر عزیزم و یاری خواهر نازنینم را که همواره راه گشای مشکلاتم در تمام مراحل زندگی بوده اند، ارج نهاده و مراتب تشکر قلبی و باطنی را از مهربانی های آنان ابراز دارم.

از زحمات بی دریغ استاد فرهیخته و توانمندم جناب آقایان دکتر شاهسونی و دکتر طاهری شهرآئینی که رهنمودها و نظرات ارزنده، صبر و حوصله فراوان، نقش مهمی در به ثمر رساندن این پروژه داشتند صمیمانه تشکر می کنم و سلامتی و موفقیت همیشگی این بزرگواران را از درگاه یزدان پاک خواستارم. امید آنکه این پایان نامه سزاوار دقت و محبت اساتید گرامی باشد.

در پایان از تمامی دوستان و کسانی که به نوعی مرا در به انجام رساندن این مهم یاری نموده اند، سپاسگزارم.

فائضه صفری

بهمن ۱۳۹۱

تعهد نامه

اینجانب **فائضه صفری** دانشجوی دوره کارشناسی ارشد رشته **آمار ریاضی** دانشکده **ریاضی** دانشگاه صنعتی شاهرود نویسنده پایان نامه مقایسه روش‌های یادگیری ماشین با روش‌های رده‌بندی متداول جهت رده‌بندی تصاویر ماهواره‌ای ابرطیفی تحت راهنمائی **دکتر داود شاهسونی و دکتر حمید طاهری شهرآئینی** متعهد می شوم.

- *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 پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .*
- *کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « **Shahrood University of Technology** » به چاپ خواهد رسید .*
- *حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.*
- *در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .*
- *در کلیه مراحل انجام این پایان نامه، 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 ۱۳۹۱/۱۱/۲۳

امضای دانشجو

مالکیت نتایج و حق نشر

- *کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .*
- *استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.*

چکیده

در گذشته برای ایجاد نقشه‌های رده‌بندی عوارض مختلف در سطح زمین از روش‌های مطالعه میدانی که بسیار وقت‌گیر و هزینه‌بر بود استفاده می‌شد. با ظهور ماهواره‌ها و برداشت تصاویر ماهواره‌ای، این مهم توسط تصاویر ماهواره‌ای و پردازش آنها انجام می‌شود. اما مشکل این است که دقت نتیجه‌های استخراج شده بستگی به عملکرد روش پردازش تصویر برای رده‌بندی دارد. روش‌های متداول رده‌بندی در گذشته توانسته‌اند به نحو مطلوبی این مهم را به انجام رسانند. با ظهور روش‌های جدید پیشرفته آماری و یادگیری ماشین، می‌توان امید داشت که این روش‌ها عملیات رده‌بندی تصاویر ماهواره‌ای را بهتر از روش‌های متداول انجام دهند. در این مطالعه جهت رده‌بندی کاربری اراضی مختلف از روش‌های پیشرفته آماری و یادگیری ماشین استفاده شده و عملکرد آنها با روش‌های متعارف مقایسه می‌شوند. بدین منظور از داده‌های تشعشع ابرطیفی استفاده می‌شود که درطول و عرض‌های جغرافیایی مشخصی جمع‌آوری شده‌اند. این داده‌ها شامل ۱۷۶ متغیر توضیحی (باند ماهواره) و یک متغیر پاسخ ۱۳رده‌ای است که نوع پوشش گیاهی در مختصات جغرافیایی منطقه مورد مطالعه را مشخص می‌کند. مقادیر متغیر توضیحی توسط سنجنده *AVIRIS* برداشت شده است. برای استخراج داده‌ها، از تصاویر ماهواره‌ای *LAND SAT* استفاده شده است. منطقه مطالعاتی در *KFC* (*Kennedy Flight center*) ایالت فلوریدا آمریکا واقع شده است. روش‌های پیشرفته آماری مورد مطالعه در این پایان نامه، شامل ماشین بردار پشتیبان و جنگل تصادفی می‌باشند. روش ماشین بردار پشتیبان دارای ۴ تابع کرنل مختلف است که هر یک از آنها دارای پارامترهای خاص خود می‌باشند. انتخاب نوع توابع کرنل و بهینه سازی مقادیر پارامترهای آن که با استفاده از جستجوی نقطه‌ای انجام می‌شود، دارای هزینه محاسباتی گزافی است. جنگل تصادفی، مجموعه‌ای از چندین درخت تصمیم است که ساخت آن وابسته به دو پارامتر تعداد درخت و تعداد متغیرهای مورد نیاز در رشد درخت می‌باشد که هزینه اجرای

آن بسیار کمتر از روش ماشین بردار پشتیبان است. این دو روش ذکر شده جزء روش‌های پیشرفته هستند و روش‌های متداول آماری شامل تحلیل ممیزی درجه دوم، تحلیل ممیزی خطی است که در علوم مهندسی به ترتیب روش ماکزیمم درستنمایی، فاصله ماهالونبیس نامیده می‌شوند. مدل نهایی از مقایسه روش‌های پیشرفته و متداول آماری حاصل می‌شود. نتیجه مقایسه این ۴ روش، مدل ماشین بردار پشتیبان با کرنل نرمال را مدل منتخب با دقت کلی ۰/۹۵۸ معرفی نموده که حساسیت بالایی برای تمامی رده‌ها دارد. به عبارت دیگر رده‌بندی برای تمامی ۱۳ رده را بخوبی انجام داده است.

فهرست مندرجات

فهرست شکل‌ها

فهرست جدول‌ها

فصل اول

مقدمه

در گذشته برای ایجاد نقشه‌های کاربری اراضی از روش‌های مطالعه میدانی که بسیار وقت‌گیر و هزینه‌بر بود استفاده می‌شد. با ظهور ماهواره‌ها و برداشت تصاویر ماهواره‌ای، این مهم توسط تصاویر ماهواره‌ای و پردازش آنها انجام می‌شود. اما مشکل این است که دقت نتایج استخراج شده از تصاویر ماهواره‌ای بستگی به عملکرد روش پردازش تصویر برای رده‌بندی دارد. روش‌های متداول رده‌بندی در گذشته توانسته‌اند به نحو مطلوبی این مهم را به انجام رسانند. با ظهور روش‌های جدید و پیشرفته آماری و یادگیری ماشین، می‌توان امید داشت که این روش‌ها عملیات رده‌بندی تصاویر ماهواره‌ای را بهتر از روش‌های متداول انجام دهند.

در این پایان نامه ابتدا دو روش جنگل تصادفی[۱] و ماشین بردار پشتیبان[۲] بعنوان دو روش پیشرفته آماری و یادگیری ماشین و دو روش تحلیل ممیزی درجه دوم[۳]، تحلیل ممیزی خطی[۴] بعنوان دو روش رده‌بندی متداول آماری معرفی شده و عملکرد آنها به منظور رده‌بندی داده‌های تشعشع ابر طیفی مورد ارزیابی قرار می‌گیرند.

۱-۱- کاربرد رده‌بندی در سنجش از راه دور

پیشرفت‌های نوین سخت‌افزاری و نرم‌افزاری در فن‌آوری سنجش‌ازدور و توسعه سنجنده[۵]های ابرطیفی ابزار مناسبی برای استخراج اطلاعات، شناسایی و مطالعه پدیده‌های گوناگون سطح زمین و بعضی سیارات دیگر فراهم آورده است. توانمندی بالقوه‌ی داده‌های تصویری ابرطیفی در کاربردهای گوناگون، افق گسترده‌ای پیش روی پژوهشگران سنجش‌ازدور گشوده است. به علت حجم نسبتاً زیاد این داده‌ها،

۱ Random Forest
۲ Support Vector Machine
۳ Quadratic Discriminant Analysis
۴ Linear Discriminant Analysis
۵ Sensor

بسیاری از تحقیقات انجام شده در گذشته، بر موضوع چگونگی فشرده‌سازی، ذخیره‌سازی و انتقال این نوع داده‌ها و انتخاب باندهای طیفی مناسب از میان تمامی باندهای موجود در الگوریتم‌های رده‌بندی متمرکز شده بود. اما اخیراً موضوعات تحقیقاتی چون توسعه و بهبود الگوریتم‌های شناسایی، جداسازی، تعیین ویژگی اهداف، رده‌بندی و آشکارسازی اهداف و ناهنجاری با استفاده از این داده‌ها، مورد توجه قرار گرفته است.

اطلاعاتی که داده‌های سنجش از دور در اختیار کاربران قرار می‌دهد توجه بسیاری از کارشناسان را به خود جلب کرده و باعث گسترش سطح استفاده از این فن‌آوری نوظهور در دنیای کنونی شده است. روش‌های سنجش از دور در مقایسه با روش‌های دیگر تولید اطلاعات مانند نقشه برداری زمینی، عکسبرداری هوایی و آمارگیری‌های محلی از مزایای بسیار زیادی برخوردار هستند. سنجش از دور نه تنها مشکل دسترسی به محل و حضور فیزیکی در آن را که لازمه روش‌های زمینی و سنتی است مرتفع ساخته و آن را به حداقل رسانده است، بلکه با ایجاد پوشش خوبی از منطقه مورد مطالعه امکان دید کلی از منطقه مطالعاتی را فراهم می‌سازد. با توجه به سطحی که یک تصویر ماهواره‌ای پوشش می‌دهد، می‌توان گفت که در مجموع، زمان ایجاد نقشه‌های مورد نیاز کاهش یافته و هزینه انجام کار پایین آمده و از لحاظ اقتصادی مقرون به‌صرفه است چرا که استفاده از این فن‌آوری به نیروی انسانی کم ولی متخصص (و عملیات زمینی بسیار محدود) نیاز دارد.

امروزه داده‌ها و کلیه پردازش‌ها و خروجی آنها در سنجش از دور به صورت رقومی بوده و همین مسئله باعث می‌شود تا از فن‌آوری کامپیوتری موجود حداکثر استفاده برده شود.

۱-۲- تاریخچه تحقیق

آگاهی درباره پوشش‌های گیاهی مختلف و فعالیت‌های انسانی روی زمین، بسیار ضروری به نظر می‌رسد. این اطلاعات برای طراحی و مدیریت بهتر استفاده از زمین‌ها بسیار مفید می‌باشد. نقشه‌ای که شامل این اطلاعات است، نقشه کاربری اراضی نامیده می‌شود.

در دهه‌های گذشته، نقشه کاربری اراضی بوسیله عکس‌های هوایی و روش‌های سنتی (مانند بازدید مکانی) استخراج می‌شده است که بسیار زمان‌بر و پرهزینه بودند. اخیراً، فرایندهای رقومی و رده-بندی عکس‌های ماهواره‌ای منجر به تولید نقشه کاربری اراضی رقومی شده است. تولید نقشه با استفاده از داده‌های ماهواره‌ای، زمان‌بر نبوده و دارای قدرت تفکیک مکانی مناسب، دقت کافی و توانایی پوشش ناحیه‌های پهناور است (مکلود و کونگالتون[۱]، ۱۹۹۸). نقشه‌های کاربری اراضی رقومی برای ترکیب با دیگر لایه‌های اطلاعات و تحلیل فضایی بیشتر، می‌توانند به محیط *GIS* انتقال داده می‌شوند. بنابراین، نقشه کاربری اراضی استخراج شده از تصاویر ماهواره‌ای، به‌طور آشکار بهتر از نقشه‌های استخراج شده با روش‌های سنتی است.

تاکنون، بسیاری از محققان نقشه کاربری اراضی را از عکس‌های ماهواره‌ای بوسیله تکنیک‌های رده‌بندی استخراج کرده‌اند. در سال ۱۹۷۰، تعدادی از تکنیک‌های رده‌بندی متداول مانند رده‌بندی تحلیل ممیزی درجه دوم، تحلیل ممیزی خطی و تکنیک‌های رده‌بندی متوازی السطوح برای رده‌بندی عکس‌های ماهواره‌ای معرفی شدند (اسچل[۲]، ۱۹۷۳؛ گودنوق و اسچلین[۳]، ۱۹۷۴؛ ریوز[۴] و همکاران، ۱۹۷۵؛ استراهلر[۵]، ۱۹۸۰). روش تحلیل ممیزی درجه دوم از معروف‌ترین روش‌های رده‌بندی متداول

[۱] Macloed and Congalton
[۲] Schell
[۳] Goodenough and Schlien
[۴] Reeves
[۵] Strahler

است که دارای نتایج رضایت بخش بوده و به طور گسترده‌ای برای رده‌بندی تصاویر ماهواره‌ای و تولید نقشه کاربری اراضی استفاده شده است (تونسند[۱] و همکاران، ۱۹۸۷؛ وانگ[۲]، ۱۹۹۰؛ مینگ[۳] و همکاران، ۱۹۹۳؛ دفریس و تونسند[۴]، ۱۹۹۴؛ هانسن[۵] و همکاران، ۱۹۹۶؛ لانگفورد[۶]، ۱۹۹۷؛ برودلی و فریدل[۷]، ۱۹۹۷؛ دمورات[۸]، ۱۹۹۸؛ اسکات و مارک[۹]، ۲۰۰۱؛ ابو الماگد و تانتون[۱۰]، ۲۰۰۳؛ شالابی و تاتیشی[۱۱]، ۲۰۰۷؛ هانگر و ریس[۱۲]، ۲۰۰۷؛ مانوندهار[۱۳] و همکاران، ۲۰۰۹).

در دهه ۸۰ و ۹۰ میلادی، روش‌های رده‌بندی هوشمندی مانند شبکه‌های عصبی مصنوعی و فازی معرفی و برای رده‌بندی تصاویر ماهواره‌ای و تولید نقشه کاربری اراضی بکار گرفته شدند (وارتون[۱۴]، ۱۹۸۹؛ بندیکتسون[۱۵] و همکاران، ۱۹۹۰؛ پائولا و اسچونگردت[۱۶]، ۱۹۹۵؛ گوپال و وودکوک[۱۷]، ۱۹۹۴؛ آتکینسون و تانتال[۱۸]، ۱۹۹۷؛ آبوئلگسیم و همکاران[۱۹]، ۱۹۹۶؛ گوپال و وودکوک، ۱۹۹۶؛ لوکاس[۲۰]، ۲۰۰۷). مقایسه بین تکنیک‌های رده‌بندی هوشمند و رده‌بندی تحلیل ممیزی درجه دوم نوید دهنده این مطلب است که روش‌های رده‌بندی هوشمند، توانایی تولید نقشه کاربری اراضی با دقت بالایی را دارا

۱ Townsend
۲ Wang
۳ Ming
۴ Defries and Townsend
۵ Hansen
۶ Langford
۷ Brodley and Friedl
۸ Demorate
۹ Scott and Mark
۱۰ Abou El-Magd and Tanton
۱۱ Shalaby and Tateishi
۱۲ Hanger and Rees
۱۳ Manundhar
۱۴ Wharton
۱۵ Benediktsson
۱۶ Paola and Schowngerdt
۱۷ Gopal and Woodcock
۱۸ Atkinson and Tantall
۱۹ Abuelgasim
۲۰ Lucas

هستند. برای مثال، کیوکو[۱](۱۹۹۳)، پال و ماتر[۲](۲۰۰۳)، مورتی[۳] و همکاران (۲۰۰۳) و سونار اربک[۴] و همکاران (۲۰۰۴) تکنیک رده‌بندی شبکه‌های عصبی را با رده‌بندی تحلیل ممیزی درجه دوم مقایسه کردند و نتایج نشان دادکه شبکه‌های عصبی روش رده‌بندی مناسب‌تری نسبت به روش تحلیل ممیزی درجه دوم برای تهیه نقشه کاربری اراضی است.

روش *درخت تصمیم*[۵] تکنیک رده‌بندی دیگری است که به طور گسترده برای تهیه نقشه کاربری اراضی از تصاویر ماهواره‌ای مورد استفاده قرار گرفته است (مثل، فریدل و برودلی[۶]، ۱۹۹۷؛ دفریس[۷] و همکاران، ۱۹۹۸؛ هانسن[۸] و همکاران، ۲۰۰۰؛ پال و ماتر[۹]، ۲۰۰۳).

در دهه ۲۰۰۰ میلادی، تکنیک‌های جدید یادگیری ماشین مانند ماشین بردار پشتیبان و جنگل تصادفی معرفی و به عنوان روش‌هایی با دقت بالا برای رده‌بندی تصاویر ماهواره‌ای و همچنین تولید نقشه کاربری اراضی به‌کار گرفته شدند. برای مثال، هرمس[۱۰] و همکاران (۱۹۹۹)، براوون[۱۱] و همکاران (۱۹۹۹)، هانگ[۱۲] و همکاران (۲۰۰۲)، کیوچل[۱۳](۲۰۰۳)، مرسیر و لنون[۱۴](۲۰۰۳)، پال و ماتر[۱۵](۲۰۰۴)، واسک و بندیکتسون[۱۶](۲۰۰۷)، والتون[۱۷](۲۰۰۸)، هانگ[۱۸] و همکاران(۲۰۰۸)، والتون[۱۹]،

---

۱ Civco
۲ Pal and Mather
۳ Murthy
۴ Sunar Erbek
۵ CART
۶ Friedl and Brodley
۷ Defries
۸ Hansen
۹ Pal and Mather
۱۰ Hermes
۱۱ Brown
۱۲ Huang
۱۳ Keuchel
۱۴ Mercier and Lennon
۱۵ Pal and Mather
۱۶ Waske and Benediktsson
۱۷ Walton
۱۸ Huang
۱۹ Walton and Yang

یانگ و همکاران(۲۰۰۸) و کاوزوگلو و کولکسن[۱](۲۰۰۹) از ماشین بردار پشتیبان به منظور رده‌بندی تصاویر ماهواره‌ای استفاده کردند. همچنین هام[۲] و همکاران(۲۰۰۵)، پال[۳] و همکاران(۲۰۰۵)،گیسلاسون[۴] و همکاران(۲۰۰۶)، پیترس[۵] و همکاران(۲۰۰۷)، کاتلر[۶] و همکاران(۲۰۰۷)، والتون[۷](۲۰۰۸) و گیمر[۸] و همکاران(۲۰۱۰) از روش جنگل تصادفی برای تولید نقشه کاربری اراضی استفاده نمودند.

۱-۳- اهداف و ضرورت انجام پایان نامه

تاکنون به‌وفور از روش‌های رده‌بندی متداول همچون روش تحلیل ممیزی خطی، تحلیل ممیزی درجه دوم جهت رده‌بندی تصاویر ماهواره‌ای استفاده شده است. اما با پیشرفت علوم و تکنولوژی، نیاز به نقشه‌های رده‌بندی با دقت بالاتر ضروری به نظر می‌رسد. نقشه کاربری اراضی، توزیع فضایی پوشش‌های مختلف زمین (مانند زمین‌های کشاورزی، مناطق شهری، جنگل و ...) در منطقه مطالعاتی را نشان می‌دهد. محققان علاقه‌مند به بررسی و تعیین تغییرات کاربری یا پوشش زمین در دهه‌های اخیر هستند (مایر و تورنر[۹]، ۱۹۹۴؛ نیجکمپ[۱۰]، ۱۹۹۷؛ اوستروم[۱۱]، ۱۹۹۰؛ پاری[۱۲]، ۱۹۹۰) زیرا این تغییرات می‌تواند به‌طور مستقیم یا غیر مستقیم به مسائل زیست محیطی، تغییر آب و هوا و قطع درختان مرتبط شود. بعلاوه، نوع و مقدار محصولات کشاورزی موجود و میوه‌ها می‌تواند بوسیله‌ی نقشه کاربری اراضی برآورد شود. بنابراین نقشه کاربری اراضی برای برنامه ریزی صادرات و واردات محصولات کشاورزی مفید و

۱ Kavzoglu and Colkesen
۲ Ham
۳ Pal
۴ Gislason
۵ Peters
۶ Cutler
۷ Walton
۸ Ghimire
۹ Meyer and Turner
۱۰ Nijkamp
۱۱ Ostrom
۱۲ Parry

سودمند می‌باشد. نقشه زمین همچنین برای صنعت بیمه از اهمیت بسزایی برخـوردار اسـت زیـرا مقـدار آسیبی که به محصولات کشاورزی بدلیل بلایای طبیعی مختلف وارد می‌شود را می‌توان با استفاده از این نقشه تعیین نموده و مقدار بیمه را با بالاترین دقت و به‌طور عادلانه محاسبه کرد.

اگر چه تاکنون روش‌های جنگل تصادفی و ماشین بردار پشتیبان برای رده‌بندی تصاویر ماهواره‌ای بکار برده شده‌اند، اما از این دو روش به ندرت جهت استخراج نقشه کاربری اراضی از داده ابرطیفی استفاده شده است و به ندرت این دو روش در این مهم با هم مورد مقایسه قرار گرفته‌اند. بنابراین، در این مطالعه، ماشین بردار پشتیبان و جنگل تصادفی برای رده‌بندی داده ابرطیفی مورد استفاده قرار می‌گیرند و توانایی، مزایا و معایب هر کدام از روش‌ها در این مطالعه مورد بررسی قرار می‌گیرد. بعلاوه، نتایج ماشین بردار پشتیبان و جنگل تصادفی با تکنیک‌های رده‌بندی متداول مقایسه می‌شود.

۱-۴- ساختار پایان نامه

بخش‌های مختلف پایان نامه بصورت زیر تهیه و تنظیم شده است.

در فصل دوم، پایگاه داده استفاده شده در این مطالعه معرفی شده است. معرفی دو روش جنگل تصادفی و ماشین بردار پشتیبان که جزو روش‌های پیشرفته آماری می‌باشند و همچنین روش‌های متداول آماری شامل تحلیل ممیزی درجه دوم و تحلیل ممیزی خطی در فصل سوم ارائه می‌شوند. در فصل چهارم، الگوریتم تحقیق بیان شده و سپس درآن معیارهای انتخاب روش مناسب، توضیح داده شده است. در فصل پنجم، ابتدا نتایج تجربی حاصل از هر روش ارائه شده و سپس مقایسه بین روش‌های پیشرفته و روش‌های متداول آماری انجام گرفته است.

فصل دوم

پایگاه داده‌ها

از دیرباز روش‌های مختلفی برای جمع آوری داده‌ها (به‌خصوص داده‌های مکانمند) وجود داشته است. انجام مشاهدات نجومی، نقشه برداری زمینی، هیدروگرافی، فتوگرامتری و سنجش از دور روش‌های عمده جمع آوری اطلاعات مکانمند می‌باشند.

سنجش از دور از زمره روش‌های جمع آوری داده‌ها محسوب می‌شود که در آن برداشت اطلاعات بدون تماس فیزیکی با اشیاء انجام می‌شود. در مقابل روش‌های زمینی که درآنها عامل انسانی وظیفه تفسیر و برداشت را بر عهده دارد و معمولاً در تماس مستقیم یا با فاصله کم از اشیاء انجام می‌شود، در سنجش از دور، جمع آوری داده برعهده سنجنده[1] است.

برای سنجش از دور تاکنون تعاریف بسیاری ارائه شده است که برخی از مهم‌ترین آنها عبارتند از:

- سنجش از دور دانش پردازش و تفسیر تصاویری است که حاصل ثبت تعامل انرژی الکترومغناطیس و اشیاء می‌باشند.
- سنجش از دور به معنای برداشت سطح زمین از فضا با استفاده از خصوصیات امواج الکترومغناطیس منعکس یا منتشر شده از سطح اشیاء است.

سنجش از دور علم و هنر(یا فن‌آوری) به‌دست آوردن اطلاعات درباره یک شیء منطقه یا پدیده، از طریق پردازش و آنالیز داده‌های اخذ شده بوسیله یک دستگاه (بدون تماس مستقیم با شیء منطقه یا پدیده مورد مطالعه) است.

با فراگیر شدن علم سنجش‌ازدور، از تصاویر ابرطیفی، به منظور رده‌بندی رده‌های پوشش زمین در کاربردهای کشاورزی، مدیریت بحران، محیط زیست، پایش محیطی و سیستم‌های اطلاعات مکانی استفاده می‌شد. در این موارد، قدرت تفکیک پایین طیفی تصاویر ابرطیفی برای آنالیز داده‌ها کافی بود.

[1] Sensor

در مقایسه با تصاویر چندطیفی، در تصاویر ابرطیفی، از باندهای طیفی با دقت طیفی بسیار بالا و تعداد باندهای بسیار زیاد برای جمع‌آوری داده‌ها استفاده می‌شود. این موارد، اصلی‌ترین تفاوت بین تصاویر چندطیفی و ابرطیفی در کاربردهای سنجش‌ازدوری است. به علت دقت طیفی پایین تصاویر چندطیفی، این تصاویر دارای محتوی اطلاعاتی کمتر نسبت به تصاویر ابرطیفی هستند.

سنجنده‌های ابرطیفی، سنجنده‌هایی هستند که قابلیت جمع‌آوری داده را در تعداد باندهای طیفی بسیار زیاد در خود دارند (بیش از ۲۰۰ باند طیفی) و بنابراین برای تجزیه و تحلیل این داده‌ها به روش‌های خاصی نیاز است.

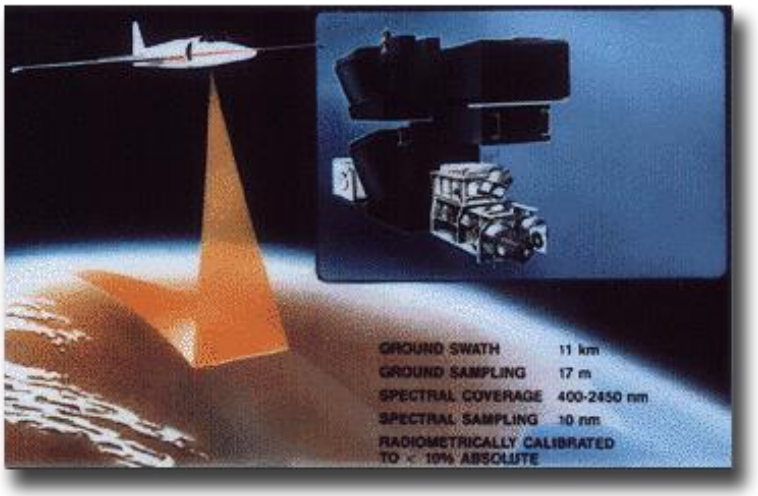
در سنجنده‌های ابرطیفی، فضای تصویر را می‌توان بصورت ماتریس ۳ بعدی در نظر گرفت که دو بعد اول پیکسل‌های تصویر است و بعد سوم، باندهای مختلف طیفی است. فضای طیفی، پاسخ طیفی رده‌های مختلف را بصورت تابعی از طول موج بیان می‌کند. در این فضا، باندهای مختلف در واقع محورهای مختصات هستند و هر پیکسل با توجه به پاسخ طیفی خود بصورت یک نقطه نمایش داده می شود.

۲-۱- تاریخچه تصویربرداری ابرطیفی

تصویربرداری ابرطیفی برای اولین بار به منظور جمع‌آوری داده‌های مناسب برای تهیه نقشه‌های زمین‌شناسی و اکتشاف معادن در اواخر دهه هفتاد میلادی در ایالات متحده آمریکا انجام شد و به سرعت توسعه و گسترش یافت. مهم‌ترین مرحله پیشرفت و تحول این فناوری، در سال ۱۹۸۷ میلادی و همزمان با ساخت اولین سیستم تصویربرداری ابرطیفی هوایی به نام *AVIRIS*[۱]، توسط مرکز *JPL* ناسا

[۱] Airborne Visible Infrared Imaging Spectrometer

صورت گرفت. اولین سنجنده ابرطیفی فضایی به نام *HYPERION* در سال ۲۰۰۰ میلادی در مـدار قـرار گرفت.



شکل(۲-۱)- نمایی از سنسور AVIRI

۲-۲- کاربرد تصاویر ماهواره‌ای

نمونه‌هایی از کاربردهای تصاویر ماهواره‌ای عبارتند از تخمین تولید و وسـعت اراضـی کشـاورزی؛ بـرآورد سطح یک محصول خاص؛ تعیین وضعیت و مشخصات انواع محصولات؛ رده‌بنـدی پوشـش و نـوع جنگـل؛ تولید نقشه‌های پوششی /کاربری؛ نقشه‌های زمین شناسی؛ نقشه‌هـای ژئومورفولـوژی؛ نقشـه‌هـای خـاک شناسی؛ نظارت و کنترل توسعه شهری؛ تنظیم شبکه جاده‌ها و دسترسی؛ مطالعه و بررسـی آلـودگی آب؛ ارزیابی ذخایر آبی؛ برف سنجی؛ مطالعه یخچال‌ها؛ اقیانوس شناسی؛ برآورد خسارت‌هـای ناشـی از زلزلـه، سیل، آتش سوزی، جنگ؛ بررسی خشکسالی؛ کاربردهای نظامی امنیتی.

گرچه این لیست به نظر طولانی می‌آید ولی هنوز بسیاری از جنبه‌های تحقیق کاربردی سـنجش از دور را شامل نمی‌شود. علاوه بر آن باید بسیاری از جنبه‌های دیگر زندگی بشری را نیز جزء کاربردهـای سنجش از دور به حساب آورد. سرانجام باید این نکته را در نظر داشت که تنها استفاده بجـا، صـحیح و در

حد قابلیت‌های داده‌های در دسترس می‌تواند موفقیت استفاده از تصاویر سنجش از دوری را تضمین نماید. توقع‌های بالا و تصورهای نادرست از داده‌های سنجش از دور همیشه باعث شده است تا بسیاری از مدیران و تصمیم گیرندگان در استفاده از این فن‌آوری احتیاط کنند و آنچنان که باید و شاید از مزایای آن بهره نبرند.

داده‌هایی که در این مسئله استفاده شده است به شرح زیر می باشد.

۲-۳- پایگاه داده‌ها

داده‌ها مربوط به ساحل غربی مرکزی فضایی کندی (*ksc*) است که در کیپ کارناوال، فلوریدا واقع شده است که زیستگاه مهمی برای گونه‌های مختلفی از پرندگان و آبزیان است. رده‌بندی پوشش زمین برای این محیط برای ایجاد پوشش گیاهی خاص با توجه به تشابه طیفی دشوار است.

برای بدست آوردن این اطلاعات طیفی از مرکز فضایی کندی، فلوریدا از تصویربرداری توسط *AVIRIS* در تاریخ ۲۳ مارس ۱۹۹۶ استفاده شده است. اطلاعات کسب شده در ۲۲۴ باند طیفی به عرض باند ۱۰ نانومتر در محدوده طیفی مرئی و مادون قرمز است. این داده‌ها که از ارتفاع حدود ۲۰ کیلومتری بدست آمده، دارای قدرت تفکیک مکانی ۱۸ متر است. متاسفانه ۴۸ باند جمع آوری شده توسط سنسور[۱] جذب آب در اتمسفر دارای خطای بالایی است و لذا قابل استفاده نمی‌باشند. لذا در تجزیه و تحلیل در این تحقیق فقط از ۱۷۶ باند باقیمانده استفاده شده است. به منظور رده‌بندی پوشش زمین، ۱۳ رده از گونه‌های مختلف پوششی در نظر گرفته شده است که اسامی آنها به همراه تعداد نمونه‌های برداشت شده در جدول (۲-۱) آمده است. این مجموعه داده بطور رایگان در آدرس اینترنتی زیر برای استفاده عموم و ارزیابی تکنیک‌های رده‌بندی قابل دسترس است.

[۱] AVIRIS

http://www.csr.utexas.edu/hyperspectral/data/KSC

جدول(۲-۱)-اسامی ۱۳ رده

	Class	No. samples
1	Scrub	761 (14.6%)
2	Willow swamp	243 (4.66%)
3	Cabbage palm hammock	256 (4.92%)
4	Cabbage palm/oak hammock	252 (4.84%)
5	Slash pine	161 (3.07%)
6	Oak/broadleaf hammock	229 (4.38%)
7	Hardwood swamp	105 (2.0%)
8	Graminoid marsh	431 (8.27%)
9	Spartina marsh	520 (9.99%)
10	Cattail marsh	404 (7.76%)
11	Salt marsh	419 (8.04%)
12	Mud flats	503 (9.66%)
13	Water	927 (17.8%)

فصل سوم

روش‌های پیشرفته و متداول آماری

۳-۱- جنگل تصادفی

یکی از روش‌های جدید مدل‌سازی، روش *جنگل تصادفی* است که جزو روش‌های یادگیری ماشین بوده و از این پس آن را به اختصار با نماد RF نمایش می‌دهیم. این روش در سال ۲۰۰۱ توسط لئو بریمن[۱] و آدله کاتلر[۲] ارائه شد به طوری که همه‌ی محققان، جنگل تصادفی را با نام این دو محقق می‌شناسند. نام این روش از نام جنگل‌های تصمیم تصادفی[۳] گرفته شده که اولین بار توسط تین کام هو[۴] پیشنهاد شد. قابلیت این روش، هنگامی که تعداد متغیرهای توضیحی زیاد است، بیشتر بروز می‌کند. اساس کار مدل جنگل تصادفی ریشه در مدل درخت تصمیم دارد، لذا لازم است ابتدا روش رده‌بندی درخت تصمیم بیان شود.

۳-۱-۱- روش درخت تصمیم

روش درخت تصمیم یک روش رده‌بندی و رگرسیونی است که با نام مستعار $CART$[۵] مشهور بوده و الگوریتم آن برای هر دو موضوع رده‌بندی و رگرسیونی تا حدودی به هم شباهت دارد. در موضوع رده‌بندی، متغیر پاسخ از نوع کیفی (اسمی) است. فرض کنید $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)\}$ مجموعه‌ای از N جفت نمونه مدل‌ساز (آموزشی) باشد که در آن $\boldsymbol{x_i} \in R^m$ بردار متغیرهای توضیحی و y_i مقدارمتغیر پاسخ i-امین مشاهده است. در این مسأله، تلاش بر این است تا فضای متغیرهای توضیحی(ابر مکعب در فضای m بعدی) به‌گونه‌ای تقسیم‌بندی شود که داده‌های واقع در هر بخش، دارای حداکثر همگونی از لحاظ رده

[۱] Leo Breiman
[۲] Adele Cutler
[۳] Random Decision Forests
[۴] Tin Kam Ho
[۵] Classification and Regression Tree

متغیر پاسخ باشند. تقسیم‌بندی فضای مذکور توسط یک الگوریتم سلسله مراتبی انجام می‌شود و در آن به سوالات زیر پاسخ داده می‌شود.

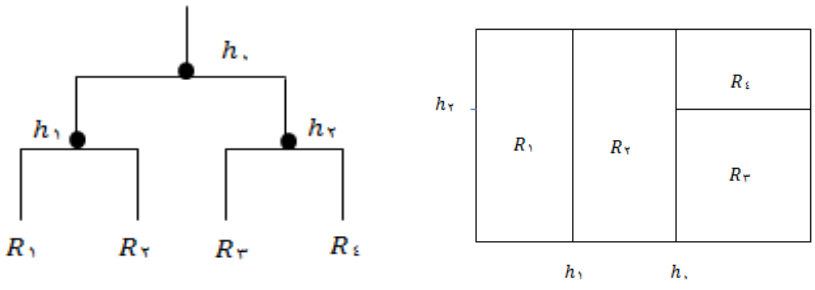
۱- تقسیم فضا در راستای محور کدام یک از متغیرهای توضیحی و در چه مقداری از آن متغیر بایستی انجام شود.

۲- معیار انتخاب راستا یاد شده چیست.

۳- تقسیم فضا، تا چه مرحله‌ای بایستی ادامه یابد.

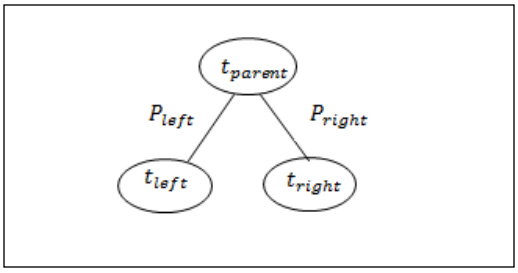
۴- پس از تقسیم فضا، رده‌بندی یک مشاهده جدید چگونه باید صورت گیرد.

به مقداری از هر متغیر که افراز در آن نقطه انجام می‌گیرد *گره* (*node*) می‌گویند. مثلاً درحالت دو متغیر توضیحی اگر تقسیم‌بندی فضا به‌صورت قاب سمت راست شکل (۳-۱) باشد آنگاه h_0, h_1, h_2 مجموعه گره‌ها را تشکیل می‌دهند. اگر به ساختار درختی این تقسیم‌بندی (قاب سمت چپ شکل ۳-۱) توجه کنیم، در واقع گره محلی است که زیرشاخه‌های درخت به یکدیگر متصل می‌شوند.



شکل (۳-۱)- (قاب سمت راست)تقسیم‌بندی فضا، (قاب سمت چپ) نمودار درختی تقسیم‌بندی فضا

لازم به ذکر است که در روش درخت تصمیم، درخت‌ها دوتایی[۱] هستند به این معنی که هر زیر فضا (شاخه) فقط می‌تواند به دو مجموعه (شاخه) تقسیم شود. این تقسیم‌بندی مطابق قانونی است که در ادامه توضیح داده خواهد شد. اگر فضای کلی داده‌ها با t_p یا t_{parent} و زیر فضاهای تقسیم شده راست و چپ به ترتیب با t_r و t_l نشان داده شوند (شکل۳-۲)، می‌توان اندازه (۳-۱) را برای این نوع تقسیم‌بندی تعریف نمود.



شکل (۳-۲)- نمایی از جداسازی الگوریتم CART

$$D\big(i(t)\big) = i(t) - \big(p_l\, i(t_l) + p_r i(t_r)\big) \qquad (۱-۳)$$

که در آن p_l و p_r به ترتیب نسبتی از کل داده‌ها هستند که در زیر فضای سمت چپ و راست حضور دارند همچنین $i(t)$ یک تابع ناخالصی[۲] است که می‌تواند یکی از گزینه‌های زیر باشد.

$$Gini - index : \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}) \qquad (۲-۳)$$

$$Misclassification - erorr : 1-\hat{p}_{mk} \qquad (۳-۳)$$

$$Cross - entropy : -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \qquad (۴-۳)$$

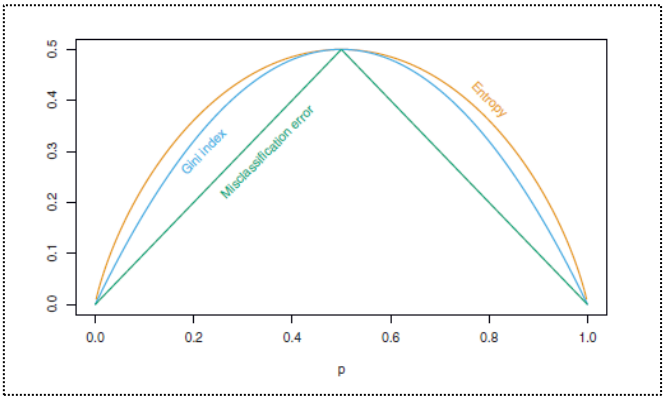
که در آن

$$\hat{p}_{mk} = \frac{1}{N_m}\sum_{m=1}^{M} I_{R_m}(\boldsymbol{x}_i \in k) \qquad (۵-۳)$$

[۱] Binary tree
[۲] Impurity Function

$$I_{R_m}(\boldsymbol{x}) = \begin{cases} ۱ , & \boldsymbol{x} \in R_m \\ ۰ , & \boldsymbol{x} \notin R_m \end{cases}$$

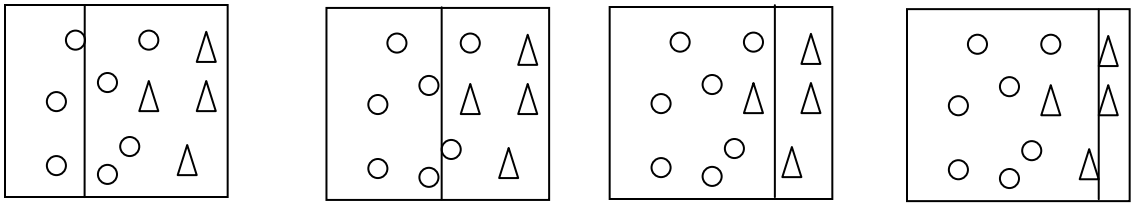
دررابطه (۳-۵) $\hat{p}_{mk}$ ، احتمال حضور داده‌های رده k-ام در زیر شاخه R_m است، N_m تعداد کل داده‌های واقع در زیر فضای R_m ، M تعداد کل زیر شاخه‌ها (فضاها) و K نشانگر تعداد کل رده‌ها می‌باشد. نمودار توابع (۳-۲) تا (۳-۴) برای دو رده، در شکل ۳-۳ نشان داده شده است. برای هر کدام از K رده، در هر زیر فضای R_m ، یک $\hat{p}_{mk}$ محاسبه می‌شود که ماکزیمم این $\hat{p}_{mk}$ ها بیانگر رده اصلی مربوط به فضای R_m ، رده K-ام است. برای حالت دو رده‌ای برای سه تابع روابط (۳-۲) تا (۳-۴) به ترتیب به صورت $۲p(۱-p)$ ، $۱ - \max\{p, ۱-p\}$ و $(-p)logp - (۱-p)\log(۱-p)$ در می‌آیند.



شکل(۳-۳)- نمودار توابع ناخالصی برای دو رده

برای پاسخ به سوال ۱ بایستی به ازای هر متغیر و به ازای تمام نقاط ممکن، افراز فضا انجام گیرد و مقدار $D(i(t))$ متناظر با آن تقسیم‌بندی محاسبه شود. مقدار ماکزیمم $D(i(t))$ تعیین کننده متغیر و مقداری از آن متغیر است که بایستی افراز فضا در این مرحله در راستای آن متغیر و از آن مقدار انجام شود. بدین ترتیب اولین تقسیم‌بندی فضا انجام شده و فضا به دو زیر فضا تقسیم شده و درخت تصمیم با دو شاخه بوجود می‌آید.

به منظور روشن شدن مطلب و پاسخ به سوال ۱، یک مثال ساده دو متغیره با دو رده k (با علامت ○) و Q (با علامت Δ) ارائه شده و نحوه تعیین نقطه تقسیم فضا تشریح می‌گردد. فرض کنید نمودار پراکنش داده‌های مسأله به‌صورت شکل (۳-۴) باشد.



شکل (۳-۴)- داده‌های مسأله دو متغیره فرضی با دو رده مختلف به همراه حالات مختلف تقسیم در جهت متغیر اول

همانطور که در شکل (۳-۴) دیده می‌شود حالات ممکن برای تفکیک فضا بطور عمودی به چهار صورت ارائه شده است. کافی است برای هر کدام از حالات مذکور، طبق رابطه (۳-۱) تابع $D(i(t))$ محاسبه شود. در هر ۴ حالت مذکور داریم.

$$i(t_p) = \left(\frac{۷}{۱۱}\right)\left(\frac{۴}{۱۱}\right) = \frac{۲۸}{۱۲۱}$$

حال برای هر کدام از افرازها، محاسبه تابع $i(t)$ بصورت زیر است:

برای حالت (۱): $i(t_r) = \left(\frac{۴}{۸}\right)\left(\frac{۴}{۸}\right) = \frac{۱۴}{۶۴}$ $\quad i(t_l) = \left(\frac{۳}{۳}\right)(۰) = ۰$

برای حالت (۲): $i(t_r) = \left(\frac{۲}{۶}\right)\left(\frac{۴}{۶}\right) = \frac{۸}{۳۶}$ $\quad i(t_l) = \left(\frac{۵}{۵}\right)(۰) = ۰$

برای حالت (۳): $i(t_l) = (۰)\left(\frac{۳}{۳}\right) = ۰$ $\quad i(t_l) = \left(\frac{۷}{۸}\right)\left(\frac{۱}{۸}\right) = \frac{۷}{۶۴}$

برای حالت (۴): $i(t_l) = (۰)\left(\frac{۲}{۲}\right) = ۰$ $\quad i(t_l) = \left(\frac{۷}{۹}\right)\left(\frac{۲}{۹}\right) = \frac{۱۴}{۸۱}$

سپس تابع $D(i(t))$ برای ۴ حالت فوق به‌ترتیب به‌صورت زیر است.

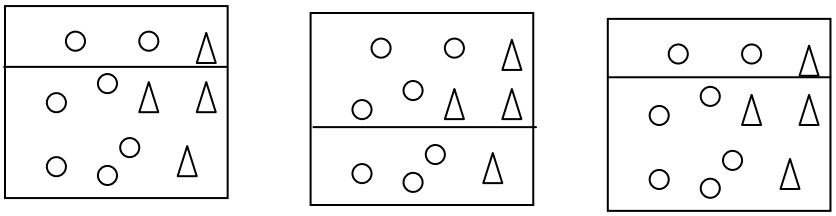
$$D(i(t_1)) = \frac{28}{121} - \left(\left(\frac{3}{11}\right)(0)\right) - \left(\left(\frac{8}{11}\right)\left(\frac{16}{64}\right)\right) = 0/049 \quad : \text{حالت اول}$$

$$D(i(t_2)) = \frac{28}{121} - \left(\left(\frac{5}{11}\right)(0)\right) - \left(\left(\frac{6}{11}\right)\left(\frac{8}{36}\right)\right) = 0/11 \quad : \text{حالت دوم}$$

$$D(i(t_3)) = \frac{28}{121} - \left(\left(\frac{8}{11}\right)\left(\frac{7}{64}\right)\right) - \left(\left(\frac{3}{11}\right)(0)\right) = 0/15 \quad : \text{حالت سوم}$$

$$D(i(t_4)) = \frac{28}{121} - \left(\left(\frac{9}{11}\right)\left(\frac{14}{81}\right)\right) - \left((0)\left(\frac{2}{11}\right)\right) = 0/09 \quad : \text{حالت چهارم}$$

با تکرار انجام این عملیات در جهت افقی تقسیماتی شبیه حالات شکل ۳-۴ نیز تکرار می‌شود.



حالت ۵ حالت ۶ حالت ۷

شکل (۳-۵)- داده‌های فرضی مساله دو متغیره با دو رده مختلف به همراه حالات مختلف تقسیم در راستای متغیر دوم

برای حالت (۵): $i(t_l) = \left(\frac{5}{8}\right)\left(\frac{3}{8}\right) = \frac{15}{64}$ $\quad i(t_r) = \left(\frac{2}{3}\right)\left(\frac{1}{3}\right) = \frac{2}{9}$

برای حالت (۶): $i(t_r) = \left(\frac{4}{7}\right)\left(\frac{3}{7}\right) = \frac{12}{49}$ $\quad i(t_l) = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right) = \frac{3}{16}$

برای حالت (۷): $i(t_r) = \left(\frac{5}{8}\right)\left(\frac{3}{8}\right) = \frac{12}{49}$ $\quad i(t_l) = \left(\frac{2}{3}\right)\left(\frac{1}{3}\right) = \frac{2}{9}$

مقادیر تابع D برای هر یک از سه حالت اخیر عبارت است از:

$$D(i(t_5)) = \frac{28}{121} - \left(\left(\frac{3}{11}\right)\left(\frac{2}{9}\right)\right) - \left(\left(\frac{8}{11}\right)\left(\frac{15}{64}\right)\right) = -0/001 \quad : \text{حالت پنجم}$$

$$D(i(t_6)) = \frac{28}{121} - \left(\left(\frac{7}{11}\right)\left(\frac{12}{49}\right)\right) - \left(\left(\frac{4}{11}\right)\left(\frac{3}{16}\right)\right) = 0/006 \quad : \text{حالت ششم}$$

$$D(i(t_7)) = \frac{28}{121} - \left(\left(\frac{8}{11}\right)\left(\frac{15}{64}\right)\right) - \left(\left(\frac{3}{11}\right)\left(\frac{2}{9}\right)\right) = -0/001 \quad : \text{حالت هفتم}$$

همانطور که قبلاً اشاره شد، راستا و نقطه‌ای برای تفکیک مناسب است که به ازای آن، تابع D ماکزیمم شود بنابراین چون $D(i(t_3)) = \max_{j=1,\dots,7} D(i(t_j))$ لذا در قدم اول، فضای متغیرها بایستی در راستای متغیر $\mathbf{x}_1$ و از نقطه t_3 انجام شود. یعنی حالت ۳ بهترین حالت برای افراز فضا است. عملیات فوق برای هر کدام از دو زیر فضای (شاخه) بوجود آمده به طور مستقل تکرار می‌شود و شاخه‌های درخت شروع به رشد می‌کند.

برای پاسخ به سوال سوم مبنی بر اینکه تقسیم‌بندی فضا تا چه مرحله‌ای باید ادامه پیدا کند و به عبارت دیگر درخت چقدر رشد پیدا کند، بایستی متذکر شد که اگر درخت خیلی بزرگ باشد پدیده بیش برازش رخ می‌دهد و درخت کوچک هم ممکن است رده‌بندی خوبی را ارائه نکند. در واقع تا جایی شاخه‌های درخت رشد پیدا می‌کنند که مقدار خطا تفاوت قابل ملاحظه‌ای نداشته باشند.

فرایند هرس کردن

در این قسمت کار هرس کردن درخت تصمیم با استفاده از فرایند *cross-validation* ارائه خواهد شد. فرایند *cross-validation* براساس بهینه مناسب بین پیچیدگی درخت و

خطای $misclassification$ است. با افزایش اندازه درخت، خطا کاهش می‌یابد و در حالت درخت ماکزیمم خطا به سمت صفر می‌رود. اما در طرف دیگر، درخت‌های تصمیم پیچیده بر روی داده مستقل ضعیف عمل می‌کنند. عملکرد درخت تصمیم روی داده مستقل قدرت پیشگوئی درست درخت نامیده می‌شود. بنابراین، کار اصلی پیداکردن بهینه مناسب بین پیچیدگی درخت و خطا است. این کار از طریق تابع هزینه-پیچیدگی بدست می‌آید.

$$R_\alpha(T) = R(T) + \alpha(\tilde{T}) \to \min_T$$

در این رابطه $R(T)$ خطای $misclassification$ درخت T، $\alpha(\tilde{T})$ اندازه پیچیدگی است که بستگی به $\tilde{T}$ (مجموع کل نودهای پایانی درخت) دارد، α-پارامتری است که از طریق دنباله‌ای در نمونه آزمون بدست می‌آید. این فرایند دفعات مختلف برای نمونه آزمون و آموزشی (به‌طور تصادفی انتخاب می‌شوند) تکرار می‌شود. این در حالی است که $cross$-$validation$ احتیاجی به اصلاح هر پارامتر ندارد. این فرایند گاهی اوقات زمان‌بر می‌باشد. لازم به ذکر است بخاطر اینکه نمونه آزمون و آموزشی به‌طور تصادفی انتخاب می‌شوند درخت نهایی ممکن است هر دفعه از دفعه قبل متفاوت باشد.

برای پاسخ به سوال چهارم مبنی بر تعیین رده بندی مشاهده جدید می‌دانیم که با اتمام مرحله قبل و تفکیک فضا به اندازه لازم، مشاهده مورد نظر بر حسب مختصاتش در یکی از نواحی واقع می‌شود. این مشاهده به رده‌ای تعلق می‌گیرد که دارای بیشترین فراوانی در آن ناحیه است.

۳-۱-۲- مدل جنگل تصادفی

حال که روش درخت تصمیم معرفی گردید می‌توان به ارائه‌ی مدل جنگل تصادفی پرداخت، چرا که اساس کار جنگل تصادفی ریشه در روش درخت تصمیم دارد. در مدل جنگل تصادفی، مجموعه‌ای از

چندین درخت تصمیم در تشکیل مدل سهیم هستند. حال به بیان دو تفاوت اصلی مدل جنگل تصادفی و مدل درخت تصمیم پرداخته می‌شود.

۱-در مدل جنگل تصادفی درخت‌های تصمیم می‌توانند کامل رشد کنند و نیازی به هرس کردن نیست.

۲- در مدل جنگل تصادفی برای جستجوی متغیر بهینه جهت تقسیم فضا، فقط تعدادی از متغیرها که به تصادف انتخاب می‌شوند، نقش دارند و بر خلاف روش درخت تصمیم، از همه متغیرها استفاده نمی‌شود.

فرض کنید $(\boldsymbol{x}_i, y_i)\ ;\ i = ۱, \dots, N$ مجموعه‌ی داده‌های آموزشی باشد که در آن $\boldsymbol{x}_i = (\boldsymbol{x}_{i۱}, \dots, \boldsymbol{x}_{iM})$ برداری از متغیرهای توضیحی و M تعداد کل این متغیرها است. اگر تعداد کل درخت های مدل با $ntree$ نشان داده شود، برای ساخت درخت i–ام $(i = ۱, \dots, ntree)$ کافیست الگوریتم زیر دنبال شود.

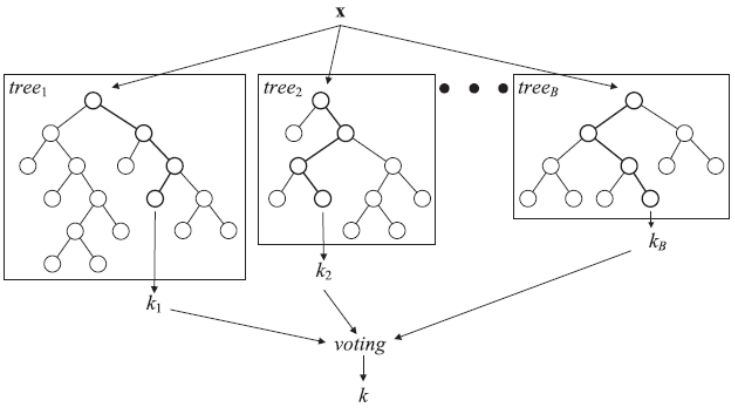
۱) در این روش نمونه اصلی به دو قسمت نمونه آموزشی و نمونه آزمون تقسیم می‌شود. سپس از نمونه آموزشی نمونه‌ای باجایگذاری به حجم N گرفته می‌شود و $\frac{۱}{۳}$ از نمونه آموزشی جدید، بعنوان نمونه خارج از کیسه (OOB) از مجموعه داده‌های آموزشی جدا می‌شود.

روش جداسازی داده‌های OOB روش نمونه گیری خودگردان است. با تکرار عملیات نمونه گیری، تعدادی مجموعه داده OOB از مجموعه آموزشی بوجود می‌آید که می‌توان برای هر مجموعه داده آموزشی یک درخت تصمیم محاسبه کرد.

۲) از بین M متغیر توضیحی، تعداد m متغیر به تصادف انتخاب می‌شوند ($m \ll M$). برای تقسیم فضای متغیرهای توضیحی به دو قسمت و بر اساس اصول درخت تصمیم، فقط از این m متغیر و نمونه‌ی N تایی انتخاب شده استفاده می‌شود تا بهترین متغیر و بهترین نقطه‌ی افراز بدست آید.

با افزایش تعداد درخت‌ها اثر بیش برازش که در روش درخت تصمیم رخ می‌دهد از بین می‌رود. معمولاً تعداد متغیرها با m نشان داده می‌شود که بایستی توسط کاربر انتخاب شود. معمولاً پیشنهاد می‌شود که در مدل رگرسیونی، $m = \frac{M}{3}$ و در مدل رده بندی $m = \sqrt{M}$ در نظر گرفته شود. وقتی که درخت ساخته شد، حال داده OOB در درخت ساخته شده قرار داده شده و میزان خطا محاسبه می‌شود. از آنجایی که تعداد درخت‌های ساخته شده بر روی نمونه های خودگردان زیاد است این میزان خطا نااریب می‌شود. از داده‌های OOB برای برآورد نااریب خطا و برآورد متغیر بااهمیت استفاده می‌شود (جنیور[۱] و همکاران، ۲۰۱۰؛ وریکاس[۲] و همکاران، ۲۰۱۰؛ تیبشیرانی و همکاران، ۲۰۰۹).

شکل (۳-۶) نمای کلی از یک جنگل تصادفی را نشان می‌دهد.



شکل (۳-۶)- نمای کلی از یک جنگل تصادفی(وریکاس و همکاران، ۲۰۱۰).

[۱] Robin Genuer
[۲] Verikas

با توجه به شکل (۳-۶) اگر یک جنگل تصادفی ساخته شده باشد چگونگی تخصیص یک رده به مشاهده جدید $\boldsymbol{x}_{\mathrm{o}} = (\boldsymbol{x}_{\mathrm{o}1}, \dots, \boldsymbol{x}_{\mathrm{o}M})$ به این صورت است که در هر درخت، این مشاهده براساس مختصات خود در یکی از گره‌های پایانی قرار گرفته و رده‌ای به آن تخصیص می‌یابد. از آنجا که رده تخصیص یافته به مشاهده $\boldsymbol{x}_{\mathrm{o}}$، در هر درخت متفاوت است لذا بیشترین فراوانی ($MODE$) این رده‌ها، رده نهایی را مشخص می‌کند.

همانطور که در روش درخت تصمیم توضیح داده شد اگر مشاهده جدید داشته باشیم به آن نود از درخت تعلق دارد که بیشترین مشاهده از آن رده را داشته باشد. در جنگل تصادفی رده‌بندی که دارای بیشترین آراء است را انتخاب می کند. در واقع رده‌بندی روی درخت‌های ساخته شده روی OOB ها انجام می‌شود. معیاری که در این قسمت برای میزان دقت در OOB استفاده می‌شود خطا است. لازم به ذکر است که تعداد درخت‌ها و تعداد متغیرهای به تصادف انتخاب شده پارامترهایی هستند که قابل تغییر بوده و توسط کاربر تعیین می‌گردند (بریمن[۱]، ۱۹۹۹).

هدف در این روش این است که آن رده‌بندی صورت گیرد که دارای مینیمم مقدار خطا باشد. در این روش به دنبال مقدار بهینه برای دو پارامتر مورد بررسی در این مدل هستیم که یکی تعداد درخت تصمیم و دیگری تعداد متغیرهایی است که در ساخت درخت تصمیم بکار می‌رود. در واقع مینیمم کردن خطا در هر درخت تصمیم معادل است با پیدا کردن مقدار بهینه برای تعداد درخت و تعداد متغیر است.

[۱] Breiman

۳-۱-۳- تعیین اهمیت متغیرها در مدل جنگل های تصادفی

در ساختار مدل جنگل تصادفی، الگوریتمی وجود دارد که می‌توان در طی آن، متغیرها را بر اساس میزان اهمیت حضورشان در مدل رتبه بندی نمود. فرض کنید خطای پیش بینی برای داده‌های OOB در درخت i- ام با نماد $ErrorOOB_i$ نشان داده شود. که در رده‌بندی برای مدل جنگل تصادفی این خطا همان *Misclassification* است. بر اساس این مقیاس می‌توان متغیرهایی که دارای نقش بیشتری در مدل هستند را بهتر شناخت. برای تعیین اهمیت متغیر j –ام، مقادیر این متغیر را به تصادف جابجا کرده و مجدداً خطا به‌ازای 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 $VI_i(\boldsymbol{X}^j)$ نشان داده و به صورت زیر تعریف می‌شود.

$$VI_i(\boldsymbol{X}^j) = \widehat{EOOB_i^j} - EOOB_i \qquad (۶-۳)$$

میزان اهمیت متغیر j–ام در مدل نهایی جنگل تصادفی به صورت زیر است.

$$VI(\boldsymbol{X}^j) = \frac{۱}{ntree}\sum_{i=۱}^{ntree}\left(VI_i(\boldsymbol{X}^j)\right) \qquad (۷-۳)$$

بدین ترتیب اهمیت هر متغیر را می‌توان در مدل نهایی بدست آورد. قابل ذکر است که مقیاس VI قابل تفسیر نبوده و فقط برای رتبه‌بندی متغیرها بر اساس اهمیت آنها به‌کار می‌رود (جنیور[۱] و همکاران،۲۰۱۰؛ وریکاس و همکاران، ۲۰۱۰).

[۱] Genuer

۳-۱-۴- مزایای روش جنگل تصادفی

۱) برای مجموعه داده‌های بزرگ بخوبی اجرا می‌شود.

۲) هزاران متغیر ورودی را می‌تواند در بر بگیرد بدون اینکه متغیری را حذف کند.

۳) یادگیری این روش آسان است.

۴) هر درخت در حالی به طور کامل رشد می‌کند و هرس نمی‌شود که مدل نهایی دچار بیش برآوردی نخواهد شد.

۳-۲- ماشین بردار پشتیبان

ماشین بردار پشتیبان[۱] یک ایده جدید در شناسایی رده‌بندی الگوهاست که از این پس به اختصار با نماد *SVM* نشان داده می‌شود. روش ماشین بردار پشتیبان در سال ۱۹۹۲ توسط واپنیک[۲] معرفی شده و بر پایه نظریه یادگیری آماری بنا گردیده است. ماشین بردار پشتیبان رده‌بندی کننده‌ای از شاخه روش‌های کرنل[۳] است که در زمره روش‌های یادگیری ماشین قرار می‌گیرد.

۳-۲-۱- اساس روش ماشین بردار پشتیبان

مبنای کاری رده‌بندی ماشین بردار پشتیبان، رده‌بندی داده‌ها به صورت خطی است. در مسائلی که داده‌ها به صورت خطی در فضای دو بعدی جداپذیر نباشند، داده‌ها با استفاده از توابع کرنل از هم جدا می‌شوند. در این بخش ما به تکنیک‌هایی برای ساختن صفحه جدا کننده بهینه که دو رده را به طور

[۱] Support Vector Machine
[۲] Vapnic Kernel method
[۳] Methods kernel

کامل از هم جدا می‌کند بحث خواهیم کرد و به حالت غیر خطی جایی که رده‌ها بصورت خطی جداپذیر نیستند تعمیم داده می‌شود (تیبشیرانی و همکاران، ۲۰۰۹).

۳-۲-۲- رده‌بندی خطی داده‌های تفکیک‌پذیر دو رده‌ای

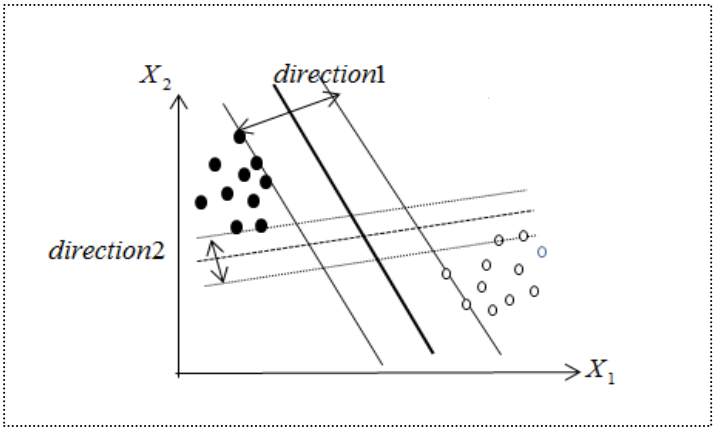
در اینجا روش SVM را برای حالتی که داده‌ها براساس رده‌هایشان کاملاً از هم جدا باشند بیان می‌کنیم. فرض کنید داده‌ها شامل N جفت:$\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_N,y_N)\}$، با $\mathbf{x}_i \in R^m$ ، همچنین $\mathbf{x}_i=\{\mathbf{x}_{i1},\mathbf{x}_{i2}\}$ شامل دو متغیر توضیحی و $y_i \in \{-1,1\}$ یعنی متغیر پاسخ شامل ۲ رده ۱ و ۱- باشد. در فضای دو بعدی صفحه مورد نظر به صورت زیر تعریف می‌شود.

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0 \qquad (3-8)$$

قانون رده‌بندی برای وقتی که بردار مشاهده جدیدی باشد بصورت زیر است.

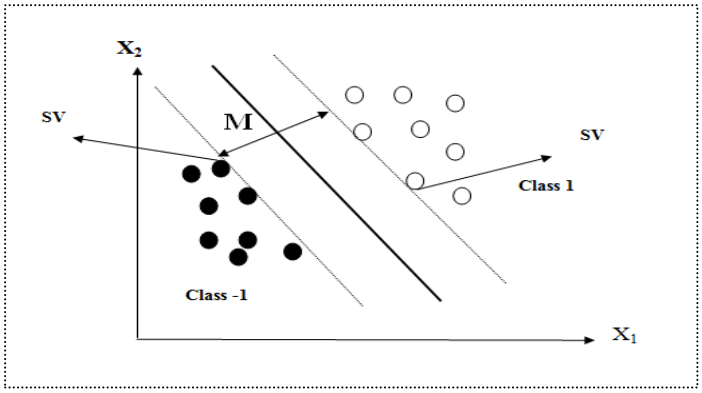
$$G(\mathbf{x}) = sign[\mathbf{w}^T\mathbf{x} + b] \qquad (3-9)$$

در رابطه (۳-۸)، $\mathbf{w} = (\mathbf{w}_1,\mathbf{w}_2)$ ضرایب متغیر توضیحی و b عرض از مبدأ است. از آنجا که $\mathbf{w}$ و b می‌تواند مقادیر متفاوت اختیار کند، لذا خطوط زیادی برای جداسازی این دو گروه داده وجود دارد. روش SVM خطی را انتخاب می‌کند که دارای کمترین خطا باشد و رده مشاهده‌ی نامعلوم $\mathbf{x}_o$ را به درستی تشخیص دهد. با توجه به متن فوق انتخاب خط جداکننده بسیار حساس است لذا خطی انتخاب می‌شود که دارای بیشترین فاصله از دو گروه باشد. زیرا با وجود بیشترین فاصله رده‌بندی به‌خوبی انجام می‌شود. همانطور که در شکل (۳-۷) دیده می‌شود خط جداکننده در دو جهت رسم شده است و هر دو خط دو گروه داده را کامل از هم جدا کرده است ولی روش SVM خط جدا کننده در جهت ۱(direction۱) را انتخاب می‌کند. زیرا بین دو گروه بیشترین فاصله وجود دارد.



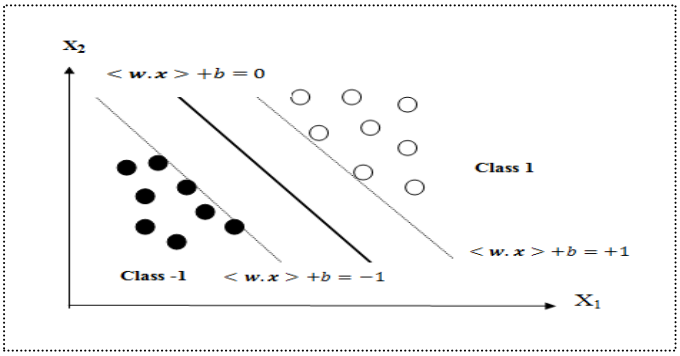
شکل(۳-۷)- نمایش خط جدا کننده بین دو گروه داده در دو جهت ۱ و ۲

فاصله بین دو خط مرزی *حاشیه* نامیده می‌شود. هدف اصلی در روش رده‌بندی SVM ماکزیمم‌سازی حاشیه است. در ضمن داده‌هایی که بر روی دو خط مرزی قرار می‌گیرند بردارهای پشتیبان (SV) نامیده می‌شوند. شکل (۳-۸) حاشیه (M) و بردارهای پشتیبان را نشان می‌دهد.



شکل (۳-۸)- نمایش بردارهای پشتیبان(SV) برای دو رده(۱و۱-)

مشاهده $\boldsymbol{x}_o$ در رده ۱- قرار می‌گیرد اگر $\boldsymbol{w^T x} + b \leq 1$، و در صورتی که $\boldsymbol{w^T x} + b \geq 1$ آنگاه در رده ۱+ قرار خواهد گرفت.



شکل (۳-۹)- خطوط جداکننده و حاشیه ، معادله خطوط

حال برای محاسبه حاشیه از خط $\boldsymbol{w^T x + b = -1}$ یک نقطه مانند $\boldsymbol{x^-}$ و از خط $\boldsymbol{w^T x + b = 1}$ یک نقطه مانند $\boldsymbol{x^+}$ در نظر گرفته می‌شود داریم.

$$\boldsymbol{w^T(x^- + Mw) + b = +1 \rightarrow w^T x^- + Mww + b = +1}$$

$$\boldsymbol{-1 + Mww = +1 \rightarrow Mww = 2 \rightarrow M = \frac{2}{\|w\|^2}}$$

در روش SVM هدف آن است که این خط به‌گونه‌ای پیدا شود که حاشیه بدست آمده از روابط مذکور که برابر این $M = \frac{2}{\|w\|^2}$ است ماکزیمم شود. قابل ذکر است که ماکزیمم سازی M معادل مینیمم سازی این رابطه $\|\boldsymbol{w}\|^2 =< \boldsymbol{w.w} >$ می‌باشد (تیبشیرانی و همکاران، ۲۰۰۹).

با توجه به اینکه ماکزیمم سازی M هم ارز، مینیمم سازی $\frac{1}{2}\|\boldsymbol{w}\|^2$ است. بنابراین مسئله بهینه سازی بصورت زیر می‌باشد.

$$Minimize \quad \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$subject\ \ to\ : y_i(\boldsymbol{w^T x_i} + b) \geq 1 \qquad (10-3)$$

این یک مسئله بهینه سازی درجه دوم است که با استفاده از روش دوگان لاگرانژ قابل محاسبه است.

حال مسئله دوگان لاگرانژ مورد نظر بصورت رابطه (۳-۱۱) می‌باشد.

$$L = \frac{۱}{۲}\|\boldsymbol{w}\|^{۲} - \sum_{i=۱}^{n} \alpha_i (y_i(\boldsymbol{w}.\boldsymbol{x} + b) - ۱) \qquad (۱۱-۳)$$

جواب بهینه این مسئله‌ی بهینه سازی بایستی در شرایط (KKT[۱]) صدق کند. از رابطه (۳-۱۱) نسبت به $\boldsymbol{w}$ و b مشتق گرفته و مساوی صفر قرار داده می‌شود.

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=۱}^{n} \alpha_i y_i \boldsymbol{x}_i = ۰ \rightarrow \boldsymbol{w} = \sum_{i \in Ns} \alpha_i y_i \boldsymbol{x}_i \qquad (۱۲-۳)$$

$$\frac{\partial L}{\partial b} = \sum \alpha_i y_i \qquad (۱۳-۳)$$

در رابطه (۳-۱۳) α_i ها ضرایب لاگرانژ هستند . α_iها مقادیر صفر یا غیر صفر می‌گیرند که مقادیر غیر صفرهمان بردار پشتیبان متناظر با بردار $\boldsymbol{x}$ مورد نظر می‌باشد. برای محاسبه α_iها کافیست در رابطه (۳-۱۱) مقدار $\boldsymbol{w} = \sum_{i \in Ns} \alpha_i y_i \boldsymbol{x}_i$ قرار داده شود آنگاه با بازنویسی، رابطه (۳-۱۴) بدست می‌آید.

$$Q(\alpha) = \frac{۱}{۲}\sum_{i=۱}^{n} \alpha_i y_j \boldsymbol{x}_i^T \sum_{j=۱}^{n} \alpha_j y_j \boldsymbol{x}_j^T + \sum_{i=۱}^{n} \alpha_i \left(۱ - y_i \left(\sum_{j=۱}^{n} \alpha_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_j + b \right) \right)$$

$$= \frac{۱}{۲}\sum_{i=۱}^{n}\sum_{j=۱}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_{i=۱}^{n} \alpha_i - \sum_{i=۱}^{n} \alpha_i \, \alpha_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_i - b \sum_{i=۱}^{n} \alpha_i y_i$$

$$= -\frac{۱}{۲}\sum_{i=۱}^{n}\sum_{j=۱}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_{i=۱}^{n} \alpha_i \qquad (۱۴-۳)$$

با توجه به این که $\sum \alpha_i y_i = 0$ است، معادله بصورت زیر خلاصه می‌شود.

$$Q(\alpha) = \sum_{i=۱}^{n} \alpha_i - \frac{۱}{۲}\sum_{i=۱}^{n}\sum_{j=۱}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \qquad (۱۵-۳)$$

[۱] Karush-Kuhn-Tucker Conditions

تابع $Q(\alpha)$ بعنوان یک حد پائینی[۱] برای تابع L در رابطه (۳-۱۱) محسوب می‌شود. برای بدست آوردن α_iها از رابطه (۳-۱۵) استفاده می‌شود (تیبشیرانی و همکاران،۲۰۰۹). برای بدست آوردن α_i ها، از الگوریتم SMO[۲] (بهینه سازی ترتیبی) می‌توان استفاده نمود این روش توسط جان پلت در سال ۱۹۹۸ ارائه شده است. پس از آن که مقادیر (α, b) که بصورت نرم افزاری قابل حل است، بدست آمد. رده‌بندی SVM را می‌توان برای مشاهده جدید بکار برد. اگر $\boldsymbol{x}$ مشاهده جدید باشد رده‌بندی آن بصورت زیر است.

$$G(\boldsymbol{x}) = sign\big(f(\boldsymbol{x}, \alpha, b)\big) \rightarrow f(\boldsymbol{x}, \alpha, b) = \sum_{i \in Ns} \alpha_i y_i \boldsymbol{x}_i^T . \boldsymbol{x} + b \qquad (۱۶-۳)$$

بعد از اینکه مقدار w بدست آمد مقدار b با جایگذاری در رابطه (۳-۸) بدست می‌آید.

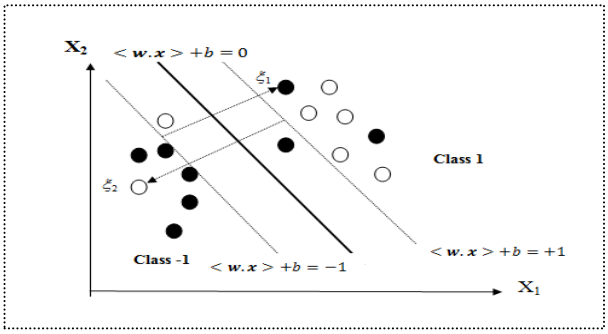
$$b = \frac{۱}{Ns} \sum_{i=Ns} (y_i - \boldsymbol{w}_i^T \boldsymbol{x}_i) \qquad (۱۷-۳)$$

در رابطه (۳-۱۷) Ns تعداد بردارهای پشتیبان می‌باشد.

۳-۲-۳- رده‌بندی خطی داده‌های تفکیک ناپذیر دو رده‌ای

یک فرض بسیار قوی در ماشین بردار پشتیبان این است که داده‌ها بصورت خطی جداپذیر هستند در حالیکه در عمل بسیاری از مواقع این فرض صحیح نمی‌باشد. در این قسمت روش ماشین بردار پشتیبان را در فضای دو بعدی برای داده‌هایی که کاملاً از هم قابل تفکیک نباشند بررسی می‌شوند. در این مواقع راه حل این است که مقدار خطا دسته‌بندی شود و این کار با معرفی متغیر کمکی (ξ) انجام می‌شود. ξ نشانگر تعداد داده‌هایی از نمونه که توسط تابع علامت G غلط ارزیابی شده‌اند. مقادیر ξ ها عبارتند از فاصله هر یک از داده‌هایی که به اشتباه در رده دیگری قرار دارند از خطوط فرضی می‌باشد برای روشن شدن این مطلب به شکل (۳-۱۰) توجه کنید.

[۱] Lower bound
[۲] Sequential Minimum Optimization



شکل (۳-۱۰)- نمایش کاربرد متغیر کمکی (ξ)در داده‌های تفکیک ناپذیر

در این حالت قانون رده‌بندی تغییری نمی‌کند تنها در روابط بهینه سازی برای بدست آوردن پارامترهای w و b ، متغیرکمکی(ξ) اضافه می‌شود. به روابط زیر توجه کنید.

$$y_i(\boldsymbol{w^T x}_i + b) \geq ۱ - \xi_i \qquad (۱۸-۳)$$

در این صورت مسئله بهینه سازی که برای مینیمم سازی $\boldsymbol{w}$ به صورت زیر می‌شود.

$$Minimize \quad L(\boldsymbol{w}) = \frac{۱}{۲}\|\boldsymbol{w}\|^۲ + c\sum \xi_i$$

$$subject \ \ to : y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) \geq ۱ - \xi_i$$

(۱۹-۳)

در این رابطه c>۰ است. مقدار مناسب c بر اساس داده‌های مسئله انتخاب می‌شود. رابطه دوگان لاگرانژ جدید برای پیدا کردن α_iها به صورت زیر است (c : تبادلی بین خط جداکننده و حاشیه است).

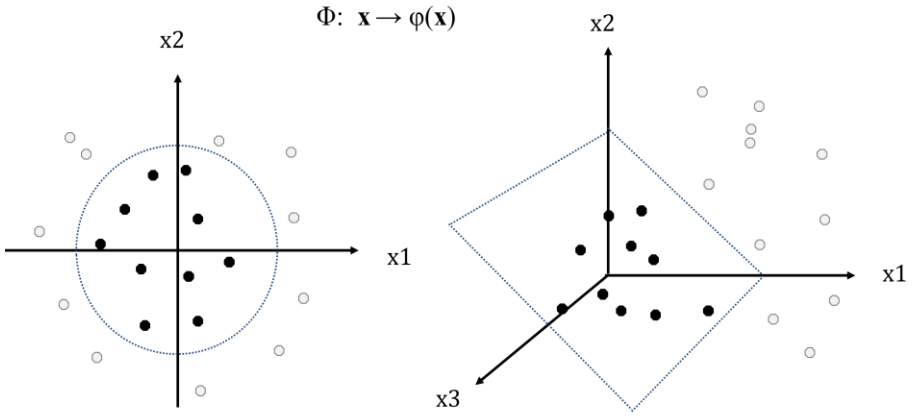
$$Q(\alpha) = \sum_i \alpha_i - \frac{۱}{۲}\sum\sum \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

$$\sum_i \alpha_i y_i = 0, 0 \leq \alpha_i \leq c \qquad (۲۰-۳)$$

۳-۲-۴- نقش توابع کرنل[۱] در روش ماشین بردار پشتیبان

هنگامی که داده‌ها در فضای اصلی خود بصورت خطی جداپذیر نباشند، بدین معنی نیست که هیچ راه دیگری برای جداسازی داده‌ها وجود ندارد. بطور مثال، ممکن است که جداسازی داده‌ها بوسیله یک منحنی چند جمله‌ای یا دوایر امکان پذیر باشد. ولی یافتن منحنی بهینه که داده‌ها را رده‌بندی کند کار مشکلی است. یک راه حل بسیار خوب برای حل این مشکل، انتقال داده‌ها به فضای دیگر و یافتن بردارهای جدیدی است که جداسازی داده‌ها در این حالت بصورت خطی امکان‌پذیر باشد و حل مساله جداسازی داده‌ها ساده‌تر می‌شود. برای انجام این کار ما یک تبدیل بصورت $z = \varphi(\boldsymbol{x})$ تعریف می‌کنیم که داده‌های ورودی $\boldsymbol{x}$ با ابعاد d به داده‌های جدیدی به نام z با ابعاد d' تبدیل می‌شوند. ما به دنبال یافتن تبدیلی خواهیم بود که امکان جداسازی خطی داده‌های آموزشی جدید $\{\varphi(\boldsymbol{x}), y\}$ بوسیله یک صفحه امکان پذیر شود (شکل ۳-۱۱). دو دسته داده‌ی موجود در نمودار سمت چپ در فضای دو بعدی قرار دارند که نمی‌توان به سادگی این دو دسته را از هم جدا کرد. حال برای این کار کافیست داده‌ها با یک نگاشت از فضای ۲ بعدی به فضای ۳ بعدی انتقال داده شوند تا بتوان به کمک یک صفحه، دو دسته داده را از هم جدا کرد.

[۱] Kernel Fuction



(ب) (الف)

شکل (۳-۱۱)- نگاشت داده‌ها در فضای دو بعدی (شکل الف) به فضای سه بعدی (شکل ب)

این روش به نظر می‌رسد که با بعضی نگرانی‌ها همراه باشد. اول اینکه، تبدیل مورد نظر برای تبدیل داده‌ها به فضای جدید چگونه باید انتخاب و تعیین شود؟ آیا باید برای هر دسته داده جدید، کار گسترده و بسیار زیادی برای یافتن این تابع تبدیل صورت گیرد؟ برای حل این مشکل با توجه به این نکته که اگر داده‌ها را به اندازه کافی به فضایی با ابعاد بالاتر منتقل شود که قابلیت جداپذیری خطی را داشته باشند، می‌توان توابع تبدیل استانداردی را برای این کار تعریف کرد. ولی انتقال داده‌ها به فضاهای بالاتر بخاطر بار محاسباتی یک مشکل و نگرانی دیگر است با انتقال داده‌ها به یک فضای با ابعاد به اندازه کافی بالا، داده‌های آموزش را می‌توان بطور کامل بصورت خطی از هم جدا کرد ولی آیا می‌توان داده‌های جدید را نیز به خوبی از هم تفکیک کرد؟ آیا طبقه بندی کننده تعریف شده دارای جواب عمومی برای داده‌های جدید نیز می‌باشد؟

رده‌بندی ماشین بردار پشتیبان قادر خواهد بود که مشکلات فوق را حل کند. پس از انتقال بردارهای ورودی $\boldsymbol{x}$ به فضای جدید بوسیله تبدیφ، صفحه جدا کننده خطی را می‌توان بصورت زیر تعریف کرد.

$$G(x) = sign\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}} \mathrm{y}_{\mathrm{i}}\, \varphi(\boldsymbol{x}_{\mathrm{i}}).\varphi(\boldsymbol{x}) + \mathrm{b}\right) = sign\left(\sum_{i=1}^{n} \alpha_i y_i \sum_{j=1}^{q} \varphi_j(\boldsymbol{x}_i).\varphi_j(\boldsymbol{x}) + b\right)$$

$$= sign\left(\sum_{i=1}^{n} \alpha_i y_i\, K_\varphi(\boldsymbol{x}_i, \boldsymbol{x}) + b\right) \qquad (۲۱-۳)$$

در رابطه فوق n نمونه‌های آموزشی هستند که به عنوان بردار پشتیبان هستند. y_i برچسب هر داده پشتیبان که مقادیر ۱- و ۱ برای رده‌های اول و دوم به ترتیب هستند نشان می‌دهد. φ تابع نگاشت داده‌ها به فضای جدید را نشان می‌دهد و تابع K را به عنوان کرنل تعریف می‌کنیم. K، تابعی است که مقدار آن به ازای دو بردار ورودی برابر با حاصلضرب داخلی داده‌های نگاشت شده، به فضای جدید می‌باشد و بصورت رابطه زیر تعریف می‌شود.

$$K_\varphi(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{j=1}^{q} \varphi_j(\boldsymbol{x}_i).\varphi_j(\boldsymbol{x}_j) \qquad (۲۲-۳)$$

با یافتن کرنل مناسب K برای حل صفحه جداکننده، نیاز به ضرب داخلی داده‌ها بطور مستقیم در فضاهای جدید که دارای ابعاد بالا هستند، نیست و کافی است که مقدار حاصل از کرنل را به ازای ورودی‌ها بدست آورده و در رابطه تابع هدف ($\boldsymbol{G}$) قرار داده شود.

در واقع قانون رده‌بندی همان تابع $\boldsymbol{G}$ است که در ابتدا تعریف شد. در چنین مواردی دیگر مدل خطی قادر به رده‌بندی مناسب نخواهد بود. در چنین مواقعی استفاده از توابع کرنل نتایج را بهبود

می‌بخشد. لازم به ذکر است که در حالت دو بعدی تابع کرنل همان $\boldsymbol{x}_i^T\boldsymbol{x}_j$ است. فرض کنید $\boldsymbol{x}_i$ و $\boldsymbol{x}_j$ دو بردار از داده‌های اصلی باشند. حال در این جا چهار تا از توابع کرنل معروف را معرفی می‌شود.

کرنل چند جمله ای $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left(coef+\gamma(\boldsymbol{x}_i^T.\boldsymbol{x}_j)\right)^{degree}$

کرنل نرمال $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=exp\left(\gamma.\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^{۲}\right)$

کرنل سیگموئید $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=tanh(\boldsymbol{\gamma}.\boldsymbol{x}_i^T\boldsymbol{x}_j+coef)$

کرنل خطی $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{x}_i^T\boldsymbol{x}_j$

$\boldsymbol{K}$ باید تابعی مثبت و متقارن باشد. $\boldsymbol{x}^T$ ترانهاده بردار $\boldsymbol{x}$ است (تیبشیرانی، ۲۰۰۹).

بنابراین مسئله دوگان لاگرانژ با استفاده از توابع کرنل به صورت زیر است. کافیست در روابط (۶-۱۲) بجای حاصلضرب داخلی بردار ورودی مشاهدات از توابع کرنل استفاده شود.

$$\boldsymbol{Q}(\boldsymbol{\alpha})=\sum_i\boldsymbol{\alpha_i}-\frac{\mathbf{1}}{\mathbf{2}}\sum\sum\boldsymbol{\alpha_i\alpha_j y_i y_j}k(\boldsymbol{x}_i,\boldsymbol{x}_j) \quad (۲۳-۳)$$

با توجه به تابع کرنل معادله تابع f به صورت زیر است

$$f(\boldsymbol{x})=\sum_i\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x})+b \quad (۲۴-۳)$$

۳-۲-۵- انتخاب کرنل مناسب

هنگامی که کاربر می‌خواهد یک ماشین بردار پشتیبان را طراحی کند با مسأله انتخاب کرنل مناسب و تعیین پارامترهای آن مواجه می‌شود. کاربر می‌تواند کرنل‌های مختلف را با پارامترهای متفاوت آزمایش

کند. سپس بهترین کرنل را با توجه به بهترین عملکرد انتخاب کند. ولی این مسأله زمان زیادی را می‌خواهد و نیز ممکن است باعث فیت شدن نتیجه شود. معمولاً در طراحی ماشین بردار پشتیبان فرض می‌شود که با مطالعات انجام گرفته توسط دیگران بهترین کرنل برای نوع خاصی از داده‌ها تعیین شده است.

نقش پارامتر C در فضای با ابعاد بزرگ واضح است. مقدار بزرگ C مانعی برای هر مقدار مثبت ξ_i می شود و به بیش برازش در مرز تصمیم گیری منجر می‌شود. مقدار کوچک C به کوچک شدن مقدار $\|w\|$ در تابع $f(x)$ می‌شود. در این حالت که مقدار C کوچک است تاکید کمتری روی مینیمم کردن خطا است و بیشتر روی ماکزیمم کردن حاشیه است. در نتیجه کاهش مقدار C افزایش حاشیه را در بر دارد. این که مقدار C را چگونه در نظر بگیریم بسته به نظر محققی دارد که از این روش استفاده می‌کند مینیمم کردن خطا یا ماکزیمم کردن حاشیه کدام برایش اهمیت داشته باشد.

۳-۲-۶- بهینه سازی پارامترهای توابع کرنل

معمولاً بهینه سازی پارامترها کاری دشوار می‌باشد برای این منظور از روش‌های مختلفی استفاده می‌شود. یکی از روش‌ها، روش جستجوی شبکه‌ای بر روی فضای مقادیر ممکن برای پارمترها می‌باشد. در این روش فضای شبکه شامل بازه‌ای از مقادیر پارامترها در نظر گرفته می‌شود و هر بار جستجو دقت رده‌بندی برآورد می‌شود (چپل و واپنیک[۱]، ۱۹۹۹؛بن-هار و واتسون[۲]، ۲۰۱۰). روش دیگر، روش بهینه سازی پارتو است که توسط جین و سندهوف[۳] در سال ۲۰۰۸ ارائه شده است. روشی دیگر، روش جدید شبیه سازی باز پخت[۴] (*SA-SVM*) می‌باشد (لین[۱] و همکاران، ۲۰۰۸؛پای و هانگ[۲]، ۲۰۰۸). ایده پیدا

۱ Chapel and Vapnic
۲ Asa Ben-Hur and Jason Weston
۳ Jin and Sendhoff
۴ Simulated annealing

کردن پارامترهای بهینه، مینیم سازی خطای کلی الگوریتم SVM می‌باشد. در این پایان نامه از ایده الکساندروس کاراتزوگلو و دیوید مایر[۳] در سال ۲۰۰۶ مبنی بر جستجوی نقطه‌ای استفاده می‌شود.

۳-۲-۷- ماشین بردار پشتیبان چند رده‌ای[۴]

در بخش‌های قبل، روش ماشین بردار پشتیبان در حالتی که متغیر پاسخ دارای دو رده بود ارائه گردید. اکنون می‌خواهیم این روش را هنگامی که تعداد رده‌های متغیر پاسخ بیش از ۲ است، بررسی کنیم بطوری که اساس کار همان حالت دو رده‌ای باشد. دو رهیافت متداول زیر در این خصوص وجود دارد.

۱. یکی- دربرابر- یکی[۵]

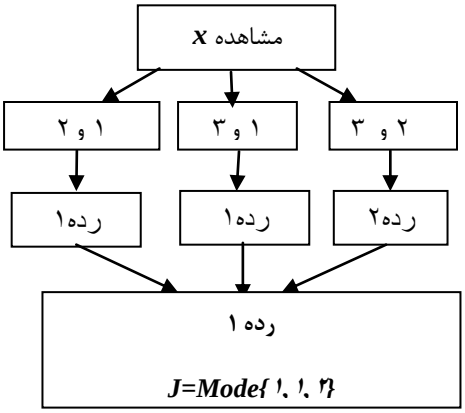
۲. یکی- دربرابر- همه[۶]

❖ رهیافت یکی- دربرابر- یکی:

دراین رهیافت، برای هر جفت از رده‌ها، روش SVM بطور مجزا انجام می‌شود. با داشتن M رده، تعداد $\frac{M(M-۱)}{۲}$ ترکیب دوتایی رده‌ها ساخته می‌شود. سپس برای هر ترکیب دو تایی، روش ماشین بردار پشتیبان اجرا شده و رده هر مشاهده با در نظر گرفتن آن ترکیب تعیین می‌گردد. رده نهایی آن مشاهده رده‌ای است که دارای ماکزیمم فراوانی در بین رده‌های برآورد شده باشد. شکل ۳-۱، رهیافت یکی- دربرابر- یکی را در حالتی که متغیر پاسخ دارای سه رده ۱، ۲ و ۳ است به عنوان مثال نشان می‌دهد.

---

۱ Lin
۲ Pai and Hang
۳ Alexandros Karatzoglou and David Mayer
۴ Multiclass SVM
۵ One-against-One
۶ One-against-All



انجام روشSVM برای ترکیب‌های مختلف دوتایی:

تعیین رده مشاهده **x**:

تعیین رده نهایی:

شکل (۳-۱۲)- نمایش رهیافت یکی در برابر یکی با سه رده

بدین ترتیب با استفاده از شاخص آماری مد(*MODE*)، مشاهده جدید به رده ۱ تعلق می‌گیرد.

❖ رهیافت یکی- دربرابر- همه:

فرض کنید $c_M, \ldots, c_2, c_1$ رده‌های متغیر پاسخ باشند. در این روش M مدل SVM ساخته می‌شود به طوری که در مدل i ام، c_i به عنوان یک رده و کلیه رده‌های $c_M, \ldots, c_{i+1}, c_{i-1}, \ldots, c_1$ به عنوان رده دوم شناخته می‌شوند. سپس رده مشاهده x تعیین شد. بدین ترتیب برای مشاهده x ، M رده برآورد می‌گردد که رده نهایی اختصاص داده شده به این مشاهده توسط شاخص مد (*MODE*) تعیین می‌شود (پلات[۱]، ۱۹۹۵).

$$i^* = arg \max_{i=1,\ldots,M} f_i(\boldsymbol{x})$$

۳-۲-۸- الگوریتم ماشین بردار پشتیبان

۱)تعیین تابع هدف

[۱] Platt

۲) حل معادله دوگان برای ماکزیمم‌سازی حاشیه و مینیمم سازی w ها و بدست آوردن α (بردار پشتیبان).

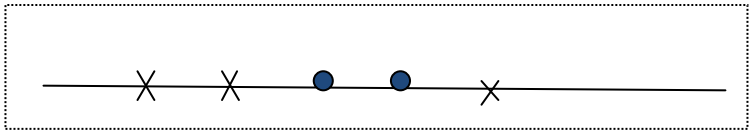
۳) استفاده از متغیر کمکی برای مواقعی که داده‌ها کاملاً خطی جدا نمی‌شوند. پارامتر c که تعادلی بین خطاها و مینیمم سازی انجام می‌دهد.*(tradeoff)*

۴) اگر نتوان داده‌ها را در فضای دو بعدی به صورت خطی جدا نمود بایستی توسط توابع کرنل به فضایی با بعد بالاتر نگاشت پیدا کنند.

مثال۱– فرض کنید داده‌های زیر را داریم با استفاده از تابع کرنل چند جمله ای با m=۲ و c=۱۰۰ داده‌ها را رده‌بندی کنید و تابع هدف را بنویسید.

X	Y
۱	۱
۲	۱
۴	-۱
۵	-۱
۶	۱

با توجه به شکل (۳-۱۳) کاملاً مشخص است که داده‌ها در فضای یک بعدی بصورت خطی جداپذیر نیستند.



شکل (۳-۱۳)- نمودار داده‌ها بر روی محور تک بعدی x

بنابراین با استفاده از تابع کرنل چندجمله‌ای درجه ۲ این داده‌ها از هم جدا می‌شوند. با حل معادله $Q(\alpha)$ و قرار دادن اطلاعات لازم داریم:

$$k(x_i, x_j) = (x_i^T x_j + 1)^2$$

$$Q(\alpha) = \sum_{i=1}^{5} \alpha_i - \frac{1}{2} \sum_{i=1}^{5} \sum_{j=1}^{5} \alpha_i \alpha_j y_i y_j (x_i x_j + 1)^2$$

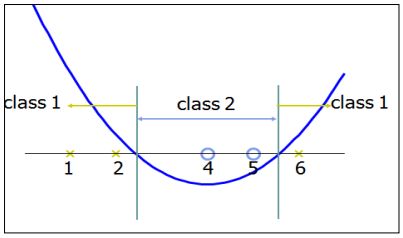
$$\sum_{i=1}^{5} \alpha_i y_i = 0$$

بدین ترتیب مقادیر α بصورت $\alpha_1 = 0, \alpha_2 = 2.5, \alpha_3 = 0, \alpha_4 = 7.33, \alpha_5 = 4.83$ می‌باشند. بنابراین بردار پشتیبان (sv) ها عبارت است از $\boldsymbol{x}_۲, \boldsymbol{x}_۴, \boldsymbol{x}_۵$ می‌باشند. حال با استفاده از این ۳ داده، تابع هدف را برای آنها رسم می‌کنیم. تابع هدف

$$f(\boldsymbol{x}) = ۲/۵ \times (۱) \times (۲\boldsymbol{x} + ۱)^۲ + ۷/۳۳ \times (-۱) \times (۵\boldsymbol{x} + ۱)^۲ + ۴/۸۳۳ \times (۱) \times$$

$$(۶\boldsymbol{x} + ۱)^۲ + b \quad , f(\boldsymbol{x}) = ۰/۶۶۷\boldsymbol{x}^۲ - ۵/۳۳۶\boldsymbol{x} + b$$

میدانیم $\boldsymbol{x}_۲$ و $\boldsymbol{x}_۵$ روی خط $\varphi(x)^T \varphi(x) + b = +1$ و $\boldsymbol{x}_۴$ روی خط $\varphi(x)^T \varphi(x) + b = -1$ قرار دارند. بنابراین با حل معادله و در نظر گرفتن اینکه $f(6) = 1, f(2) = 1, f(5) = -1$ ، سه مقدار برای b بدست می‌آید که میانگین این سه مقدار را به عنوان مقدار b نهایی در نظر می‌گیریم.سرانجام معادله خط بصورت $f(x) = ۰/۶۶۷\boldsymbol{x}^۲ - ۵/۳۳۶\boldsymbol{x} + ۹$ درمی‌آید.



شکل (۳-۱۴)-رده‌بندی داده‌های یک متغیره به روش SVM با استفاده از کرنل چندجمله‌ای درجه۲

ماشین بردار پشتیبان روشی مناسب برای مدل‌سازی داده‌ها می‌باشد. این روش کلیات رده‌بندی را با تکنیک‌های نگاشت به بعدهای بالاتر ترکیب می‌کند. فرمول‌بندی نتایج به یک مسئله بهینه سازی درجه دوم منجر می‌شود. توابع کرنل چارچوبی پیوسته فراهم می‌کند که برای بیشتر مدل‌ها بکار می‌رود و قادر به مقایسه کردن با بقیه مدل‌ها می‌باشد. در مسئله رده‌بندی عموماً این کار با ماکزیمم کردن حاشیه، یا با مینیمم کردن بردار وزنی انجام می‌شود. با حل معادله مورد نظر مجموعه‌ای از بردار پشتیبان که می‌تواند تعدادش کم باشد بدست می‌آید. آن‌ها بر روی خطوط مرزی قرار می‌گیرند و با استفاده از آن‌ها که خلاصه شده اطلاعات هستند می‌توان جداسازی را انجام داد.

۳-۳- روش تحلیل ممیزی درجه دوم

تحلیل ممیزی درجه دوم (QDA) یک روش آماری است که تاکنون در مسائل مختلف مهندسی به‌وفور مورد استفاده قرار گرفته است. این روش در علوم مهندسی به روش «ماکزیمم درستنمایی» موسوم است. در این روش براساس داده‌های مدل‌ساز (آموزشی) مدلی برای رده‌بندی داده‌ها برآورد می‌شود تا بتوان از آن برای تخمین رده مشاهداتی که در ساخت مدل نقشی ندارند (آزمون) استفاده نمود. در این روش فرض می‌شود که بردار متغیرهای توضیحی $(X_1, \dots, X_p)$ در جامعه‌ی رده k-ام دارای توزیع نرمال $N_p(\mu_k, \Sigma_k)$ است. بنیان این روش براساس قانون بیز است که در آن احتمال تعلق مشاهده x به رده k بصورت زیر تعریف می‌شود.

$$p(G = k|X = \boldsymbol{x}) = \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{l=1}^{K} f_l(\boldsymbol{x})\pi_l} \qquad (۲۵-۳)$$

که در آن $f_k(\boldsymbol{x})$ تابع چگالی احتمال توام $(X_1, \dots, X_p)$ در رده k است.

$$f_k(\boldsymbol{x}) = \frac{1}{2\pi^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} . \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu_k)^T{\Sigma_k}^{-1}(\boldsymbol{x}-\mu_k)\right) \qquad (۲۶-۳)$$

$\pi_k = p(G=k)$ توزیع پیشین رده k-ام و K تعداد کل رده‌ها است. مسلماً در عمل پس از برآورد میانگین و ماتریس کواریانس هر رده و داشتن توزیع پیشین احتمال تعلق یک مشاهده به رده k-ام، احتمال (۳-۲۵) مشخص خواهد شد. بنابراین رده برآورد شده برای هر مشاهده رده‌ای است که دارای ماکزیمم مقدار عبارت (۳-۲۵) باشد. به عبارت دیگر

$$\hat{G}(\boldsymbol{x}) = arg\max_k \delta_k(\boldsymbol{x}) \qquad (۲۶-۳)$$

که در آن $\delta_k(\boldsymbol{x})$ به‌صورت زیر است.

$$\delta_k = -\frac{1}{2}log|\Sigma_k| - \frac{1}{2}(\boldsymbol{x}-\mu_k)^T\Sigma_k^{-1}(\boldsymbol{x}-\mu_k) + log\pi_k$$

لازم به ذکر است که وقتی $\Sigma_i \neq \Sigma_j \; i \neq j = 1, \dots, K$ باشد، تابع ممیزی دارای صورت درجه دو خواهد بود. دلیل این موضوع به‌صورت زیر تشریح می‌شود.

فرض کنید مشاهده معلوم $\boldsymbol{x}$ با دو رده k و l وجود دارد. این مشاهده $\boldsymbol{x}$ متعلق به رده k است اگر

$$p(G=k|X=\boldsymbol{x}) > p(G=l|X=\boldsymbol{x}) \qquad (۲۷-۳)$$

با جایگذاری رابطه (۳-۲۵) در رابطه (۳-۲۷) نتیجه زیر حاصل می‌شود.

$$p(G=k|X=\boldsymbol{x}) = \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{i=1}^{K} f_i(\boldsymbol{x})\pi_i} > p(G=l|X=\boldsymbol{x}) = \frac{f_l(\boldsymbol{x})\pi_l}{\sum_{i=1}^{K} f_i(\boldsymbol{x})\pi_i} \qquad (۲۸-۳)$$

در رابطه (۳-۲۸) عبارت $\sum_{i=1}^{K} f_i(\boldsymbol{x})\pi_i$ از هر دو طرف ساده می‌شود سپس از دو طرف لگاریتم گرفته می‌شود.

$$log\frac{f_k(\boldsymbol{x})\pi_k}{f_l(\boldsymbol{x})\pi_l} = log\frac{f_k(\boldsymbol{x})}{f_l(\boldsymbol{x})} + log\frac{\pi_k}{\pi_l} > ۰ \qquad (۲۹-۳)$$

در رابطه (۳-۲۹) همانطور که در قبل گفته شد π_k احتمال پیشین رده k-ام است. با جایگذاری تابع چگالی نرمال به‌جای جمله $f(\boldsymbol{x})$ داریم.

$$\left(log\pi_k - \frac{۱}{۲}(\boldsymbol{x}-\mu_k)^T\Sigma_k^{-۱}(\boldsymbol{x}-\mu_k) - \frac{۱}{۲}log|\Sigma_k|\right) -$$

$$\left(log\pi_l - \frac{۱}{۲}(\boldsymbol{x}-\mu_l)^T\Sigma_l^{-۱}(\boldsymbol{x}-\mu_l) - \frac{۱}{۲}log|\Sigma_l|\right) > ۰ \qquad (۳۰-۳)$$

در حالتی که تمامی احتمالات پیشین برای رده‌ها یکی باشد جمله $log\frac{\pi_k}{\pi_l}$ حذف می‌شود. وجود عبارت $(x-\mu_k)^T\Sigma_k^{-۱}(x-\mu_k)$ یک فرم درجه دوم است.

۳-۴- روش تحلیل ممیزی خطی

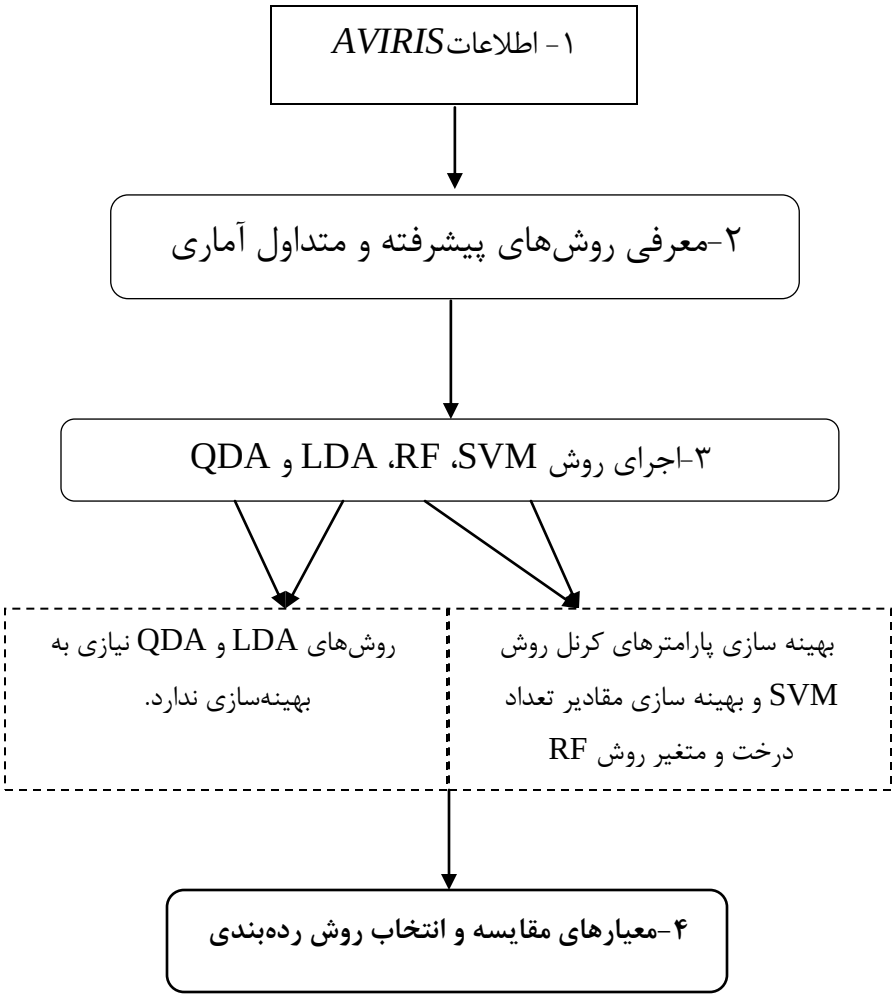
روش تحلیل ممیزی خطی (LDA) حالت خاصی از روش درجه دوم است که در آن فرض میشود که ماتریس واریانس-کواریانس تمامی رده‌ها برابر است $\Sigma_i = \Sigma \ , i = ۱,..,N$ (در عمل، این برابری توسط آزمون‌های فرض قابل بررسی است). بنابراین در رابطه (۳-۳۳) جملات درجه دوم $\boldsymbol{x}^T\Sigma^{-۱}\boldsymbol{x}$ حذف شده و صورت خطی زیر بدست می‌آید.

$$\delta_k = x^T\Sigma^{-۱}\mu_k - \frac{۱}{۲}\mu_k^T\Sigma^{-۱}\mu_k + log\pi_k \qquad (۳۱-۳)$$

فصل چهارم

الگوریتم تحقیق

در این فصل الگوریتم تحقیق با استفاده از اطلاعات جمع‌آوری شده تشریح می‌گردد. در شکل (۴-۱) الگوریتم این تحقیق نشان داده شده است. توضیحات مربوط به قسمت‌های مختلف این الگوریتم در بخش بعدی این فصل ارائه می‌گردد.



شکل (۴-۱)- الگوریتم تحقیق

۴-۱- جمع آوری اطلاعات مکانی کاربری اراضی

در این تحقیق، اطلاعات موجود در مسئله، داده‌های تشعشع طیفی هستند که در طول و عرض جغرافیایی مشخص جمع آوری شده‌اند. این داده ها شامل ۱۷۶ متغیر توضیحی و یک متغیر پاسخ است که پارامتر مورد رده‌بندی را در یک مختصات جغرافیایی مشخص می‌کند. هر متغیر توضیحی متناظر با یک طول موج مشخص است که مقدار انعکاس سطح زمین را در همان مختصات جغرافیایی و آن طول موج، تعیین می‌کند. مقادیر متغیر توضیحی از طریق داده‌های ماهواره‌ای جمع‌آوری شده است. همچنین از تصاویر *Land Sat* بهره گرفته شده است. در این مسئله متغیر پاسخ شامل ۱۳رده است.

همانطور که در فصل دوم گفته شد تعداد داده‌های ابرطیفی مجموعاً برابر با ۵۲۱۱ است که این مجموعه را به طور تصادفی به دومجموعه داده "آموزشی" (به نسبت ۶۷ درصد) و داده "آزمون" (به نسبت ۳۳ درصد) تقسیم شده است. بنابراین از مجموع کل داده‌ها ۳۵۷۷ تا برای مجموعه آموزشی و ۱۵۴۴ تا برای مجموعه آزمون در نظر گرفته شده است. بعد از تقسیم‌بندی داده‌ها به ارائه خلاصه‌ای از روش‌های پیشرفته آماری و متداول آماری پرداخته می‌شود.

۴-۲- معرفی روش‌های رده‌بندی پیشرفته و متداول آماری

۴-۲-۱- رده‌بندی به روش ماشین بردار پشتیبان (*SVM*)

این روش بر اساس محاسبه ابر صفحه جداکننده کار رده‌بندی را انجام می‌دهد. معادله ممیزکننده به صورت زیر است.

$$G(\boldsymbol{x}) = sign\left(\sum_{i=1}^{n} \alpha_i y_i \, K_\varphi (\boldsymbol{x_i}, \boldsymbol{x}) + b \right) \qquad (1-4)$$

در رابطه (۴-۱)، α_i ضرایب معادله لاگرانژ و K تابع کرنل می‌باشد. مهمترین مسئله‌ای که کاربر با آن رو برو است، مسأله انتخاب کرنل مناسب و تعیین پارامترهای آن می‌باشد. کاربر می‌تواند کرنل‌های مختلف را با پارامترهای متفاوت آزمایش کند. سپس بهترین کرنل را با توجه به بهترین عملکرد، داشتن کمترین خطا، انتخاب می‌شود.

۴-۲-۲- رده‌بندی به روش جنگل تصادفی (RF)

اساس روش جنگل تصادفی ریشه در روش درخت تصمیم دارد. در واقع در روش درخت تصمیم تنها یک درخت تصمیم تشکیل می‌شود و مدل نهایی هم بر اساس همان یک درخت شکل می‌گیرد، در حالیکه در مدل جنگل تصادفی، مجموعه‌ای از چندین درخت تصمیم در تشکیل مدل سهیم هستند.

۴-۲-۳- رده‌بندی به روش تحلیل ممیزی درجه دوم (QDA)

در این روش که در علوم مهندسی به ماکزیمم درستنمایی معروف است. فرض بر این است که متغیرهای توضیحی در رده i-ام دارای توزیع $N(\mu_i, \Sigma_i)$ می‌باشند. در این روش $\Sigma_i \neq \Sigma_j$ است و براساس محاسبه احتمال تعلق یک نمونه به یک رده خاص می‌باشد.

$$D_i = Ln(\pi_i) - \left[\frac{1}{2} Ln(|\mathbf{\Sigma_i}|)\right] - \left[(\boldsymbol{x} - \mu_i)^T \left(\Sigma_i^{-1}\right)(\boldsymbol{x} - \mu_i)\right] \qquad (۴-۲)$$

با توجه به رابطه (۴-۲) مشاهده جدید در صورتی به رده j تعلق می‌گیرد که بیشترین مقدار D_i را داشته باشد.

$$\mathrm{j} = \arg \max_{\mathrm{i}=1,\ldots,\mathrm{M}} \mathrm{D_i}$$

۴-۲-۴- رده‌بندی به روش تحلیل ممیزی خطی (*LDA*)

تفاوت این روش با روش تحلیل ممیزی درجه دوم در ماتریس کواریانس است. بعبارت دیگر $\Sigma_{i=1}^{M} = \Sigma$ می‌باشد.

$$D_i = (\boldsymbol{x} - \mu_i)^{T} \left(\boldsymbol{\Sigma^{-1}}\right) (\boldsymbol{x} - \mu_i) \qquad (۳-۴)$$

با توجه به رابطه (۴-۳)، مشاهده جدید به رده j تعلق می‌گیرد وقتی که بیشترین مقدار D_i را دارا باشد.

$$j = \arg \max_{i=1,\ldots,M} D_i$$

۴-۳- اجرای روش‌های پیشرفته و متداول

مرحله بعدی اجرای روش‌های ماشین بردار پشتیبان، جنگل تصادفی و روش‌های متداول آماری با استفاده از داده‌های آموزشی و انتخاب مدل مناسب می‌باشد. ابتدا برای انتخاب مدل در روش ماشین بردار پشتیبان، ۴ کرنل مورد نظر است که هر کدام از کرنل‌ها دارای پارامترهایی می‌باشند. بهینه سازی پارامترها مطابق الگوریتم جستجوی نقطه‌ای انجام می‌شود. الگوریتم جستجوی نقطه‌ای مجموعه داده‌های آموزشی را به دو مجموعه آموزشی ثانویه و مجموعه اعتبارسنجی به نسبت $\frac{۲}{۳}$ و $\frac{۱}{۳}$ تقسیم می‌کند. قابل ذکر است که برای اجرای دو روش *SVM* و *RF* به ترتیب از بسته‌های نرم‌افزاری *e۱۰۷۱* و *RandomForest* در نسخه‌ی ۲.۱۵.۱ نرم‌افزار *R* استفاده شده است. برای هر یک از مقادیر پارامترها بازه‌ای در نظر گرفته می‌شود، سپس برای هر یک از توابع کرنل با توجه به مقادیر مختلف پارامترها، مدل‌های زیادی ساخته می‌شود. از بین این مدل‌ها مدلی با مقادیر بهینه انتخاب می‌شود که دارای کمترین مقدار *Misclassification* باشد. سپس مدل بدست آمده با استفاده از داده‌های آزمون مورد ارزیابی قرار می‌گیرد. بدین ترتیب برای هر یک از توابع کرنل مدلی با بهترین مقادیر پارامتر ایجاد

می‌شود و در نهایت از بین ۴ مدل، مدلی که دارای بیشترین دقت کلی و حساسیت باشد انتخاب می‌گردد.

مدل جنگل تصادفی شامل دو پارامتر تعداد درخت و تعداد متغیر می‌باشد. برای بهینه سازی مقادیر دو پارامتر، بازه‌ای از مقادیر برای هر یک در نظر گرفته می‌شود. قابل ذکر است که بازه در نظر گرفته شده برای تعداد متغیر باید شامل مقدار پیش فرض $\sqrt{M}$ باشد. هر یک از این مدل‌ها دارای خطای OOB می‌باشند. سرانجام مدلی انتخاب می‌شود که دارای کمترین خطای OOB باشد. در نهایت برای ارزیابی عملکرد مدل انتخاب شده از داده‌های آزمون استفاده می‌شود.

برای روش‌های متداول آماری پارامتری برای بهینه سازی وجود ندارد. برای هر یک از این ۲ روش ابتدا مدل با داده‌های آموزشی و سپس با استفاده از داده‌های آزمون مورد ارزیابی قرار می‌گیرد. برای اجرای این دو روش از نرم‌افزار SAS استفاده شده است.

در روش‌های متداول آماری تنها روش تحلیل ممیزی خطی دارای جواب می‌باشد. همانطور که در فصل ۳ گفته شد در روش تحلیل ممیزی خطی برای تمامی رده‌ها از یک ماتریس کواریانس استفاده می‌شود که آن ماتریس دارای وارون می‌باشد. ولی در روش تحلیل ممیزی درجه دوم برای هر کدام از رده‌ها ماتریس کواریانس مجزا استفاده می‌شود. که وارون ماتریس کواریانس وجود ندارد زیرا بین ستون‌های ماتریس (باندهای ماهواره) همبستگی بالایی وجود دارد. بدین ترتیب ۳ مدل با استفاده از روش‌های SVM، RF و LDA بدست می‌آید.

۴-۴- معیارهای انتخاب روش رده‌بندی

انتخاب روش رده‌بندی با توجه به مقدار خطا و دقت[۱] کلی انجام می‌شود. در مسئله رده‌بندی مهمترین مسئله، محاسبه دقت کلی[۲] توسط ماتریس پیش بینی می‌باشد. ماتریس مربعی پیش بینی $P_{k\times k}$ که K تعداد کل رده‌هاست به صورت زیر تعریف می‌شود.

P_{ij} = تعداد مشاهدات متعلق به رده i ام که مدل پیش‌بینی آنها را متعلق به رده j ام پیش‌بینی کرده‌است

بنابراین P_{ii}، یعنی عناصر روی قطر اصلی، مبیین تعداد داده‌هایی هستند که رده آن‌ها به درستی تشخیص داده شده است. از این ماتریس دو معیار دقت کلی و حساسیت[۳] (دقت هر رده) را می‌توان محاسبه نمود. معیار حساسیت، درستی رده‌بندی داده‌ها را نشان می‌دهد. معیار حساسیت را با نام دیگر (*True Positive*) هم نشان می‌دهند. فرمول معیار حساسیت (*TP*) برابر است با:

$$TPrate = TP/(TP + FN)$$

که TP تعداد داده‌هایی است که درست رده‌بندی شده‌اند وFN (*Fals Negative*) تعدادی از داده‌هایی است که اشتباه رده‌بندی شده‌اند. فرمول دقت کلی به زبان ساده به‌صورت زیر می‌باشد.

$$Overall - accuracy = \frac{\sum_{i=1}^{K} p_{ii}}{N}$$

در رابطه فوق، p_{ii} عناصر روی قطر اصلی، K تعداد کل رده‌ها و N تعداد کل نمونه است (داسکالاکی[۴] و همکاران ،۲۰۰۶) و (پروست[۵] و همکاران، ۱۹۹۷).

[۱] Accuracy
[۲] Overall- accuracy
[۳] sensitivity
[۴] Daskalaki
[۵] Provost

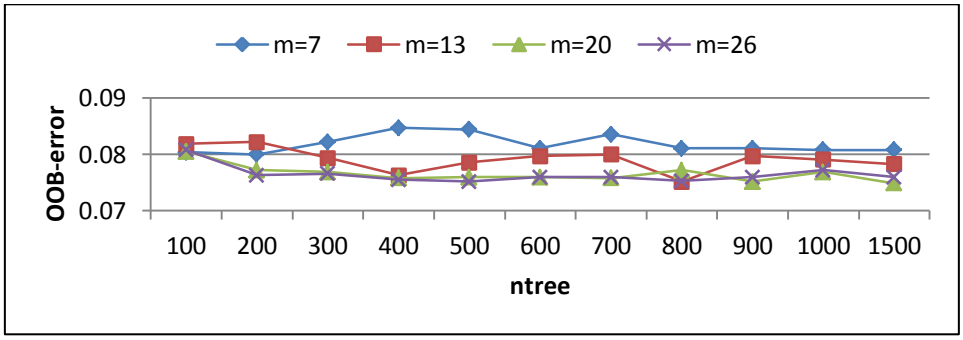
فصل پنجم

ارایه و تحلیل نتایج

در این فصل نتایج پیاده‌سازی روش‌های رده‌بندی مختلف بر روی داده‌ها ارائه می‌شود. نتایج بدست آمده مورد تحلیل قرار گرفته و روش رده‌بندی مناسب انتخاب می‌شود.

۵-۱- نتایج حاصل از اجرای روش جنگل تصادفی

ابتدا روش جنگل تصادفی را بر روی داده‌های آموزشی اجرا می‌کنیم. همانطور که در فصل ۳ توضیح داده شد، در روش جنگل تصادفی بایستی مقدار بهینه برای دو پارامتر تعداد متغیرها در هر نود از درخت (m) وتعداد درخت ($ntree$) تعیین شود. این کار با در نظر گرفتن دامنه‌ای برای m انجام می‌شود. از آنجایی که مقدار پیشنهادی m برابر $\sqrt{M}$ است لذا بازه‌ای شامل $\sqrt{M}$ را برای مقادیر m در نظر می‌گیریم. در شکل (۵-۱) مقادیر خطای داده‌های OOB به ازاء مقادیر مختلف m و $ntree$ نشان داده شده است. همانطور که در شکل (۵-۱) مشاهده می‌شود خطای OOB در بازه (۰/۰۷۵۵,۰/۰۸۴۷) قرار داشته و مقدار آن به ازای m=۲۶ کمترین است. همچنین به ازای m=۲۶ ، تعداد درخت برابر با ۵۰۰ حاوی کمترین مقدار خطا است.



شکل(۵-۱)- نمودارخطای OOB برحسب تعداد درخت و تعداد متغیر به تصادف انتخاب شده، با استفاده از روش RF برای داده‌های آموزشی.

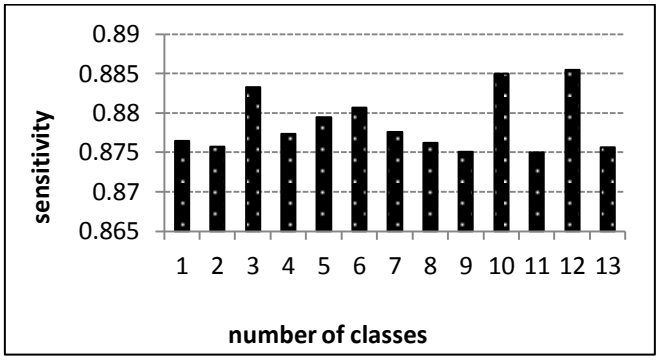
با بدست آوردن مقادیر بهینه برای پارامترهای روش جنگل تصادفی، ماتریس پیش بینی بدست آمده و دقت پیش بینی هر رده تحت عنوان "حساسیت"، مشخص می‌شود. درجدول (۵-۱) ماتریس پیش بینی مدل جنگل تصادفی با m=۲۶ و $ntree$=۵۰۰ آمده است.

جدول (۵-۱)- ماتریس پیش بینی مدل جنگل تصادفی با m=۲۶ و $ntree$=۵۰۰ برای داده آموزشی

رده‌پیش‌بینی / رده واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۵۱۲	۱	۰	۲	۰	۸	۰	۳	۱	۰	۰	۰	۰	۰/۹۷۲
۲	۰	۱۵۵	۱	۱	۰	۰	۱۲	۰	۰	۲	۰	۲	۰	۰/۸۹۶
۳	۰	۰	۱۶۷	۸	۱	۰	۰	۳	۲	۰	۰	۰	۰	۰/۹۲۳
۴	۴	۱	۱۰	۱۴۰	۱۴	۱۱	۰	۴	۰	۰	۰	۰	۰	۰/۷۶۱
۵	۰	۰	۲	۲۹	۷۶	۷	۰	۲	۱	۰	۰	۲	۰	۰/۶۳۹
۶	۲۸	۲	۳	۲۰	۴	۱۰۰	۵	۰	۰	۰	۰	۰	۰	۰/۶۱۸
۷	۰	۸	۰	۱	۰	۱	۶۰	۰	۰	۰	۰	۰	۰	۰/۸۵۷
۸	۱	۵	۱	۲	۰	۰	۰	۲۷۱	۱۲	۰	۰	۳	۰	۰/۹۱۹
۹	۱	۰	۰	۰	۰	۰	۰	۱۱	۳۵۲	۰	۰	۰	۰	۰/۹۶۸
۱۰	۰	۰	۱	۰	۰	۰	۰	۳	۸	۲۵۳	۲	۲	۰	۰/۹۴۱
۱۱	۰	۰	۰	۰	۰	۰	۰	۱	۰	۱	۲۸۲	۱	۰	۰/۹۹
۱۲	۰	۰	۰	۰	۰	۰	۰	۹	۰	۰	۰	۲۹۴	۰	۰/۹۷۱
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۶۴۴	۰/۹۹۹
دقت کلی														۰/۹۲۴

یادآوری می‌شود که درایه‌های قطر اصلی نشان‌دهنده تعدادی از داده‌هاست که رده آن‌ها به‌درستی تشخیص داده شده و درایه‌های پائین و بالای قطر اصلی تعداد داده‌هایی هستند که رده آن‌ها به اشتباه تشخیص داده شده است. از آنجا که روش جنگل تصادفی در اجراهای متعدد، خروجی‌های متفاوتی را نتیجه می‌دهد، لذا مقادیر حساسیت، توام با عدم قطعیت می‌باشند. بدین منظور کافیست با پارامتر های m=۲۶ و $ntree$=۵۰۰ روش جنگل تصادفی را مثلاً به تعداد ۵۰ مرتبه اجرا کرده و سپس میانگین

مقادیر حساسیت هر رده را مورد نظر قرار دهیم. مقادیر این میانگین‌ها در شکل (۵-۲) آمده است. همانطور که ملاحظه می‌شود رده‌های ۳ ، ۱۰ و۱۲ حساسیت نسبتاً بالایی دارند. به عبارت دیگر از بین رده‌ها این ۳ رده بیشتر درست تشخیص داده شده‌اند. همچنین رده‌های۲، ۹ و ۱۱ نسبت به بقیه رده‌ها دارای حساسیت کمتری هستند. البته این نکته قابل ذکر است که مقادیر حساسیت ۱۳ رده بسیار نزدیک به هم می‌باشند.



شکل (۵-۲)- نمودار میله‌ای میانگین حساسیت رده‌ها براساس داده‌های آموزشی با استفاده از روش RF با ۵۰ بار تکرار

با در نظر گرفتن مقادیر بهینه m=۲۶ و $ntree$=۵۰۰ برای داده‌های آموزشی می‌توان ماتریس پیش‌بینی را برای داده‌های آزمون بدست آورد. نتایج در جدول (۵-۳) ارائه شده است. همانطور که در شکل (۵-۳) مشاهده می‌شود رده‌های ۳ ، ۱۰ و۱۲ حساسیت نسبتاً بالایی دارند که در مقایسه با داده‌های آموزشی شکل (۵-۲) تقریباً نتایج یکسانی بدست آمده است. در جدول (۵-۲) معیار دقت کلی برای داده‌های آموزشی و آزمون ارائه شده است.

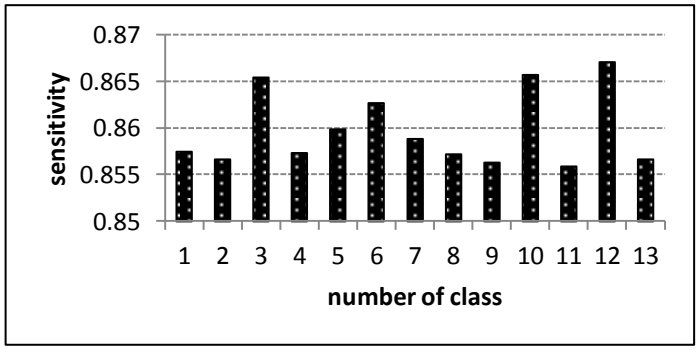
جدول (۵-۲)-مقداردقت کلی برای دو گروه داده آموزشی و آزمون

داده آزمون	داده آموزشی	
۰/۹۱۳	۰/۹۲۴	دقت کلی

جدول (۵-۳)- ماتریس پیش بینی مدل جنگل تصادفی براساس داده‌های آزمون با دو پارامتر m=۲۶ و $ntree$=۵۰۰

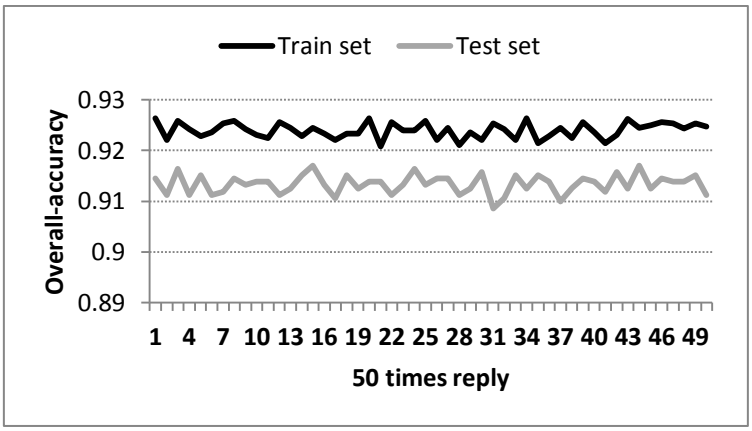
رده پیش‌بینی / رده واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۲۲۴	۲	۰	۰	۰	۶	۰	۱	۱	۰	۰	۰	۰	۰/۹۵۷
۲	۰	۵۹	۰	۱	۰	۰	۵	۱	۱	۰	۰	۱	۰	۰/۸۶۷
۳	۰	۰	۶۷	۵	۰	۰	۰	۱	۲	۰	۰	۰	۰	۰/۹۰۶
۴	۲	۰	۵	۵۲	۲	۳	۰	۱	۰	۲	۰	۰	۰	۰/۷۴۶
۵	۲	۱	۰	۱۰	۲۴	۵	۰	۰	۰	۰	۰	۰	۰	۰/۶۱۹
۶	۱۲	۱	۱	۷	۱	۳۹	۴	۱	۱	۰	۰	۰	۰	۰/۵۵۲۲
۷	۰	۲	۰	۱	۰	۱	۳۱	۰	۰	۰	۰	۰	۰	۰/۹۱۴
۸	۰	۳	۱	۰	۰	۰	۰	۱۲۲	۸	۱	۰	۰	۱	۰/۸۹۷
۹	۰	۰	۱	۰	۰	۰	۰	۴	۱۵۰	۰	۱	۰	۰	۰/۹۶۷
۱۰	۰	۰	۰	۰	۰	۰	۰	۳	۴	۹۷	۳	۱	۰	۰/۹۰۹۹
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۱۳۱	۲	۰	۰/۹۷۷
۱۲	۰	۱	۰	۰	۰	۰	۰	۲	۱	۲	۰	۱۵۲	۱	۰/۹۶۲۵
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۲۶۲	۰/۹۹۲۶
دقت کلی														۰/۹۱۳

مقادیر حساسیت جدول(۵-۳)، جهت مقایسه با یکدیگر در شکل (۵-۳) توسط نمودار میله‌ای نشان داده شده است.



شکل(۵-۳)- نمودار میله‌ای میانگین حساسیت برای رده‌ها براساس داده‌های آزمون با استفاده از روش RF با ۵۰ بار تکرار

با توجه به مقادیر بهینه‌ای که برای تعداد متغیر و تعداد درخت از قبل بدست آمد(m=۲۶ و $ntree$=۵۰۰) این مدل ۵۰ بار اجرا شده و در هر بار اجرا دقت کلی محاسبه می‌شود. نتیجه آن بصورت شکل(۵-۴) می‌باشد.



شکل(۵-۴)- مقدار دقت کلی مدل RF در ۵۰ بار تکرار با $ntree$=۵۰۰ و m=۲۶

همانطور که در شکل (۵-۴) دیده می‌شود، مقادیر دقت کلی داده‌های آموزشی بیشتر از داده‌های آزمون است و این مقادیر برای هر کدام از تکرارها بسیار به هم نزدیک می‌باشند. در واقع برای داده آموزشی تلورانسی حول مقدار ۰/۹۲ دارند و اختلاف آن‌ها بسیار کم و در برخی از مراحل تکرار، برابر می‌باشند.

۵-۲- نتایج حاصل از اجرای روش ماشین بردار پشتیبان

همانطور که در بخش ۳-۲ ذکر شد، در روش ماشین بردار پشتیبان می‌توان از ۴ تابع کرنل خطی، نرمال، چندجمله‌ای و سیگموئید استفاده نمود. هر کدام از این توابع دارای پارمترهایی هستند که با بدست آوردن مقادیر بهینه آن‌ها می‌توان کرنل مناسب را انتخاب نمود. در جدول (۵-۴)، مجدداً مشخصات توابع کرنل و پارامترهای مربوطه آن‌ها یادآوری می‌شود.

جدول (۵-۴)-توابع کرنل و پارامترهای مربوطه

توابع کرنل		پارامترها			
خطی	$x'.x$	*Cost*			
نرمال	$exp(-\gamma.(x'.x))^2$	*gamma*	*Cost*		
چندجمله ای	$(\gamma.x'.x+coef)^{degree}$	*gamma*	*Degree*	*cost*	*Coef*
سیگموئید	$Tanh(\gamma.x'.x+coef)$	*Gamma*	*Cost*	*Coef*	

در هرکرنل ابتدا بهترین مقادیر برای پارامترها بر اساس داده‌های آموزشی بدست می‌آید و پس از آن، مدل بدست آمده توسط داده‌های آزمون مورد ارزیابی قرار می‌گیرد. مقادیر بهینه پارامترها برای هر تابع کرنل، با اعمال فرآیند بهینه سازی بخش(۴-۳)، در جدول (۵-۵) آمده است. مقدار *Misclassification* بر مبنای خطای حاصل از رده‌بندی نادرست با مقادیر بهینه می‌باشد.

جدول (۵-۵)- مقادیر بهینه توابع کرنل در روش ماشین بردار پشتیبان مبتنی بر داده‌های آموزشی

توابع کرنل		*Gamma*	*Cost*	*Degree*	*Coef*	*Misclass*	*SV*
خطی	$x'.x$	۰/۰۰۵۶	۱	-	-	۰/۰۹۵۵	۷۹۶
نرمال	$exp(-\gamma.(x'.x))^2$	۰/۰۰۱	۱۰۰۰	-	-	۰/۰۴۱۹	۷۸۰
چندجمله ای	$(\gamma.x'.x+coef)^{degree}$	۰/۰۰۰۱	۱۰۰۰۰	۲	۲	۰/۰۳۹۷	۷۲۲
سیگموئید	$Tanh(\gamma.x'.x+coef)$	۰/۰۱	۱۰۰۰	-	۲	۰/۰۴۱۰	۲۷۴۶

با جایگذاری مقادیر بهینه جدول(۵-۵) درروابط توابع کرنل (بخش ۳-۲) و ساخت تابع ممیزی نتایج رده‌بندی به تفکیک نوع کرنل در جداول (۵-۶) تا (۵-۹) آمده است. مجدداً لازم به یادآوری است که معیار حساسیت، میزان درستی رده‌بندی را در هر رده تعیین می‌کند. به عنوان مثال در جدول (۵-۶) حساسیت رده به صورت زیر محاسبه شده است.

$$\frac{۵۱۹}{۵۱۹ + ۲۲} = ۰/۹۵۵$$

جدول (۵-۶)- ماتریس پیش بینی داده‌های آموزشی با تابع کرنل خطی

رده‌پیش‌بینی / رده‌واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۵۱۹	۰	۰	۰	۰	۲۲	۰	۰	۰	۰	۰	۰	۰	۰/۹۵۹
۲	۰	۱۷۰	۰	۰	۰	۰	۲	۰	۰	۰	۰	۰	۰	۰/۹۸۸
۳	۰	۰	۱۷۱	۳	۳	۰	۰	۳	۰	۰	۰	۰	۰	۰/۹۵
۴	۰	۰	۶	۱۷۶	۱۷	۱	۰	۲	۰	۰	۰	۰	۰	۰/۹۸۹۷
۵	۰	۰	۱	۳	۹۹	۰	۰	۰	۰	۰	۰	۰	۰	۰/۹۶۱
۶	۸	۰	۰	۱	۰	۱۳۹	۰	۰	۰	۰	۰	۰	۰	۰/۹۳۹
۷	۰	۳	۰	۰	۰	۰	۶۸	۰	۰	۰	۰	۰	۰	۰/۹۵۷
۸	۰	۰	۱	۱	۰	۰	۰	۲۸۶	۱	۰	۰	۰	۰	۰/۹۸۹
۹	۰	۰	۲	۰	۰	۰	۰	۴	۳۶۳	۰	۰	۰	۰	۰/۹۸۳
۱۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۶۹	۰	۰	۰	۱
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۸۴	۰	۰	۱
۱۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۳۰۳	۰	۰/۹۹
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۶۴۵	۱
دقت کلی														۰/۹۷۶

جدول (۵-۷)-ماتریس پیش بینی داده‌های آموزشی با تابع کرنل نرمال

رده‌پیش‌بینی / رده‌واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۵۲۱	۰	۰	۰	۰	۱۳	۰	۰	۰	۰	۰	۰	۰	۰/۹۷۵
۲	۰	۱۷۱	۰	۰	۰	۰	۱	۰	۰	۰	۰	۰	۰	۰/۹۹۴
۳	۰	۰	۱۷۶	۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰/۹۸۸
۴	۰	۰	۲	۱۷۸	۷	۱	۰	۰	۰	۰	۰	۰	۰	۰/۹۴۶
۵	۰	۰	۰	۴	۱۱۲	۰	۰	۰	۰	۰	۰	۰	۰	۰/۹۶۵
۶	۶	۰	۰	۰	۰	۱۴۸	۰	۰	۰	۰	۰	۰	۰	۰/۹۶۱
۷	۰	۲	۰	۰	۰	۰	۶۹	۰	۰	۰	۰	۰	۰	۰/۹۷۲
۸	۰	۰	۱	۰	۰	۰	۰	۲۹۳	۰	۰	۰	۰	۰	۰/۹۹۶
۹	۰	۰	۲	۰	۰	۰	۰	۲	۳۶۴	۰	۰	۰	۰	۰/۹۸۹
۱۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۶۹	۰	۰	۰	۱
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۸۵	۰	۰	۱
۱۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۳۰۳	۰	۱
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۶۴۵	۱
دقت کلی														۰/۹۸۷

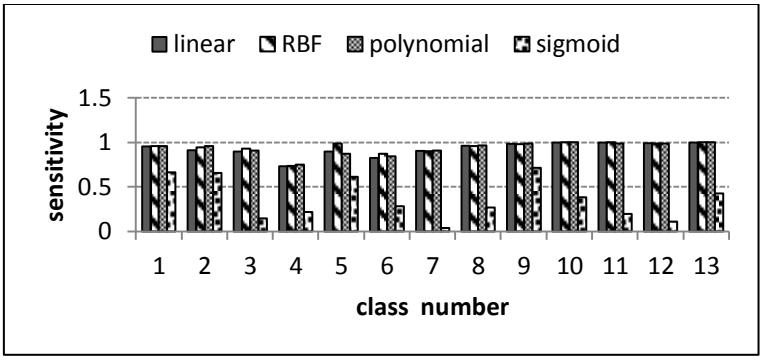
جدول(۵-۸)-ماتریس پیش بینی برای داده‌های آموزشی با تابع کرنل چند جمله‌ای

رده‌پیش‌بینی / رده واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۵۲۱	۰	۰	۰	۰	۱۳	۰	۰	۰	۰	۰	۰	۰	۰/۹۷۵
۲	۰	۱۷۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱
۳	۰	۰	۱۸۰	۲	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰/۹۸۳
۴	۰	۰	۰	۱۸۰	۸	۱	۰	۰	۰	۰	۰	۰	۰	۰/۹۵۲
۵	۰	۰	۰	۲	۱۱۰	۰	۰	۰	۰	۰	۰	۰	۰	۰/۹۸۲
۶	۶	۰	۰	۰	۰	۱۴۸	۰	۰	۰	۰	۰	۰	۰	۰/۹۶۱
۷	۰	۱	۰	۰	۰	۰	۷۰	۰	۰	۰	۰	۰	۰	۰/۹۸۵
۸	۰	۰	۰	۰	۰	۰	۰	۲۹۲	۰	۰	۰	۰	۰	۱
۹	۰	۰	۱	۰	۰	۰	۰	۳	۳۶۴	۰	۰	۰	۰	۰/۹۸۹
۱۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۶۹	۰	۰	۰	۱
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۸۵	۰	۰	۱
۱۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۳۰۳	۰	۱
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۶۴۵	۱
دقت کلی														۰/۹۸۹

جدول(۵-۹)-ماتریس پیش بینی داده‌های آموزشی با تابع کرنل سیگموئید

رده پیش‌بینی / رده واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۳۳۰	۱۵	۴	۳۹	۱۳	۸۰	۱۱	۰	۰	۰	۲	۰	۰	۰/۶۶۸
۲	۱۴	۱۲۵	۰	۳	۳	۵	۲۸	۸	۱	۱	۰	۰	۰	۰/۶۶۴
۳	۱	۱	۵۳	۰	۴	۴	۱	۱	۸	۴	۱۹۴	۰	۰	۰/۱۹۵
۴	۲	۷	۱۱۳	۷۹	۳۹	۲۲	۲۶	۴	۰	۰	۰	۰	۰	۰/۲۷۰
۵	۰	۰	۰	۲۳	۲۹	۱	۰	۰	۰	۰	۰	۰	۰	۰/۵۴۷
۶	۲۰	۰	۰	۲۳	۴	۳۲	۱	۰	۰	۰	۰	۰	۰	۰/۴
۷	۳۸	۸	۰	۸	۲۰	۷	۳	۱	۰	۰	۰	۰	۰	۰/۰۳
۸	۰	۱	۱	۰	۰	۰	۰	۷۶	۷۴	۲۳	۰	۴۷	۱۵۲	۰/۲۰۳
۹	۰	۱	۲	۰	۰	۰	۰	۴۵	۲۲۲	۱۳	۴۳	۴	۰	۰/۶۷۲
۱۰	۴	۰	۱	۲	۴	۳	۰	۲۹	۱۱	۴۵	۱	۳	۰	۰/۴۳۶
۱۱	۱۱۷	۵	۷	۵	۲	۸	۰	۷	۴۰	۱۵	۴۴	۵	۰	۰/۱۷۲
۱۲	۰	۱	۰	۰	۰	۰	۰	۲۸	۸	۹	۰	۱۹	۱۳۲	۰/۰۹۶
۱۳	۱	۹	۰	۲	۰	۰	۰	۹۶	۰	۱۵۹	۱	۲۲۵	۳۶۱	۰/۴۲۲
دقت کلی														۰/۳۹۶

پس از محاسبه ماتریس‌های پیش بینی، برای انتخاب تابع کرنل در روش *SVM* کافی است دو معیار حساسیت و دقت کلی برای چهار تابع کرنل مورد نظر مقایسه شود. بدین منظور ابتدا نمودار حساسیت ۱۳ رده برای چهار تابع کرنل به تفکیک هر رده رسم می‌شود (شکل ۵-۵). از آنجا که مقدار حساسیت، درستی رده‌بندی مشاهدات را نشان می‌دهد و در شکل (۵-۵) تابع کرنل نرمال تقریباً برای تمامی ۱۳ رده مقدار حساسیت بالایی دارد لذا این تابع با مقادیر بهینه پارامترهایش بعنوان انتخاب نهایی مطرح است.



شکل (۵-۵)-نمودار میله‌ای حساسیت رده‌ها با استفاده از روش SVM برای داده‌های آموزشی

در جدول (۵-۱۰) مقادیر دقت کلی برای هر ۴ تابع کرنل آمده است.

جدول (۵-۱۰)- دقت کلی برای توابع کرنل برای داده آموزشی

	کرنل خطی	کرنل نرمال	کرنل چندجمله‌ای	کرنل سیگموئید
دقت کلی	۰/۹۷۶	۰/۹۸۷	۰/۹۸۹	۰/۳۹۶

برای ارزیابی مدل انتخابی بر اساس مقادیر بهینه پارامترها در جدول (۵-۵) از داده‌های آزمون استفاده می‌شود. نتایج آزمون روش *SVM* تحت ۴ کرنل مختلف در جداول (۵-۱۱) تا (۵-۱۴) ارائه شده است.

جدول (۵-۱۱)- ماتریس پیش‌بینی داده‌های آزمون با تابع کرنل خطی

رده پیش‌بینی / رده واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۲۲۷	۰	۰	۲	۱	۹	۰	۰	۰	۰	۰	۰	۰	۰/۹۵۳
۲	۰	۶۵	۰	۰	۰	۰	۶	۰	۰	۰	۰	۰	۰	۰/۹۱۵
۳	۰	۰	۶۹	۳	۰	۳	۰	۱	۱	۰	۰	۰	۰	۰/۸۹۶
۴	۰	۰	۴	۵۴	۱۵	۱	۰	۰	۰	۰	۰	۰	۰	۰/۷۲۹
۵	۰	۰	۰	۲	۲۶	۰	۰	۰	۰	۰	۰	۱	۰	۰/۸۹۶
۶	۷	۰	۰	۳	۰	۵۳	۱	۰	۰	۰	۰	۰	۰	۰/۸۲۸
۷	۰	۳	۰	۰	۰	۰	۲۸	۰	۰	۰	۰	۰	۰	۰/۹۰۳
۸	۰	۰	۰	۳	۰	۱	۰	۱۳۴	۱	۰	۰	۰	۰	۰/۹۶۴
۹	۰	۰	۲	۰	۰	۰	۰	۱	۱۵۴	۰	۰	۰	۰	۰/۹۸۰
۱۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۰۸	۰	۰	۰	۱
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۳۳	۰	۰	۱
۱۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۱۵۸	۰	۰/۹۹
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۶۳	۱
دقت کلی														۰/۹۵۳

جدول (۵-۱۲)- ماتریس پیش بینی داده های آزمون با تابع کرنل نرمال

رده پیش‌بینی / رده‌واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۲۲۸	۰	۰	۱	۰	۷	۰	۲	۰	۰	۰	۰	۰	۰/۹۵۷
۲	۰	۶۵	۰	۰	۰	۰	۶	۰	۰	۰	۰	۰	۰	۰/۹۴۲
۳	۰	۰	۷۰	۳	۰	۱	۰	۱	۱	۰	۰	۰	۰	۰/۹۳۲
۴	۰	۰	۳	۵۷	۱۳	۲	۰	۰	۰	۰	۰	۱	۰	۰/۷۳۵
۵	۰	۰	۰	۲	۲۹	۱	۰	۰	۰	۰	۰	۰	۰	۰/۹۹
۶	۶	۰	۰	۱	۰	۵۵	۱	۰	۰	۰	۰	۰	۰	۰۸۷۳
۷	۰	۳	۰	۰	۰	۰	۲۸	۰	۰	۰	۰	۰	۰	۰/۹۰۳
۸	۰	۰	۰	۳	۰	۱	۰	۱۳۲	۲	۰	۰	۰	۰	۰/۹۵۶
۹	۰	۰	۲	۰	۰	۰	۰	۱	۱۵۳	۰	۰	۰	۰	۰/۹۸
۱۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۰۸	۰	۰	۰	۱
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۳۳	۰	۰	۱
۱۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۱۵۸	۰	۰/۹۹
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۶۳	۱
دقت کلی														۰/۹۵۸

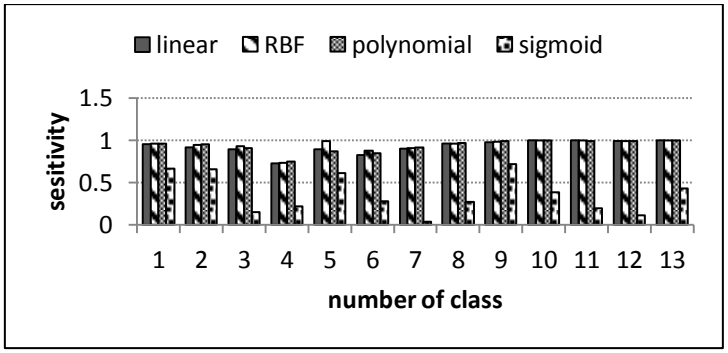
جدول (۵-۱۳)- ماتریس پیش بینی داده‌های آزمون با تابع کرنل چند جمله‌ای

رده پیش‌بینی / رده واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۲۲۷	۰	۰	۱	۰	۹	۰	۰	۰	۰	۰	۰	۰	۰/۹۵۷
۲	۰	۶۵	۰	۰	۰	۰	۳	۰	۰	۰	۰	۰	۰	۰/۹۵۵
۳	۰	۰	۷۰	۱	۱	۲	۰	۲	۱	۰	۰	۰	۰	۰/۹۰۹
۴	۰	۰	۳	۵۷	۱۴	۱	۰	۰	۰	۰	۰	۱	۰	۰/۷۵
۵	۰	۰	۰	۳	۲۷	۰	۰	۰	۰	۰	۰	۱	۰	۰/۸۷۰
۶	۷	۰	۰	۲	۰	۵۴	۱	۰	۰	۰	۰	۰	۰	۰/۸۴۳
۷	۰	۳	۰	۰	۰	۰	۳۱	۰	۰	۰	۰	۰	۰	۰/۹۱۱
۸	۰	۰	۱	۳	۰	۱	۰	۱۳۳	۰	۰	۰	۰	۰	۰/۹۶۴
۹	۰	۰	۱	۰	۰	۰	۰	۱	۱۵۵	۰	۰	۰	۰	۰/۹۸۷
۱۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۰۸	۰	۰	۰	۱
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱۳۳	۱	۰	۰/۹۹
۱۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۱۵۶	۰	۰/۹۹
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۲۶۳	۱
دقت کلی														۰/۹۵۷

جدول (۵-۱۴)- ماتریس پیش بینی داده‌های آزمون با تابع کرنل سیگموئید

رده پیش‌بینی / رده واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۱۴۱	۶	۱	۱۲	۹	۳۳	۸	۰	۰	۰	۲	۱	۰	۰/۶۶۲
۲	۳	۴۸	۰	۱	۰	۲	۸	۶	۳	۱	۰	۰	۰	۰/۶۶
۳	۲	۰	۱۹	۱	۲	۵	۱	۰	۲	۰	۹۸	۰	۰	۰/۱۴۶
۴	۰	۳	۵۲	۲۸	۱۶	۱۱	۱۷	۱	۰	۰	۰	۰	۰	۰/۲۱۸
۵	۰	۰	۰	۷	۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۰/۶۱۱
۶	۱۱	۰	۰	۱۱	۱	۹	۰	۰	۰	۰	۰	۰	۰	۰/۲۸۱
۷	۱۹	۲	۰	۰	۳	۲	۱	۰	۰	۰	۰	۰	۰	۰/۰۳۷
۸	۰	۰	۰	۲	۰	۰	۰	۴۴	۳۷	۸	۰	۲۹	۴۴	۰/۲۶۸
۹	۰	۰	۱	۰	۰	۰	۰	۱۴	۸۶	۷	۱۰	۲	۰	۰/۷۱۶
۱۰	۰	۱	۰	۲	۰	۱	۰	۱۷	۷	۱۹	۰	۲	۰	۰/۳۸۷
۱۱	۵۷	۲	۲	۲	۰	۴	۰	۱	۲۰	۵	۲۳	۰	۰	۰/۱۹۸
۱۲	۰	۰	۰	۱	۰	۰	۰	۱۱	۱	۵	۰	۸	۴۸	۰/۱۰۸
۱۳	۱	۶	۰	۰	۰	۰	۰	۴۲	۰	۶۳	۱	۱۱۷	۱۷۱	۰/۴۲۶
دقت کلی														۰/۳۹۳

با توجه به جداول (۵-۱۱) تا (۵-۱۴) به راحتی می‌توان قضاوت نمود که کرنل نرمال(*RBF*) نسبت به دیگر توابع، دقت کلی بالاتری دارد. نمودار میله‌ای حساسیت رده‌ها برای چهار تابع کرنل در شکل (۵-۶) نشان داده شده است. مشاهده می‌شود، که تابع سیگموئید نسبت به ۳ تابع دیگر حساسیت رده‌ها را بخوبی نشان نداده است. کرنل نرمال (*RBF*) در مجموع از سایر کرنل‌ها بهتر می‌باشد.



شکل (۵-۶)- نمودار میله‌ای حساسیت رده‌ها برای روش *SVM* با استفاده از داده‌های آزمون

جدول (۵-۱۵)، خلاصه نتایج اعمال روش SVM را به ازای ۴ تابع کرنل را نشان می‌دهد.

جدول (۵-۱۵)- دقت کلی روش SVM برای داده‌های آموزشی و آزمون

توابع کرنل	دقت کلی	
	داده آموزشی	داده آزمون
خطی	۰/۹۷۶	۰/۹۵۳
نرمال	۰/۹۸۷	۰/۹۵۸
چندجمله ای	۰/۹۸۹	۰/۹۵۷
سیگموئید	۰/۳۹۶	۰/۳۹۳

۵-۳- نتایج حاصل از اجرای روش تحلیل ممیزی درجه دوم

در مسئله رده‌بندی تحلیل ممیزی درجه دوم، معکوس ماتریس کواریانس مورد نیاز می‌باشد. بدین ترتیب معکوس ماتریس کواریانس موجود نمی‌باشد. بنابراین این روش برای داده‌های ابرطیفی موجود نمی‌باشد.

۵-۴- نتایج حاصل از اجرای روش تحلیل ممیزی خطی

ماتریس پیش بینی نتایج حاصل از این روش با استفاده از داده‌های آموزشی و آزمون به ترتیب در جداول (۵-۱۶) و (۵-۱۷) آمده است.

جدول(۵-۱۶)- ماتریس پیش بینی براساس داده آموزشی با استفاده از روش تحلیل ممیزی خطی

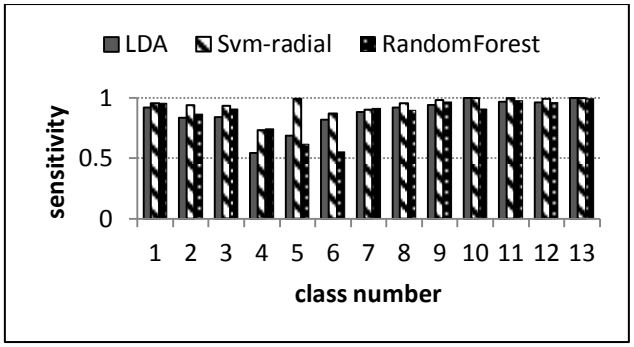
رده‌پیش‌بینی / رده‌واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۴۸۱	۴	۰	۲	۰	۳۴	۰	۶	۰	۰	۰	۰	۰	۰/۹۱۲۷
۲	۰	۱۵۵	۰	۰	۰	۰	۲۰	۰	۰	۰	۰	۰	۰	۰/۸۸۵۷
۳	۰	۰	۱۶۶	۷	۴	۰	۰	۲	۲	۰	۰	۰	۰	۰/۹۱۷۱
۴	۰	۰	۲۶	۱۳۴	۱۱	۱۱	۰	۰	۲	۱	۰	۰	۰	۰/۷۲۴۳
۵	۰	۰	۶	۱۸	۹۴	۰	۰	۰	۱	۰	۰	۰	۰	۰/۷۸۹۹
۶	۱۷	۰	۰	۵	۱	۱۳۳	۴	۲	۰	۰	۰	۰	۰	۰/۸۲۰۹
۷	۰	۴	۰	۰	۰	۱	۶۵	۰	۰	۰	۰	۰	۰	۰/۹۲۸۵
۸	۰	۷	۱	۵	۲	۰	۰	۲۵۵	۲۵	۰	۰	۰	۰	۰/۸۶۴۴
۹	۰	۰	۰	۰	۰	۰	۰	۱۲	۳۵۲	۰	۰	۰	۰	۰/۹۶۷۰
۱۰	۰	۰	۰	۰	۰	۰	۰	۰	۲	۲۸۳	۱	۱	۱	۰/۹۸۲۶
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۶	۰	۲۷۰	۹	۰	۰,۹۴۷۳
۱۲	۰	۰	۰	۰	۱	۰	۰	۵	۳	۰	۳	۳۲۰	۱	۰/۹۶۰۹
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۶۶۱	۱
دقت کلی														۰/۹۲۴

جدول (۵-۱۷)- ماتریس پیش بینی براساس داده آزمون با استفاده از روش تحلیل ممیزی خطی

رده‌پیش‌بینی / رده واقعی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	حساسیت
۱	۲۱۶	۱	۰	۱	۰	۱۳	۰	۳	۰	۰	۰	۰	۰	۰/۹۲۳۷
۲	۰	۵۷	۰	۰	۰	۰	۱۰	۰	۱	۰	۰	۰	۰	۰/۸۳۸۲
۳	۰	۰	۶۳	۵	۵	۰	۰	۱	۱	۰	۰	۰	۰	۰/۸۴
۴	۰	۰	۱۲	۴۳	۸	۱۰	۳	۳	۰	۰	۰	۰	۰	۰/۵۴۴۳
۵	۰	۰	۲	۱۱	۲۹	۰	۰	۰	۰	۰	۰	۰	۰	۰/۶۹۰۴
۶	۱۰	۰	۰	۱	۰	۵۵	۰	۱	۰	۰	۰	۰	۰	۰/۸۲۰۸
۷	۰	۲	۰	۰	۱	۱	۳۱	۰	۰	۰	۰	۰	۰	۰/۸۸۵۷
۸	۰	۱	۱	۰	۰	۱	۰	۱۲۵	۸	۰	۰	۰	۰	۰/۹۱۹۱
۹	۰	۰	۱	۰	۰	۰	۰	۸	۱۴۷	۰	۰	۰	۰	۰/۹۴۲۳
۱۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۱۱۱	۰	۰	۰	۰/۹۹۹
۱۱	۰	۰	۰	۰	۰	۰	۰	۰	۱	۰	۱۳۰	۳	۰	۰/۹۷۰۱
۱۲	۰	۰	۰	۱	۰	۰	۰	۰	۰	۰	۳	۱۶۴	۲	۰/۹۶۴۷
۱۳	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۰	۲۶۵	۰/۹۹۹
دقت کلی														۰/۹۱۹

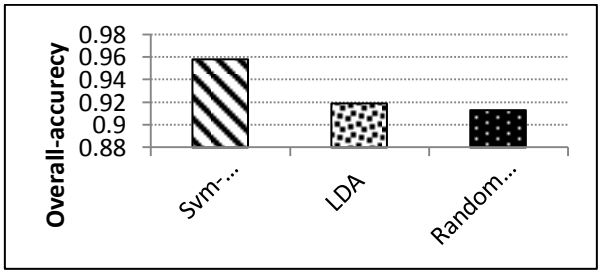
۵-۵-مقایسه روش‌های پیشرفته و متداول آماری

در این قسمت به مقایسه روش متداول و روش های پیشرفته آماری پرداخته می‌شود. شکل(۵-۷) عملکرد سه روش *RF*، *SVM* و *LDA* را در خصوص معیار حساسیت نشان می‌دهد. روش *SVM* با کرنل نرمال در مقایسه با دو روش *RF* و *LDA* حساسیت بالایی برای تمامی ۱۳ رده دارد. یا بعبارتی رده‌بندی را بخوبی انجام می‌دهد.



شکل (۵-۷)- مقایسه حساسیت رده‌ها برای داده‌های آزمون با استفاده از روش‌های جنگل تصادفی(RF)، ماشین بردار پشتیبان(SVM-RBF) و روش تحلیل ممیزی خطی(LDA).

در شکل (۵-۸) دقت کلی این ۳روش مقایسه شده است که در آن روش ماشین بردار پشتیبان با کرنل نرمال بالاترین میزان دقت کلی را دارد.



شکل (۵-۸)-مقایسه دقت کلی روش‌های جنگل تصادفی(RF)، ماشین بردار پشتیبان(SVM-RBF) و روش تحلیل ممیزی خطی(LDA).

فصل ششم

نتیجه گیری و پیشنهادها

۶-۱- بحث و نتیجه گیری

هدف اصلی این مطالعه مقایسه دو روش پیشرفته آماری با دو روش متداول آماری است. ماشین بردار پشتیبان روشی مناسب برای رده‌بندی داده‌ها هستند. در مسئله رده‌بندی SVM عموماً این کار با ماکزیمم کردن حاشیه که در یک چارچوب قانونی می‌باشد را فراهم می‌کند. یکی از مسائل مهم و اساسی در اجرای روش SVM بهینه سازی مقادیر پارامتر کرنل‌ها می‌باشد. بهینه سازی مقادیر پارامتر ها کار بسیار پرهزینه و زمان‌بر می‌باشد. که در این مطالعه بررسی شد.

برای تحلیل داده‌ها در این آزمایش، مقدار دقت کلی تابع کرنل نرمال از بقیه توابع کرنل چندجمله‌ای، خطی و سیگموئید بالاتر است. در مورد معیار حساسیت هم می‌توان گفت تابع کرنل نرمال حساسیت بالایی دارد به این معنی که رده‌بندی را به خوبی انجام می‌دهد. تابع کرنل نرمال دارای دو پارامتر $gamma=۰/۰۰۱$ و $cost=۱۰۰۰$ می‌باشد.

در روش RF یکی از مسائل مهم و اساسی یافتن مقادیر بهینه برای دو پارمتر، تعداد درخت و متغیر می‌باشد. برای هر کدام از این دو پارامتر مقادیر مختلفی در نظر گرفته می‌شود. شکل (۵-۱) نمودار خطای OOB برای این دو پارامتر را نشان می‌دهد و می توان مقدار بهینه برای هر یک از دو پارامتر را بدست آورد. استنباطی که از شکل (۵-۱) می‌توان کرد این است که با زیاد شدن تعداد درخت مقدار خطای OOB به پایداری می‌رسد.

دقت کلی روش‌های RF و LDA در مقایسه با روش SVM کمتر است. البته این تنها ملاک برتری روش SVM نمی‌باشد. زیرا عواملی مانند پر هزینه و زمان‌بر بودن مقادیر بهینه پارامترهای کرنل می‌تواند بعنوان نقصی برای این روش باشد. اما با مقایسه‌هایی که برای همه روش‌ها انجام شد روش SVM با تابع کرنل نرمال برای رده‌بندی داده‌های مسئله انتخاب می‌شود.

۶-۲- پیشنهادها

۱) استفاده از روش‌های بهینه سازی مختلف در تعیین پارامترهای مدل‌های RF و SVM و مقایسه و ارزیابی روش‌های بهینه سازی مختلف

۲) استفاده از روش‌های انتخاب الگو (Feature Selection) ی مختلف در تعیین ورودی‌های مهم برای مدل‌های توسعه داده شده

۳) مقایسه نتایج روشRF از لحاظ انتخاب متغیرهای مهم با روش‌های دیگر انتخاب الگو

۴) استفاده از پایگاه داده‌های مختلف مانند پایگاه داده‌های چند طیفی و ارزیابی عملکرد روش‌های مختلف رده‌بندی بر روی آنها و مقایسه نتایج پایگاه داده‌های ابرطیفی و چند طیفی

۵) استفاده از دیگر روشهای یادگیری ماشین در رده‌بندی پایگاه داده‌های ماهواره‌ای و مقایسه نتایج آنها با روش‌های مطالعه شده در این تحقیق

فهرست منابع

*Abou El-Magd, I. and Tanton, T.W., (2003). "Improvements in land use mapping for irrigated agriculture from satellite sensor data using a multi-stage maximum likelihood classification", **International Journal of Remote Sensing**, 24(21), 4197-4206.*

*Verikas, A. A.Gelzinis, M.Bacauskiene., (2010). "Mining data with random forests", A survey and results of new tests, **Pattern Recognition**.*

Abuelgasim, A., Gopal, S., Irons, J. and Strahler, A., (1996). "An artificial neural network for the classification of ASAS directional measurements". Remote Sens. Environ., 57, 79-87.

*Atkinson, P.M. and Tatnall, A.R.L., (1997). "Neural networks in remote sensing", **International Journal of Remote Sensing**, 18, 699–709.*

*Benediktsson, J., Swain, P. and Ersoy, O. K., (1990). "Neural network approaches versus statistical methods in classification of multisource remote sensing data". **IEEE Trans. Geosci**. Remote Sens, 28, 540-552.*

*Brown, M., Gunn, S.R. and Lewis, H. G., (1999). "Support vector machines for optimal classification and spectral unmixing", **Ecological Modelling**, 120(2-3), 167-179.*

Ben-Hur, Asa., and Jason Weston., (2010). "A user's guide to support vector machine. Methods in Molecular 43 Biology 609, 223-239.

Chapelle, O. and Vapnik, V., (1999). "Model selection for support vector machines". In Advances in neural information processing systems.

*Civco, D.L., (1993). "Artificial neural networks for land-cover classification and mapping", **International Journal of Geographical Information**, 7(2), 173-186.*

Cutler D.R., Edwards, T.C. Jr., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J., (2007). "Random forests for classification in ecology", Ecology, 88, 2783–2792.

*DeFries, R. S., Hansen, M., Townshend, J. R. G. and Sohlberg, R., (1998). "Global land cover classifications at ١km spatial resolution", the use of training data derived from Landsat imagery in decision tree classifiers, **International Journal of Remote Sensing**, 19, 3141–3168.*

*Defries, H. S. and Townsend, J. R. C., (1994). "NDVI-derived and cover classifications at a global scale", **Int. J. Remote Sens**., 15, 3567-3586.*

*Demorate, F., (1998). "Land cover mapping estimated in Rendonia", Brazil, **Int. J. Remote Sensing,** 19(5).*

Davood Mahmoodi, Ali Soleimani, Hossein Khosravi, Mehdi Taghizadeh., (2011). "FPGA Simulation of Linear and Nonlinear Support Vector Machine".

Dr.Wolfgang Hardle., (2004). "Classification and Regression Tree(CART)", Humboldt University, Berlin.

Daskalaki S, Kopanas I, Avouris N., (2006). "Evaluation of classifiers for an uneven class distribution problem". Applied Artificial Intelligence ,20:381–417.

Friedl, M.A. and Brodley, C.E., (1997). "Decision Tree Classification of Land Cover from Remotely Sensed Data". Remote Sensing of Environment, 61, 399-409.

*Ghimire, B., Rogan J. and Miller, J., (2010). "Contextual land-cover classification, incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic". **Remote Sensing Letters**, 1:1, 45-54.*

*Gislason, P. O., Benediktsson, J. A. and Sveinsson, J. R., (2006). "Random forests for land cover classification". **Pattern Recognition Letters**, 27, 294–300.*

*Goodenough, D. and Schlien, S., (1974). "Results of cover-type classification by maximum likelihood and parallelepiped methods". Proceedings of the Second Canadian Symposium on Remote Sensing, 1:136-164. **Canada Centre for Remote Sensing, Ottawa**.*

Gopal, S., and Woodcock, C. E., (1994). "Theory and methods for accuracy assessment of thematic maps using fuzzy sets". Photogramm. Eng. Remote Sens., 60, 181-188.

*Gopal, S., and Woodcock, C. E., (1996). "Remote sensing of forest change using artificial neural networks". **IEEE Trans. Geosci. Remote Sens**., 34, 119-140.*

*Ham, J.S., Chen, Y., Crawford, M.M. and Ghosh, J., (2005). "Investigation of the random forest framework for classification of hyperspectral data". **IEEE Transaction on Geoscience and Remote Sensing**, 43(3), 492–501.*

*Hansen, M., DeFries, R. S., Townshend, J. R. G., and Sohlberg, R., (2000). "Global land cover classification at ١ km spatial resolution using a classification tree approach". **International Journal of Remote Sensing**, 21, 1331–1364.*

*Hansen, M., Dubayah, R., and DeFries, R., (1996). "Classification trees: an alternative to traditional land cover classifiers". **International Journal of Remote Sensing**, 17, 1075–1081.*

*Hermes, L., Frieauff, D., Puzicha, J. and Buhmann, J. M., (1999). "Support vector machines for land usage classification in Landsat TM imagery". **Proc. of the IEEE International Geoscience and Remote Sensing Symposium**, vol. 1 (348– 350), **Hamburg**.*

*Huang, C., Davis, L.S. and Townshend, J.R.G., (2002). "An assessment of support vector machines for land cover classification". **International Journal of Remote Sensing, 23(4), 725-749.***

Huang, C., Song. K., Kim, S., Townshend, J.R.G., Davis, P., Masek, J. G., Goward, S. N., (2008). "Use of a dark object concept and support vector machines to automate forest cover change analysis". Remote Sensing of Environment, 112 (3), 970-985.

*Ham, J., Y., Chen, M. M. Crawford, and J. Ghosh., (2005). "Investigation of the random forest framework for classification of hyperspectral data". **IEEE Trans. Geosci. Remote Sens., vol. 43, no. 3, 492–501.***

*Kavzoglu, T. and Colkesen, I., (2009). "A kernel functions analysis for support vector machines for land cover classification". **International Journal of Applied Earth Observation**, 11(5), 352-359.*

Keuchel, J., Naumann, S., Heiler, M. and Siegmund, A., (2003). "Automatic land cover analysis for Tenerife by supervised classification using remotely sensed data". Remote Sensing of Environment, 86, 530-541.

*Langford, M., (1997). "Land cover mapping in tropical hillsides in environment". a case study in the Caucau region, Colombia, **Int. J. Remote Sensing**, 18(6).*

Lucas, R., Rowlands, A., Brown, A., Keyworth, S. and Bunting, P., (2007). "Rule-based classification of multi-temporal satellite imagery for habitat and agricultural land

*cover mapping". **ISPRS Journal of Photogrammetry & Remote Sensing**, 62, 165-185.*

Lutz Hame., (2009). "Knowledge Discovery With Support Vector Machines". Wiley.

Breiman, L., (1999). "Random forests features". Technical Report 567, Department of Statistics. University of California, Berkeley.

Macloed, R.S. and Congalton, R.G., (1998). "A quantitative comparison of change detection algorithms for monitoring grass from remotely sensed data". Photogrammetric and Engineering Remote Sensing, 64(3), 207-216.

Mercier, G. and Lennon, M., (2003). "Support vector machines for hyperspectral image classification with spectral-based kernels". Geoscience and Remote Sensing Symposium, 1, 288-290.

Meyer, W.B. and Turner B.L., (1994). "Change in land use and land cover". a global perspective, Cambridge University Press, Cambridge.

*Ming, Z., Goosens, R. and Daels, L., (1993). "Application of satellite remote sensing to soil and land use mapping in the Rolling Hilly Areas of Nanjing". Eastern China, **EARSeL Advances in Remote Sensing**, 2(3), 34-44.*

*Murthy, C.S., Raju, P.V. and Badrinath, K.V.S., (2003). "Classification of wheat crop with multi-temporal images". performance of maximum likelihood and artificial neural networks, **International Journal of Remote Sensing**, 24(23), 4871-4890.*

Nijkamp, P., (1997). "Environmental security and sustainability in national resource management". a decision support framework, Research memorandum, Dep. of Economics, Free University, Amsterdam.

Ostrom, E., (1990). "Governing the commons". Cambridge University Press, Cambridge.

*Pal, M. and Mather, M., (2003). "An assessment of the effectiveness of decision tree methods for land cover classification". **Remote Sensing of Environment**, 86, 554-565.*

Pal, M. and Mather, M., (2004). "Assessment of the effectiveness of support vector machines for hyperspectral data". Future Generation Computer Systems, 20(7), 1215-1225.

Pai, P.-F., W.-C. Hong., (2005). "Support vector machines with simulated annealing algorithms in electricity load forecasting". Energy Conversion Manage.

Pai, P.-F., W.-C. Hong., (2005). "Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms". Electric Power Syst. Res.

*Pal, M., (2005). "Random forest classifier for remote sensing classification". **International Journal of Remote Sensing**, 26, 217–222.*

*Paola, J.D. and Schowengerdt, R.A., (1995). "A review and analysis of backpropagation neural networks for classification of remotely sensed multi-spectral imagery". **International Journal of Remote Sensing, 16, 3033–3058.***

Parry, M.L., (1990). "Climate change and world agriculture". EarthSacan, London.

*Peters, J., Baets, B. D., Verhoesta, N. E.C., Samson, R., Degroeve, S., Becker, P. D. and Huybrechts, W., (2007). "Random forests as a tool for ecohydrological distribution modelin". **Ecological Modeling, 207, 304-318.***

*Provost F, Fawcett T., (1997). "Analysis and visualization of classifier performance. Comparison under imprecise class and cost distributions". In: **Proc. 3rd Intl. Conf. on Knowledge Discovery and Data Mining, pp. 43–48.***

Platt, J., (1999). "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods". In: Smola, A.J., Bartlett, P., Sch¨olkopf, B., Schuurmans, D. (eds.): Advances in Large Margin Classifiers. MIT Press, 61–74.

*Platt, J., Cristanini, N., Shawe-Taylor, J., (2000). "Large margin DAGs for multiclass classification". Advances in Neural Information Processing Systems 12. **MIT Press** 543–557.*

*Reeves, R. G., Anson, A. and Landen, D., (1975). "Manual of Remote Sensing". Amer. Soc. of Photogramm., Falls Church, VA, 2 **vols.**, 2144..*

*Robin Genuer, Jean-Michel Poggi, ChristineTuleau-Malot., (2010). "Variable selection using Random Forests". **Pattern Recognition Letters**.*

Schell, J. A., (1973). "In Remote Sensing of Earth Resources". Volume I (F. Shahrokhi, ed.), Univ. of Tennessee Space Inst., Tullahoma, TN, 374-394.

*Scott, G. B. and Mark, R. G., (2001). "Classification of land cover types for the Fort Bening eco-region using enhanced thematic mapper data". **Strategic** Environmental Research and Development program (SERDP), ERDC/ET TN-ECMI-01-01. 9.*

*Shalaby, A. and Tateishi, R., (2007). "Remote Sensing and GIS for mapping and monitoring land cover and land use changes in the Northwestern coastal zone of Egypt". **Applied Geography**, 27, 28-41.*

Shih-Wei Lin, Zne-Jung Lee, Shih-Chieh Chen, Tsung-Yuan Tseng., (2008). "Parameter determination of support vector machine and feature selection using simulated annealing approach".

Strahler, A. H., (1980). "The use of prior probabilities in maximum likelihood classification of remotely sensed data". Remote Sensing of Environment, 10, 135-163.

*Sunar Erbek, F., Ozkan, C. and Taberner, M., (2004). "Comparison of maximum likelihood classification method with supervised artificial neural network algorithms for land use activities". **International Journal of Remote Sensing,** 25(9), 1733-1748.*

*Steve R.Gunn., (1998). "Support Vector Machine for Classification and Regression". **Technical report**.*

*Townsend, J. R. G., Justice, C. O. and Kalb, V. T., (1987). "Characterization and classification of South American land cover types using satellite data". **Int. J. Remote Sens**., 8, 1189-1207.*

*Trevor Hastie, Robert Tibshirani, Jerome Friedman., (2009). "The Elements of Statistical Learning". **۲nd edition, Spring**.*

Walton, J. T., (2008). "Subpixel urban land cover estimation. comparing cubist, random forests and support vector regression". Photogrammetric Engineering & Remote Sensing, 74(10), 1213-1222.

*Wang, F., (1990). "Fuzzy supervised classification of remote sensing images". **IEEE Transactions on Geoscience and Remote Sensing**, 28, 194–201.*

*Waske, B. and Benediktsson, A., (2007). "Fusion of Support Vector Machines for Classification of Multisensor Data". **IEEE Transactions on Geoscience and Remote Sensing, 45(12), 3858-3865.***

*Wharton, S.B., (1989). "Knowledge-based spectral classification of remotely sensed image data". In Theory and Applications of Optical Remote Sensing, **G. Asrar Ed., Wiley, New York**.*

*Jia, X., and J.A. Richards., (1997). "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification". **IEEE Trans. Geosci. Rem. Sens.,37(1), 538-42.***

Yang, Q., Li, X. and Shi, X., (2008). "Cellular automata for simulating land use changes based on support vector machines". Computers & Geosciences, 34(6), 592-602.

*Jin, Y., and B. Sendhoff., (2008). "Pareto-Based Multi-Objective Machine Learning An Overview and Case Studies". **IEEE Transactions on Systems**, Man and Cybernetics: Part C, vol. 38, no. 3, 397-415.*

Abstract:

In the past to create a map class packing different effects on the surface of a field of study that is very time consuming and expensive was to be used. With the advent of satellites and satellite images a it by satellite images and their processing is done . But the problem is that the precision of the extracted result for class classification is dependent on the performance of the image processing method. Common methods Enterprises ranking in the past have been able to accomplish this in an appropriate manner. With the advent of new and advanced statistical and machine learning methods, is hoped that these methods can classified satellite images a better class operations conventional methods do. In this study, the class of land use classification method using advanced statistical and recruiting machines and their performance compared with conventional methods, are. Thus, the radiation Abrtyfy be used during and Collected within a specific geographic are collected. The data included 176 explanatory variable (band satellite) and a response variable that is 13 categories of vegetation in the study area is characterized by geographic coordinates to.Values of the explanatory variable has been removed by the AVIRIS sensor. to extract the data from the satellite imagery LAND SAT was used. The study area includes KFC (Kennedy Flight center) is located in the state of Florida in America. Advanced statistical methods studied in this thesis, including support vector machines and random forests are .Support Vector Machine has 4 different kernel function method is that each of them has its own parameters that. Choice of kernel functions and optimize its parameter values using the grid search done, it has a high computational cost. Another method studied in this thesis is a using random forests. Several sets of random forest method the decision tree for the model parameters and variable number of trees required. Further optimization of these two parameters to do the costly and time can be. These two methods are listed in the Advanced Materials and Methods conventional statistical quadratic discriminant analysis, linear discriminant analysis in the engineering sciences to the maximum likelihood method, the distance be Mahalonobis and called. The final model compared to conventional statistical methods,

advanced statistical is obtained. The comparison of the four models are Gaussian kernel Support Vector Machine is. The model parameters and overall accuracy 0/958, gamma = ٠/001 and cost = 1000 and also has a high sensitivity for all class . In other words, the classification class of for all 13 class will be doing well.

Shahrood University of Technology
Facualty of Mathematical Sciences

Dissertation Submitted in Partial
Fulfillment of the Requirements For The
Degree of Master of Science in Applied mathematics

comparison between machine learning and typical classification methods for the classification of satellite images

Faezeh Safari

Supervisors:

Dr.Davood Shahsavani

Dr.Hamid Taheri Shahr Aeini

Advisor:

Dr.Barat Mojaradi

Feb 2013