
بسم الله الرحمن الرحيم

دانشکده علوم ریاضی

گروه آمار

پایان‌نامه کارشناسی ارشد

تشخیص اثرات متقابل موثر برای داده‌های با بعد بالا

سعید محمدی

استادان راهنما

دکتر داود شاهسونی و دکتر حسین باغیشنی

بهمن ۱۳۹۴

تقدیم به:

پدر و مادرم به خاطر زحمات بی دریغشان

چیدم گلی ز باغ ادب تا بروز عید
در بارگاه میر ادب پرور آورم

حیف است با خسان گل دانش کنی نثار
من گل نثار مردم دانشور آورم

استاد شهریار

هر قدر چالیشدیم یا خچیلیقا آلتشقیا مادیم

فقط گوزیاشی توک بیلدیم اما دانشا مادیم

سپاس‌گزاری

خداوند را شاکر هستم که به من قدرت انجام این پایان‌نامه را عطا کرد.

در ابتدا، از زحمات بی‌دریغ اساتید راهنمای محترم، جناب آقایان دکتر داود شاهسونی و دکتر حسین باغیشنی که از ابتدای راه و در طی انجام تحقیق، با راهنمایی‌های خود مرا در نگارش این اثر یاری نمودند، تشکر می‌کنم. بر خود لازم می‌دانم که از اساتید محترم داور، جناب آقای دکتر محمد آرشی و سرکار خانم دکتر هدی مشایخی، و همچنین از تمام اساتید فرهیخته گروه آمار دانشگاه شاهرود، سرکار خانم دکتر اقبال، جناب آقای دکتر احمد نزاکتی و دکتر محمد رضا ربیعی، تشکر و قدردانی نمایم.

وظیفه خود می‌دانم که از زحمات پدر و مادر بسیار عزیز، دلسوز و فداکارم که پیوسته راهنمای من در سختی‌ها و مشکلات بوده‌اند، قدردانی به عمل آورم. همچنین از برادران و خواهرم که همواره پشتیبان من بوده‌اند، صمیمانه سپاس‌گذاری می‌کنم.

با سپاس بی‌دریغ خدمت دوستان و همکلاسی‌های گران‌مایه‌ام که مرا صمیمانه و مشفقانه یاری داده‌اند.

در پایان از تمامی دانشجویان آمار دانشکده ریاضی و مسئولین محترم آموزش دانشکده، تشکر و قدردانی می‌نمایم.

سعید محمدی

بهمن ۱۳۹۴

تعهدنامه

این‌جانب سعید محمدی دانشجوی کارشناسی ارشد رشته آمار دانشکده علوم ریاضی دانشگاه شاهرود، نویسنده پایان‌نامه با عنوان تشخیص اثرات متقابل موثر برای داده‌های با بعد بالا تحت راهنمایی دکتر داود شاهسونی و دکتر حسین باغیشنی متعهد می‌شوم:

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

سعید محمدی
بهمن ۱۳۹۴

مالکیت نتایج و حق نشر

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

چکیده

فن‌آوری‌های نوین، داده‌های حجیمی را با بعد بالا تولید می‌کنند که مسئله تحلیل آن‌ها موجب ابداع روش‌های آماری جدید شده است. قبل از اعمال این روش‌ها، معمول است که بعد داده‌های اصلی توسط روش‌هایی نظیر تحلیل مولفه‌های اصلی و تحلیل عاملی، کاهش داده شده و متغیرهای موثر شناسایی شوند. برای داده‌های با بعد بالا، روش‌هایی نظیر غربال‌گری مستقل مطمئن (SIS) و جنگل‌های تصادفی، متغیرها را با توجه به اهمیت آن‌ها رتبه‌بندی کرده و اثرات مهم را شناسایی می‌کنند. این روش‌ها هزینه محاسباتی بالایی ندارند اما اثرات متقابل را در نظر نمی‌گیرند، در حالی که تجربه محقق ممکن است حاکی از لزوم وجود این اثرات باشد. در این تحقیق، با ارائه رویکرد ضریب همبستگی تعمیم‌یافته (درستنمایی ماکسیمم)، اثرات اصلی و متقابل موثر را با ارائه یک الگوریتم دومرحله‌ای شناسایی می‌کنیم. یکی از مزایای این رهیافت نسبت به روش SIS آن است که مجموعه اثرات منتخب، می‌تواند عملکرد روش‌های رده‌بندی از جمله k–نزدیک‌ترین همسایگی و مرکز–مبنا را بهبود بخشد.

کلمات کلیدی: غربال‌گری مستقل مطمئن، رده‌بندی، رگرسیون لجستیک، رتبه‌دهی متغیرها، ضریب همبستگی تعمیم‌یافته، اثرات متقابل.

پیش گفتار

امروزه با پیشرفت تکنولوژی و ظهور فن‌آوری‌های نو، انقلابی در تولید و ذخیره‌سازی داده‌ها رخ داده و این امر موجب پیدایش پایگاه‌های داده بزرگ و متنوع شده است که برخی از آن‌ها بعد بالا محسوب می‌شوند. بدیهی است که این بعد بالا از داده‌ها به‌سادگی قابل بهره‌برداری نیست. بنابراین نیاز به روش‌های کاهش بعد است تا بتوان ابعاد این داده‌ها را کاهش داده و از روش‌های معمول آماری در تحلیل آن‌ها استفاده نمود.

یکی از مباحث پرکاربرد در داده‌های با بعد بالا مانند داده‌های پزشکی، پردازش تصویر، ژنوم و ریزآرایه، رده‌بندی است. یکی از عوامل خیلی مهم در دقت رده‌بندی، کاهش ابعاد فضای متغیرهای توضیحی است. کاهش هزینه محاسباتی و دقت رده‌بندی، دو دلیل عمده کاهش بعد است. به‌طور کلی دو روش برای کاهش بعد وجود دارد: انتخاب متغیر و استخراج متغیر. در روش انتخاب متغیر، متغیرهایی که در رده‌بندی مفید هستند، از مجموعه کامل متغیرها انتخاب می‌شوند. استخراج متغیر به معنی استخراج متغیرهای جدید از متغیرهای اصلی است که از طریق انجام برخی تبدیلات روی متغیرهای اصلی صورت می‌گیرد. تحلیل مولفه‌های اصلی نمونه‌ای از روش‌های استخراج متغیر است. هر دو رهیافت انتخاب و استخراج متغیرها در اثربخشی روش‌های رگرسیونی و رده‌بندی و تفسیر بهتر نتایج نقش به‌سزایی را ایفا می‌کنند.

در این تحقیق دو روش انتخاب متغیر، با نام‌های غربال‌گری مستقل مطمئن و روش مبتنی بر ضریب همبستگی تعمیم‌یافته را معرفی می‌کنیم. اولین روش تنها اثرات اصلی را انتخاب می‌کند، در حالی‌که روش دوم انتخاب همزمان اثرات اصلی و متقابل را مد نظر دارد. عملکرد این دو روش را، توسط چند مثال شبیه‌سازی و واقعی، در یک مساله رده‌بندی با متغیر پاسخ دوسطحی، ارزیابی می‌کنیم. با این مقدمه، ساختار پایان‌نامه به‌صورت زیر تنظیم شده است:

- در فصل اول، مساله داده‌های با بعد بالا را تشریح کرده و برخی از روش‌های انتخاب متغیر را معرفی می‌کنیم.
- در فصل دوم، روش‌های مختلف رده‌بندی و ارزیابی آن‌ها بر اساس داده‌های آموزشی و آزمون را مطرح می‌کنیم.
- در فصل سوم، ابتدا روش غربال‌گری مستقل مطمئن را معرفی کرده و سپس به بیان روش پیشنهادی مبتنی بر ضریب همبستگی تعمیم‌یافته، می‌پردازیم. هم‌چنین، با دو مثال شبیه‌سازی عملکرد خوب روش پیشنهادی را نمایش می‌دهیم.
- در فصل چهارم، عملکرد دو روش انتخاب متغیر (کاهش بعد) غربال‌گری مستقل مطمئن و ضریب همبستگی تعمیم‌یافته را، به کمک چند مجموعه داده واقعی، بر دقت رده‌بندی متغیر پاسخ

(دوسطحی) ارزیابی می‌کنیم. در مساله رده‌بندی از دو روش رده‌بندی مرکز-مبنا و k-نزدیک‌ترین همسایگی استفاده خواهیم کرد.

- این پایان‌نامه شامل دو پیوست است. پیوست آ شامل تعاریف فضای برداری و توابع به‌طور یکنواخت کراندار می‌باشد. پیوست ب نیز شامل کدهای نوشته‌شده در محیط نرم‌افزار R برای اجرای تحلیل‌های مجموعه داده تصویر برداری قلبی فصل ۴ است.

فهرست مقاله‌های مستخرج از پایان‌نامه

۱. محمدی، س.، شاهسونی د.، باغیشنی ح. (۱۳۹۴)، ”انتخاب متغیر در داده‌های بزرگ“، نخستین کنفرانس ملی محاسبات نرم، دانشگاه گیلان، گیلان، ایران.

۲. محمدی، س.، شاهسونی د.، باغیشنی ح. (۱۳۹۴)، ”تشخیص اثرات متقابل در کاهش خطای رده‌بندی برای داده‌های با بعد بالا“، سیزدهمین کنفرانس آمار ایران، دانشگاه کرمان، کرمان، ایران.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مروری بر روش‌های انتخاب متغیر

۱.۱ مقدمه

امروزه با پیشرفت تکنولوژی و ظهور فن‌آوری‌های نو، انقلابی در تولید و ذخیره‌سازی داده‌ها رخ داده و این امر موجب پیدایش پایگاه‌های داده بزرگ و متنوع شده است. بدیهی است که این حجم بالا از داده‌ها به سادگی قابل بهره‌برداری نیست. بنابراین نیاز به دانشی است که در بطن داده‌ها قرار دارد. این دانش، داده‌کاوی[۱] نامیده شده که با به‌کارگیری علم آمار، هوش مصنوعی[۲] و یادگیری ماشین[۳] در صدد است تا روابط پنهان در داده‌ها را تا حد ممکن آشکار نماید. یکی از مباحث پرکاربرد در داده‌کاوی، رده‌بندی[۴] است که یک روش گروه‌بندی با نظارت[۵] است و در آن گروه‌ها از قبل معلوم هستند. تفاوت اصلی رده‌بندی با خوشه‌بندی (گروه‌بندی بدون نظارت[۶])، وجود متغیر پاسخ است که مشخص می‌کند هر مشاهده به کدام گروه (رده) تعلق دارد. یافتن رابطه یا قاعده‌ای برای تخصیص مشاهدات جدید به رده‌های موجود، نتیجه عملی است که به آن مدل رده‌بندی می‌گویند. در حال حاضر، رده‌بندی در حیطه گسترده‌ای از زمینه‌های علوم مانند پزشکی، شیمی، اقتصاد و مهندسی، کاربرد دارد. پردازش متن[۷] از اسناد اینترنتی، پردازش تصویر[۸] و تحلیل آرایه‌های بیان ژنی[۹] نیز از جمله زمینه‌های هدف موضوع رده‌بندی هستند. در برخی کاربردها مانند رده‌بندی داده‌های ژنوم، مستندات متنی و داده‌های ریزآرایه، محققان با داده‌هایی مواجه هستند که تعداد متغیرهای توضیحی آن‌ها در مقایسه با مشاهدات موجود بسیار زیاد است (داده‌های با

---

۱ Data mining
۲ Artifitial intelligence
۳ Machine learning
۴ Classification
۵ Superviesed
۶ Unsuperviesed
۷ Text processing
۸ Image processing
۹ Gene expression arrays

بعد بالا[۱]) و این امر تحلیل و تفسیرپذیری این نوع داده‌ها را با چالش جدی روبرو ساخته است. بنابراین برای تفسیرپذیری بهتر و درک عمیق‌تر این داده‌ها، نیاز به روش‌های کاهش بعد[۲] و شناسایی متغیرهای موثر است.

انتخاب ویژگی[۳] که آن را در آمار و یادگیری ماشین، انتخاب متغیر[۴] و انتخاب صفت[۵] می‌نامند، نقش مهمی در مدل‌سازی آماری برای داده‌های با بعد بالا دارد. وجود متغیرهای بی اهمیت یا کم اهمیت، که بار اطلاعاتی مفیدی ندارند، می‌تواند کارایی و تفسیر مدل‌های رگرسیونی و رده‌بندی را تحت تاثیر جدی قرار دهد. ممکن است حذف نکردن آن‌ها علاوه بر افزایش محاسبات، باعث ذخیره‌سازی اطلاعات غیرمفید زیادی شود. برای مساله انتخاب متغیر، راه‌حل‌ها و الگوریتم‌های فراوانی ارائه شده‌اند که بعضی از آن‌ها قدمتی سی یا چهل ساله دارند. مشکل بعضی از روش‌ها در زمانی که ارائه شده بودند، بار محاسباتی زیاد آن‌ها بود. امروزه با وجود رایانه‌های سریع و منابع ذخیره‌سازی بزرگ، این مشکل به چشم نمی‌آید ولی از طرف دیگر برای مسائل جدید، مجموعه داده‌های بسیار بزرگ باعث شده است که هم‌چنان پیدا کردن الگوریتم‌های سریع به‌عنوان یک مساله مهم مطرح باشد.

روش‌های معمول انتخاب متغیر، شناسایی متغیرها را بدون در نظر گرفتن اثر متقابل[۶] آن‌ها انجام می‌دهند؛ در حالی که تجربه محقق ممکن است حاکی از لزوم وجود اثر متقابل در تحلیل باشد. در سراسر این پایان‌نامه منظور از داده‌های با بعد بالا، مجموعه داده‌هایی است که در آن تعداد متغیرهای توضیحی بیش‌تر از تعداد مشاهدات در دسترس است.

در این پایان‌نامه قصد داریم ضمن معرفی دو روش انتخاب متغیر؛ یعنی روش غربال‌گری مستقل مطمئن[۷] (SIS) و روش پیشنهادی مبتنی بر ضریب همبستگی تعمیم‌یافته[۸]، ابتدا بعد داده‌ها را کاهش داده و سپس به مساله رده‌بندی آن‌ها بپردازیم. منظور از کاهش بعد، همان تقلیل تعداد متغیرهای توضیحی است، که خود به عنوان یک بخش از روش‌های کاهش داده در داده‌کاوی محسوب می‌شود. شایان ذکر است که در تحلیل داده‌های با بعد بالا، بسیاری از روش‌های کلاسیک آماری همچون مدل‌های رگرسیونی و رده‌بندی دچار چالش شده و در مواردی غیر قابل بهره برداری‌اند. بنابراین ضروری است ابتدا بعد این داده‌ها کاهش داده شود تا بتوان از این ابزارهای آماری استفاده بهینه داشت. عملکرد دو روش مذکور را نیز بر اساس مثال‌های شبیه‌سازی و مطالعات موردی مورد بررسی قرار خواهیم داد.

---

۱ High dimensional data
۲ Dimension reduction
۳ Feature selection
۴ Variable selection
۵ Attribute selection
۶ Interaction effect
۷ Sure Independence Screening
۸ Generalized correlation coefficient

۲.۱ پیشینه تحقیق

روش‌های انتخاب متغیر بسیار متنوع هستند که با دیدگاه‌های متفاوت مورد بحث و مقایسه قرارگرفته‌اند. فن و لیو (۲۰۱۰) علاوه بر مرور روش‌های مختلف انتخاب متغیر، بیش‌تر به روش‌های انتخاب متغیر مبتنی بر برآوردهای درستنمایی پرداختند. این روش‌ها در مدل برازش داده‌شده به داده‌ها، با اضافه کردن یک تابع تاوان[۱] (جریمه) به تابع درستنمایی، ضرایب کوچک نزدیک به صفر را حذف کرده و بدین وسیله فقط متغیرهای توضیحی با ضرایب رگرسیونی معنادار را در مدل باقی نگه می‌دارند که نتیجه آن کاهش بعد می‌باشد. آن‌ها علاوه بر معرفی چند تابع تاوان، ویژگی‌هایی نیز برای توابع تاوان خوب بر شمردند که از جمله می‌توان به میزان اریبی برآوردها و ویژگی تنک بودن[۲] اشاره کرد. فن و همکاران (۲۰۰۸) برای رده‌بندی داده‌های با بعد بالا، مثل داده‌های ریزآرایه[۳]، روش قواعد مستقل درنوردیدن خصیصه‌ها[۴] (FAIR) را به عنوان یک روش کارا نسبت به روش رده‌بندی فیشر معرفی کردند. فن و لیو (۲۰۰۸) یک روش انتخاب متغیر مبتنی بر همبستگی، به نام SIS را معرفی کردند و نشان دادند که این روش با احتمال متمایل به یک، بعد از کاهش بعد، متغیرهای با اهمیت را حفظ می‌کند. آن‌ها روش SIS را با توابع تاوان لاسو[۵] نیز ادغام کردند. فن و همکاران (۲۰۰۹) تعمیمی از روش SIS را با نام غربال‌گری مستقل مطمئن تکراری[۶] (ISIS) معرفی و از آن در رده مدل‌های خطی تعمیم‌یافته استفاده کردند. گویان و الیس‌سیف (۲۰۰۳) روش‌های انتخاب متغیر صافی[۷] و پوشه[۸] را معرفی و از آن‌ها برای کاهش بعد داده‌های بیان ژنی، ریزآرایه و رده‌بندی متون استفاده کردند. هال (۲۰۰۰) از یک روش انتخاب متغیر برای داده‌های رده‌بندی تحت عنوان انتخاب متغیر مبتنی بر همبستگی[۹](CFS) استفاده کرد. هیو و همکاران (۲۰۰۹) عملکرد روش‌های انتخاب متغیر مانند صافی، پوشه و چند روش دیگر را برای رده‌بندی داده‌های با بعد بالا مورد بررسی قرار دادند. دودایت و همکاران (۲۰۰۲) به مقایسه روش‌های رده‌بندی تومورها با استفاده از داده‌های بیان ژنی پرداختند. سایز و همکاران (۲۰۰۷) تکنیک‌های انتخاب متغیر مانند روش‌های پوشه، صافی و جداسازی[۱۰] را در بیوانفورماتیک مورد مقایسه قرار دادند. تیب شیرانی (۱۹۶۶) و ژو (۲۰۰۶) برای کاهش بعد داده‌ها از روش‌های انتخاب متغیر مبتنی بر مدل، با نام‌های لاسو و لاسو تطبیقی[۱۱] که مبتنی بر پیش‌بینی هستند، استفاده کردند.

اغلب رهیافت‌های انتخاب متغیر مبتنی بر رتبه‌دهی متغیرها، مورد توجه ویژه‌ای هستند. در میان آن‌ها روش‌های مبتنی بر همبستگی از اهمیت بالایی برخوردارند (سایز و همکاران، ۲۰۰۷). این نوع

---

۱ Penalty function
۲ Sparsity
۳ Microarray
۴ Features Annealed Independence Rules
۵ LASSO
۶ Iterative SIS
۷ Filter
۸ Wrapper
۹ Correlation-beasd Feature Selection
۱۰ Embedded
۱۱ Adaptive LASSO

روش‌ها با دو دیدگاه یک متغیره (رتبه‌دهی متغیرها به‌صورت تکی) و چند متغیره (رتبه‌دهی زیرمجموعه‌ای از متغیرها)، عمل رتبه‌دهی را انجام می‌دهند (هال و مایلر، ۲۰۰۹). حضور اثرات متقابل در انتخاب مهم‌ترین متغیرها، تلاش جدیدی است که توسط هال و ژو (۲۰۱۴) مطرح شد و مجموعه روش‌های انتخاب متغیر را غنی‌تر کرده. برای مرور بیش‌تر در زمینه انتخاب متغیر، مطالعه تحقیقات فن و پینگ (۲۰۰۴)، فن و رین (۲۰۰۶) و فن و همکاران (۲۰۰۸) نیز مفید می‌باشند.

۳.۱ ضرورت انجام تحقیق

برای داده‌های با بعد بالا، اغلب روش‌های انتخاب متغیر مانند SIS، رگرسیون گام به گام[۱]، رگرسیون چند متغیره سازوار اسپلاین[۲] (MARS) و روش صافی، متغیرهای توضیحی را با توجه به همبستگی آن‌ها با متغیر پاسخ به‌صورت تکی رتبه‌بندی می‌کنند و سپس متغیرهای توضیحی شاخص را برای ساخت مدل رده‌بندی در نظر می‌گیرند. روش‌های مذکور هزینه محاسباتی چندانی ندارند ولی مشکل اصلی آن‌ها در نظر نگرفتن اثرات متقابل متغیرها می‌باشد. این در حالی است که در بخشی از حوزه‌های علوم، مانند پزشکی، بررسی این اثرات گاهی اوقات قابل چشم‌پوشی نیست. در بین روش‌های رتبه‌دهی متغیرها، روش‌های مبتنی بر همبستگی دارای تکنیک‌های ساده و نتایج ظریف هستند که به‌طور گسترده در عمل استفاده می‌شوند (هال، ۲۰۰۰؛ هال و مایلر، ۲۰۰۹ و هال و ژو، ۲۰۱۴). در این تحقیق با به کارگیری رویکرد ضریب همبستگی تعمیم‌یافته هال و ژو (۲۰۱۴)، اثرات متقابل موثر را با ارائه یک الگوریتم، شناسایی می‌کنیم. یکی از فواید این رهیافت آن است که می‌تواند عملکرد روش‌های رده‌بندی ساده نظیر k- نزدیک‌ترین همسایگی و هم‌چنین مرکز-مبنا[۳] را بهبود بخشد.

۴.۱ داده‌های با بعد بالا

یکی از مشکلات جدی در فرآیند تحلیل داده‌ها، رویارویی با حجم خیلی زیاد و متنوعی از داده‌ها است، که هر یک دارای رفتار متفاوتی هستند. جمع‌آوری متغیرهای توضیحی در پایگاه داده‌ها به‌طور طبیعی افزایش ابعاد داده‌ها را به دنبال دارد و این یکی از مشکلات اساسی شناخته‌شده در تحلیل داده‌ها است. البته این مشکل، فرصت‌های ناب متنوعی برای گسترش روش‌های آماری مناسب برای آماردان‌ها ایجاد کرده است. ویژگی‌های فضاهایی با ابعاد بالا برای ما ناملموس است، چرا که ما در طبیعت با فضاهای دو و سه بعدی مانوس هستیم.

برای روشن‌تر شدن مسئله، چند ویژگی برای داده‌هایی با ابعاد بالا را بررسی می‌کنیم (اسماعیلی مهدی، ۱۳۹۳). به منظور تحلیل داده‌ها به حجم مناسبی از داده‌ها نیاز داریم. طبیعی است که با تنوع بیش‌تر در ویژگی داده‌ها که ما آن را با ابعاد بالاتر نشان می‌دهیم، مقدار داده‌های موجود در پایگاه داده‌ها

[۱] Stepwise regression

[۲] Multivariate Adaptive Regression Spline

[۳] Centroid-based

نیز باید بیش‌تر شوند. بنابراین با افزایش ابعاد به‌طور حتم افزایش داده‌ها را نیز خواهیم داشت. با یک مثال این موضوع را می‌توان به وضوح مشاهده نمود.

اگر در یک فضای یک‌بعدی ۱۰۰ نمونه داده داشته باشیم، با حفظ چگالی، تعداد این نمونه‌ها در یک فضای ۵ بعدی به $۱۰۰^۵ = ۱۰^{۱۰}$ می‌رسد. به‌عبارت دیگر در فضاهایی با ابعاد بالا برای جمع‌آوری کسری از داده‌ها، شعاع بیش‌تری از داده‌ها باید انتخاب شوند. فرمول زیر این خصوصیت را روشن‌تر بیان می‌کند:

$$E_d(p) = p^{\frac{۱}{d}}$$

که در آن d تعداد ابعاد داده‌ها و p کسری از نمونه‌هاست که مقدار آن مشخص می‌باشد. برای مثال اگر بخواهیم فقط ۱۰ درصد از نمونه‌ها را جمع آوری کنیم، سطح‌های انتخابی در داده‌هایی با ابعاد ۲، ۳ و ۱۰ عبارتند از

$$E_۲(۰/۱۰) = ۰/۱۰^{\frac{۱}{۲}} = ۰/۳۲$$
$$E_۳(۰/۱۰) = ۰/۱۰^{\frac{۱}{۳}} = ۰/۴۶$$
$$E_{۱۰}(۰/۱۰) = ۰/۱۰^{\frac{۱}{۱۰}} = ۰/۸۰.$$

این نتیجه نشان می‌دهد که حجم زیادی از داده‌های کنار هم، فقط بخش کوچکی از داده‌ها را در فضاهایی با ابعاد بالاتر شامل می‌شوند. خصوصیت دیگر داده‌های حجیم، فاصله بیش‌تر نمونه‌ها در ابعاد بالاتر است. برای n نمونه داده در بعد d، فاصله D به‌صورت زیر محاسبه می‌شود:

$$D(d,n) = \frac{۱}{۲} \times (\frac{۱}{n})^{\frac{۱}{d}}.$$

با داشتن ۱۰۰۰۰ نمونه، این فاصله برای ابعاد ۲ و ۱۰ برابر است با

$$D(۲, ۱۰۰۰۰) = \frac{[(\frac{۱}{۱۰۰۰۰})^{\frac{۱}{۲}}]}{۲} = ۰/۰۰۰۵$$
$$D(۱۰, ۱۰۰۰۰) = \frac{[(\frac{۱}{۱۰۰۰۰})^{\frac{۱}{۱۰}}]}{۲} = ۰/۴.$$

بنابراین با افزایش بعد و ثابت نگه داشتن تعداد نمونه‌ها، فاصله میان داده‌ها نیز افزایش می‌یابد. اغلب این خصوصیات، هنگامی که با تعداد محدودی از نمونه‌ها در فضاهایی با ابعاد بسیار بالا رو برو هستیم، اثر و نتیجه‌ای جدی به‌دنبال دارند. تعداد نمونه‌ها، تعداد متغیرهای توضیحی و هم‌چنین تعداد مقادیری که یک متغیر توضیحی می‌تواند داشته باشد، حجم بالای داده‌ها را سبب می‌شوند. در یک فروشگاه زنجیره‌ای بزرگ با میلیون‌ها نمونه روبرو هستیم، یا در علم بیوانفورماتیک مجموعه داده‌ها با هزاران متغیر توضیحی توصیف می‌شوند. بدین علت علی‌رغم وجود تعداد زیاد تکنیک‌های تحلیل، برخی از آن‌ها قادر به رویارویی با این حجم وسیع از داده‌ها نیستند. برای جزئیات بیش‌تر در زمینه داده‌های با بعد بالا می‌توان کارهای هانگ و همکاران (۲۰۰۳)، فن دگیر (۲۰۰۸)، بوهلمن و فن دگیر (۲۰۱۱)، لی و زو (۲۰۰۹) و تونی‌کا و شین (۲۰۱۱) را مطالعه نمود.

۱.۴.۱ چالش‌های داده‌های با بعد بالا

فرض کنید $\mathbf{X} = (X_1^T, X_2^T, \ldots, X_n^T)^T$ ماتریس طرح $n \times p$ با ستون‌های استانداردشده است به طوری که $X_i = (X_{i1}, \ldots, X_{ip})$ بردار p - بعدی متغیرهای توضیحی است. هم‌چنین فرض کنید Y_i متغیر پاسخ دو سطحی با دو رده صفر و یک برای مشاهده iام، $i = 1, \ldots, n$، است. در حالت بعد بالا، وقتی تعداد متغیرهای توضیحی بسیار بزرگ‌تر از تعداد نمونه‌ها است، یعنی $p \gg n$، چهار مشکل عمده فهرست می‌شوند:

الف- ماتریس طرح $\mathbf{X}$ مستطیلی‌شکل بوده به‌طوری که تعداد ستون‌ها (متغیرها) بیشتر از تعداد سطرها (مشاهدات) است. در این مواقع، ماتریس $\mathbf{X}^T\mathbf{X}$ وارون‌پذیر نیست. به دلیل بعد بالا و این‌که متغیرهای توضیحی بی‌اهمیت، در حضور متغیرهای توضیحی با اهمیت، می‌توانند همبستگی بالایی با متغیر پاسخ داشته باشند، ماکسیمم مقدار ضرایب همبستگی بین متغیر پاسخ با متغیرهای توضیحی می‌تواند به‌صورت کاذب بزرگ باشد. این مساله شرایط انتخاب متغیر را با مشکلات جدی مواجه می‌سازد.

ب- در مدل رگرسیونی ممکن است بعضی از متغیرهای توضیحی مهم و موثر به دلیل داشتن ضرایب بسیار کوچک، حذف شوند.

ج- توزیع $Z = \hat{\Sigma}^{-\frac{1}{2}}$ می‌تواند دم‌سنگین باشد، که در آن Σ ماتریس کوواریانس متغیرهای توضیحی است. این به آن معنی است که دقت بردار ضرایب رگرسیونی برآوردشده با چالش جدی روبرو می‌شود.

د- ممکن است با افزایش n ماتریس کوواریانس وارون‌پذیر نباشد.

۵.۱ انتخاب متغیر

فرآیند انتخاب زیرمجموعه مناسبی از متغیرها را انتخاب متغیر می‌نامند. به سه دلیل از روش‌های انتخاب و استخراج متغیر استفاده می‌شود:

۱. ساده کردن مدل برای تفسیرپذیری آسان آن توسط محقق و کاربر

۲. کوتاه شدن زمان آموزش مدل (برازش مدل)

۳. افزایش تعمیم‌پذیری مدل به وسیله کاهش واریانس با کم کردن بیش‌برازش[۱]

بنابراین فرآیند انتخاب متغیر، حجم کار را تا حد زیادی کاهش می‌دهد و مدل‌سازی را تسهیل می‌کند. استخراج متغیر[۲] به معنی استخراج متغیرهای جدید از متغیرهای اصلی است که از طریق انجام برخی تبدیلات روی متغیرهای اصلی صورت می‌گیرد. تحلیل مولفه‌های اصلی[۳] (PCA) نمونه‌ای از روش‌های

[۱] Over fitting

[۲] Feature extraction

[۳] Principal Component Analysis

استخراج متغیر است. هر دو رهیافت انتخاب و استخراج متغیرها در اثربخشی روش‌های رگرسیونی و رده‌بندی و تفسیر بهتر نتایج نقش به سزایی را ایفا می‌کنند. در هنگام استفاده از داده‌های با بعد بالا، ممکن است بسیاری از متغیرهای توضیحی در داده‌ها، غیر موثر و غیر مفید باشند. منظور از متغیرهای توضیحی غیر مفید، آن است که اطلاعات و تغییرات این متغیرها به وسیله متغیرهای دیگر نیز قابل بیان است. منظور از متغیرهای توضیحی غیر موثر نیز متغیرهایی هستند که هیچ اطلاعات مفیدی در رابطه با متغیر پاسخ به ما نمی‌دهند.

به‌طور کلی روش‌های انتخاب متغیر باید متمایز از روش‌های استخراج متغیر در نظر گرفته شوند، زیرا در روش‌های انتخاب متغیر، یک زیرمجموعه از متغیرهای توضیحی انتخاب می‌شود، در حالی که در روش‌های استخراج متغیر، متغیرهای جدیدی بر اساس توابعی از متغیرهای موجود، شکل می‌گیرد. تکنیک‌های انتخاب متغیر اغلب در حوزه‌هایی که تعداد متغیرهای توضیحی در مقایسه با تعداد مشاهدات بسیار زیاد است، کاربرد بیش‌تری دارند. تحلیل داده‌های ریزآرایه و متن کاوی که شامل هزاران متغیر و ده‌ها یا صدها نمونه (مشاهده) است، مثالی از این گونه موارد است. رهیافت‌های انتخاب متغیر می‌توانند مبتنی بر روش‌های زیر باشند:

۱. ضریب همبستگی

۲. مدل‌های رگرسیونی پیشگو و مدل‌های تاوان‌دار (مثل لاسو)

۳. درست‌نمایی ماکسیمم

۴. آزمون‌های آماری

در میان رهیافت‌های انتخاب متغیر، روش‌های مبتنی بر همبستگی بار محاسباتی کمی دارند و به‌طور گسترده مورد استفاده قرار می‌گیرند. روش SIS نمونه‌ای از روش‌های مبتنی بر ضریب همبستگی است. مطالعات زیادی در این زمینه صورت گرفته‌اند که به عنوان چند نمونه می‌توان به بولس‌تیکس (۲۰۰۴)، برادیس و همکاران (۲۰۰۹)، کائو (۲۰۰۷)، چایرومونت و مارتین‌لی (۲۰۰۲)، دونوهو و جین (۲۰۰۸)، فن و همکاران (۲۰۰۹) و گرینش‌تین (۲۰۰۶) اشاره کرد.

روش‌های انتخاب متغیر معمولا بر اساس چگونگی ترکیب، انتخاب الگوریتم و برازش مدل، که منجر به انتخاب بهترین زیرمجموعه از متغیرها می‌شوند، در سه رده انتخاب متغیر زیر مورد بررسی قرار می‌گیرند:

۱. روش‌های پوشه

۲. روش‌های صافی

۳. روش‌های جداسازی

در ادامه، هر یک از این سه روش را به اختصار معرفی می‌کنیم.

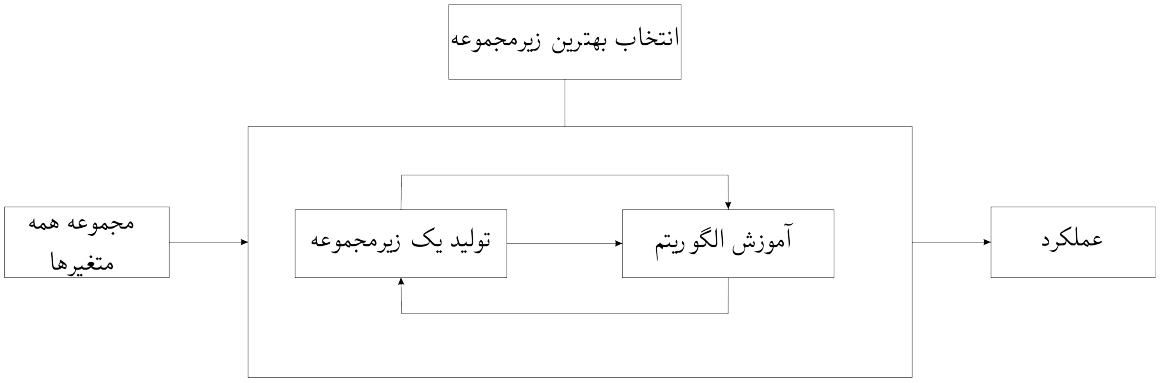
۱.۵.۱ روش پوشه

در این روش، ابتدا زیرمجموعه‌هایی از متغیرهای توضیحی یا همه زیرمجموعه‌ها تولید می‌شوند. سپس بر اساس روش‌هایی مانند الگوریتم ژنتیک و انواع روش‌های رده‌بندی، به جست و جوی بهترین زیرمجموعه از متغیرها پرداخته می‌شود (لو و موتودا، ۲۰۰۸). در واقع، روش‌های پوشه برای هر زیرمجموعه از متغیرهای توضیحی با استفاده از یک مدل، رتبه‌ای می‌دهند که این رتبه‌ها براساس میزان خطای مدل برای داده‌های آزمون، محاسبه می‌شوند. بنابراین زیرمجموعه با کم‌ترین رتبه به‌عنوان بهترین زیرمجموعه در نظر گرفته می‌شود. این روش‌ها بار محاسباتی خیلی بالایی دارند، اما معمولا بهترین زیرمجموعه از متغیرها را در یک رده خاص از مدل‌ها، مانند رده مدل‌های رگرسیونی و رده‌بندی نتیجه می‌دهند (نیلسون، ۲۰۰۷). دو عیب اصلی روش‌های پوشه عبارتند از

۱. خطر افزایش بیش‌برازش وقتی که تعداد نمونه‌ها (مشاهدات) کافی نیست.

۲. مدت زمان بالای محاسبات زمانی‌که تعداد متغیرها بزرگ است.

در شکل ۱.۱، نحوه عملکرد روش پوشه به صورت یک الگوریتم بیان شده است.



شکل ۱.۱: مراحل اجرای روش‌های پوشه

۲.۵.۱ روش صافی

روش‌های صافی، برخلاف روش‌های پوشه که با معیار نرخ خطا به رتبه‌دهی زیرمجموعه‌های مختلف از متغیرهای توضیحی می‌پردازند، بر اساس معیارهای دیگر مانند انواع ضریب همبستگی و آزمون‌های آماری، متغیرهای توضیحی را به‌صورت تکی و گروهی رتبه‌بندی می‌کنند (لی و هان، ۲۰۰۳). این روش‌ها در مواردی که هنوز مجموعه متغیرهای مفید به دست نیامده‌اند، برای انجام محاسبات سریع انتخاب می‌شوند. روش‌های صافی به‌طور کلی به دو دسته تقسیم می‌شوند:

- رتبه‌بندی متغیرها به صورت تکی:

پارامتری: آزمون‌های t -استودنت و فیشر، تحلیل واریانس و رگرسیون

ناپارامتری: آزمون گلموگروف-اسمیرنوف[۱]، مجموع رتبه ویلکاکسن[۲]

- رتبه‌بندی متغیرها به صورت گروهی:

روش‌های CFS، MRMR[۳]، USC[۴] و ضریب همبستگی پیرسون یا ضریب همبستگی گشتاور حاصل ضرب پیرسون[۵] از جمله روش‌های رتبه‌بندی متغیرهای توضیحی به‌صورت گروه‌های دوتایی می‌باشند.

روش‌های صافی معمولا بار محاسباتی کمتری نسبت به روش‌های پوشه دارند اما مجموعه‌هایی از متغیرها را ایجاد می‌کنند که مناسب برازش یک مدل پیشگو نیستند، زیرا تعداد مجموعه متغیرهای حاصل از این روش بیشتر از تعداد متغیرهای به‌دست آمده از روش‌های پوشه می‌باشد و معمولا عملکرد پیش‌بینی ضعیفی نسبت به روش‌های پوشه دارند (لو و موتودا، ۲۰۰۸ و نیلسون، ۲۰۰۷). با این‌که روش‌های صافی موجب کاهش هزینه محاسباتی می‌شوند، اما تمایل به انتخاب متغیرهای بی‌اهمیت دارند. دلیل آن هم نادیده گرفتن ارتباط بین متغیرها (هم‌خطی) است. البته این روش‌ها می‌توانند به عنوان یک گام برای روش‌های پوشه استفاده شوند و اجازه می‌دهند تا روش‌های پوشه برای مسائل بزرگ و پیچیده به‌کار گرفته شوند. در شکل ۲.۱ نحوه عملکرد روش صافی به صورت شماتیک نمایش داده شده است.

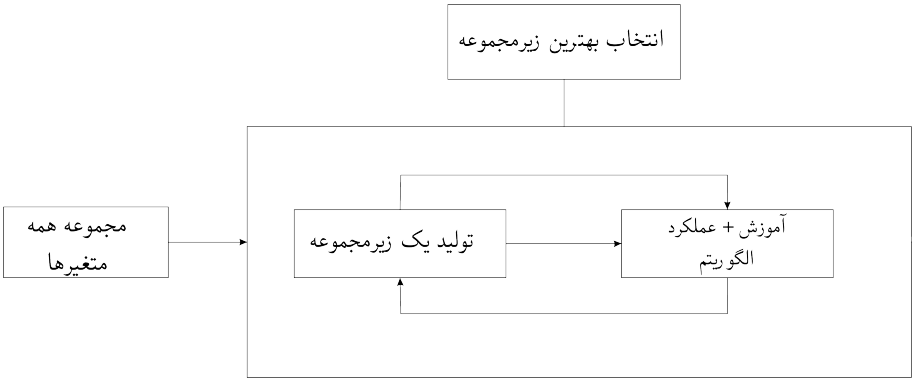
شکل ۲.۱: مراحل اجرای روش‌های صافی

۳.۵.۱ روش جداسازی

در این روش‌ها، عمل انتخاب متغیر همراه با برازش مدل انجام می‌شود. روش لاسو برای ساخت مدل خطی، به‌عنوان نمونه‌ای از این رویکرد می‌باشد که با جریمه کردن ضرایب رگرسیونی موجب حذف بسیاری از ضرایب کوچک و نزدیک به صفر می‌شود و هر متغیری که ضریب رگرسیونی غیر صفر دارد، باقی می‌ماند (مایر و همکاران، ۲۰۰۸ و بونا و همکاران، ۲۰۰۷). یکی از روش‌های دیگر الگوریتم حذف بازگشتی است، که معمولا با ماشین بردار پشتیبان[۶] (SVM) به‌طور مکرر برای ساخت یک مدل و حذف متغیرهای با وزن کم استفاده می‌شود (آبی، ۲۰۰۵ و وپنیک، ۱۹۹۵). روش رگرسیونی MARS، جنگل‌های تصادفی[۷]، درخت تصمیم[۸]، و رگرسیون لجستیک وزنی از جمله نمونه‌های دیگر از روش

[۱] Kolmogorov-smirnov test
[۲] Wilcoxon rank sum test
[۳] Minimum redundancy-maximum relevance
[۴] Uncorrelated shrunken centroid
[۵] Pearson product-moment correlation coefficient
[۶] Support Vector Machines
[۷] Random forests
[۸] Decision tree

جداسازی می‌باشند (نیلسون، ۲۰۰۷). رهیافت جداسازی در مواردی که محاسبات پیچیده است، بین دو روش صافی و پوشه عمل می‌کند. روش انتخاب متغیر گام به گام در رگرسیون، می‌تواند به‌عنوان یک روش جداسازی قلمداد شود. این روش، در هر تکرار بهترین متغیر را اضافه می‌کند (و بدترین متغیر را حذف می‌کند). مهم‌ترین مسئله تشخیص زمان توقف الگوریتم است. در یادگیری ماشین، معمولا، این کار را با اعتبارسنجی متقابل انجام می‌دهند. اخیرا روش‌های جداسازی به‌منظور کاهش خطای رده‌بندی ناشی از برازش مدل، ارائه شده‌اند که این روش‌ها سعی می‌کنند مزایای هر دو روش صافی و پوشه را در ساختار خود به‌کار گیرند. شکل ۳.۱، مراحل اجرای معمول روش‌های جداسازی را نشان می‌دهد.



شکل ۳.۱: مراحل اجرای روش‌های جداسازی

۶.۱ کاهش داده‌ها

پالایش داده‌ها و تغییر شکل آن‌ها و هم‌چنین تکنیک‌هایی که در مرحله جمع‌آوری داده‌ها استفاده می‌شوند، برای داده‌هایی با حجم پایین و حتی با حجم متوسط مناسب هستند. استخراج دانش از داده‌هایی با حجم بسیار بالا مستلزم صرف زمان زیادی است. بنابراین منطقی به‌نظر می‌رسد که ما روش‌هایی را برای کاهش حجم داده‌ها به‌کار ببریم. شاید با تعداد زیاد داده‌ها نتایج بهتری را بتوان به‌دست آورد، اما نمی‌توان به جرات گفت داده‌های با حجم کم، دارای بار اطلاعاتی کمی هستند. تکنیک‌های کاهش داده‌ها می‌توانند بدون از دست دادن درستی داده‌ها و بدون به مخاطره انداختن نتایج نهایی تحلیل، وارد عمل شوند. کاوش بر روی داده‌های با حجم کمتر، هم سریع‌تر است و هم کاراتر. به‌طور حتم کاهش داده‌ها، سادگی ارائه و نمایش داده‌ها را نیز به‌همراه خواهد داشت، به‌نحوی که مدل قابل فهم‌تر خواهد بود. در تکنیک‌های کاهش داده‌ها عمل کاهش می‌تواند در سه سطح انجام شود:

الف- کاهش متغیرهای توضیحی (ستون‌های ماتریس داده)

ب- کاهش نمونه‌ها (سطرهای ماتریس داده)

ج- کاهش در مقادیر یک متغیر توضیحی

الگوریتم‌هایی که از هر سه عمل فوق پشتیبانی می‌کنند، ساده نیستند، به‌خصوص هنگامی که باید بر روی حجم زیادی از داده‌ها اجرا شوند. در ادامه به توضیح هر کدام از آن‌ها می‌پردازیم.

۱.۶.۱ کاهش متغیرهای توضیحی

تکنیک‌های کاهش متغیرهای توضیحی یا به‌عبارت دیگر روش‌های انتخاب متغیرهای توضیحی مناسب، بر اساس یک معیار، می‌توانند زیرمجموعه‌ای مناسب از متغیرهای توضیحی اولیه را انتخاب کنند. چنان‌چه تعداد متغیرهای توضیحی کوچک باشد، ممکن است بتوان تمام زیرمجموعه‌های متغیرهای توضیحی، جهت یافتن متغیرهای توضیحی بهینه، را بررسی کرد. در این حالت، با تعریف پارامتر یا تابعی برای ارزیابی، کلیه حالت‌های ممکن قابل کنترل هستند. پارامترهای ارزیابی بر حسب سازگاری نمونه‌ها، محتویات آن‌ها، فاصله بین نمونه‌ها و وابستگی‌های آماری تعریف می‌شوند. مجموعه داده‌ها می‌توانند شامل صدها یا بیش‌تر متغیر توضیحی باشد، که خیلی از آن‌ها بی‌اهمیت یا اضافه هستند. به عنوان مثال، در یک پایگاه داده دانشگاهی، برای ارزیابی سطح دانش دانشجویان، کلیه نمرات درس‌ها جمع‌آوری می‌شوند. در حالی‌که می‌توان با جمع‌آوری نمرات دروس مشترک در رشته‌های مختلف مثل دروس ریاضی، فیزیک، زبان و حتی مبانی کامپیوتر تحلیل خوبی در زمینه سطح دانش دانشجویان ارائه نمود. توجه داشته باشید، متغیرهای توضیحی بی‌اهمیت مثل شماره دانشجویی و آدرس در مرحله جمع‌آوری داده‌ها یا دیگر مراحل کنار گذاشته می‌شوند.

انتخاب متغیرهای توضیحی مفید می‌تواند مشکل و زمان‌بر باشد، به‌خصوص هنگامی که رفتار داده‌ها خوب شناخته شده نیست. نکته حائز اهمیت این است که کاهش متغیرهای توضیحی مرتبط و نگه‌داری متغیرهای توضیحی بی‌اهمیت، باعث عمل‌کرد نامناسب الگوریتم‌های مورد استفاده شده و منجر به استخراج الگویی ضعیف یا نادرست می‌شود. عمل کاهش بعد در پایگاه داده‌ها، با حذف متغیرهای توضیحی (زاید) آغاز می‌شود و نتیجه احتمالی آن قابل فهم‌تر شدن الگوهای استخراج‌شده است.

از یک نقطه نظر می‌توان انتخاب یک زیرمجموعه مناسب را به الگوریتم‌های جست و جو تشبیه کرد. در واقع باید از یک مجموعه p عضوی از متغیرهای توضیحی، زیرمجموعه‌ای مناسب را مشخص کنیم. طبیعی است اگر همه زیرمجموعه‌ها را در نظر بگیریم، باید $۲^p - ۱$ حالت بررسی شوند. واضح است که برای p بزرگ، رده همه زیرمجموعه‌ها برای جست و جوی بهترین، بسیار گسترده خواهد بود. در حالی‌که با p کوچک می‌توان همه فضا را برای انتخاب بهتر جست و جو کرد. به هر حال با داشتن تعداد زیاد متغیرهای توضیحی، جست و جوی جامع در این فضا برای پیدا کردن زیرمجموعه مناسب، هزینه‌بر خواهد بود و شاید حتی امکان‌پذیر نباشد.

برای کاهش بعد (انتخاب بهترین مجموعه)، می‌توان با مجموعه کاملی از متغیرهای توضیحی شروع و با اجرای یک الگوریتم در هر مرحله یک متغیر توضیحی را حذف کرد تا به مجموعه بهینه رسید. یا با مجموعه تهی شروع و با اجرای الگوریتمی در هر مرحله متغیرهای توضیحی موثر را به این مجموعه اضافه کرد تا به مجموعه بهینه نزدیک شد. به عبارت دیگر، روش انتخاب متغیر می‌تواند از بالا به پایین یا بالعکس، از پایین به بالا عمل کند. تقریبا اکثر الگوریتم‌های موجود برای بررسی و انتخاب متغیرهای توضیحی مناسب، به‌وسیله یکی از این دو روش عمل می‌کنند. گاهی الگوریتم‌های رده‌بندی نیز می‌توانند

برای کاهش بعد استفاده شوند. الگوریتم‌های درخت‌های تصمیم مثل ID3 و C4.5 (هستی و همکاران، ۲۰۰۹) به‌طور طبیعی نامزد خوبی برای این کار هستند. این درخت‌ها ساختاری همانند فلوچارت‌ها دارند که هر گره غیر برگ، شرطی بر روی یک متغیر توضیحی و هر شاخه با توجه به نتیجه شرط، به گره بعدی اشاره می‌کند. در گره‌های برگ یا گره‌های پایانی، برچسب رده مشخص می‌شود و در گره‌های میانی بهترین متغیرهای توضیحی به منظور رده‌بندی داده‌ها انتخاب می‌شوند. هر درخت تصمیم معمولا شامل تمام متغیرهای توضیحی نیست و متغیرهای توضیحی ظاهرشده در درخت، حداقل متغیرهای توضیحی هستند که می‌توان برای تحلیل از آن‌ها استفاده نمود. بنابراین درخت‌های تصمیم می‌توانند به‌عنوان روشی برای کاهش بعد داده‌ها استفاده شوند. سایر الگوریتم‌های رده‌بندی نیز به‌همین طریق می‌توانند مفید باشند. البته فراموش نکنید داده‌های ورودی این روش‌ها دارای برچسب رده برای هر نمونه است.

روش‌های کاهش بعد داده‌ها به دو دسته روش‌های مبتنی بر استخراج متغیر و روش‌های مبتنی بر انتخاب متغیرهای توضیحی، تقسیم می‌شوند. در روش‌های استخراج، یک فضای چندبعدی به یک فضای با ابعاد کمتر نگاشت می‌شود. در واقع، این روش‌ها با ترکیب متغیرهای توضیحی موجود، تعداد کمتری متغیر به‌وجود می‌آورند به‌طوری که متغیرهای حاصل دارای تمام یا بخش عمده‌ای از اطلاعات موجود در متغیرهای توضیحی اولیه باشند. این روش‌ها به دو دسته خطی و غیر خطی تقسیم می‌شوند و تکنیک PCA بهترین روش برای کاهش بعد داده‌ها به‌صورت خطی است (اندرسون، ۱۹۸۴). در روش‌های مبتنی بر انتخاب متغیر، تلاش می‌شود با انتخاب زیرمجموعه‌ای از متغیرهای توضیحی، بعد داده‌ها کاهش داده شود. روش‌های بسیار زیادی وجود دارند که می‌توانند جهت کاهش بعد داده‌ها، متغیرهای توضیحی مناسب را انتخاب کنند. در این پایان‌نامه، برخی از روش‌های انتخاب متغیر را تشریح خواهیم کرد.

۷.۱ کاهش نمونه‌ها

بدون شک در بسیاری از داده‌ها که با آن مواجه می‌شویم، تعداد نمونه‌ها بیش‌ترین مقدار حجم داده‌ای (نسبت به تعداد متغیرهای توضیحی) هستند که ما در حجم داده‌ها با آن مواجه هستیم. بنابراین باید با استفاده از تکنیک‌های کاهش داده‌ها، حجم آن را در حد قابل کنترلی قرار دهیم. کاهش نمونه‌ها یکی از پیچیده‌ترین وظایف در روش‌های کاهش داده‌ها محسوب می‌شود. با این وجود، راه‌هایی هستند که می‌توانیم با کمک آن‌ها حجم نمونه‌ها را کم کنیم. یکی از ساده‌ترین و معمولی‌ترین آن‌ها استفاده از روش‌های نمونه‌گیری[۱] است.

۱.۷.۱ تکنیک‌های نمونه‌گیری

اگر بخواهیم از نمونه‌گیری نتیجه رضایت‌بخشی به‌دست آوریم، باید از مراحلی که نمونه‌ها طی آن‌ها انتخاب می‌شوند، آگاهی داشته باشیم. عمده‌ترین موضوع نمونه‌گیری انتخاب واقعی نمونه است. جلوگیری از جهت‌گیری در فرآیند انتخاب و افزایش دقت در تمام مراحل، ما را به انتخاب واقعی نزدیک می‌کند.

[۱] Sampling

اندازه نمونه، یعنی تعداد نمونه‌های انتخابی، یکی از پارامترهای مهم در روش‌های نمونه‌گیری است که متاسفانه راهکار قاطعی برای یافتن آن وجود ندارد. اما در برخی از کتاب‌ها حداقل حجم نمونه‌ها را با توجه به کاربردهای مختلف پیشنهاد می‌کنند.

در صورتی‌که حجم نمونه بزرگ باشد، شباهت نمونه‌ها با داده‌های اصلی افزایش خواهد یافت، اما کاهش چشم‌گیری در حجم داده‌ها دیده نمی‌شود. از طرفی دیگر با تعداد کم نمونه‌ها برخی از مشخصات داده‌های اولیه در نمونه‌ها وجود نخواهد داشت. نمونه‌گیری تصادفی از ساده‌ترین روش‌های نمونه‌گیری به شمار می‌رود که در دو صورت با جایگزینی[۱] و بدون جایگزینی[۲] انجام می‌شود. در این‌گونه روش‌ها هر یک از نمونه‌ها دارای شانسی مساوی و مستقل برای انتخاب هستند. در نوع بدون جایگزینی هر نمونه فقط یک‌بار شانس انتخاب شدن را دارد، در صورتی‌که در نمونه‌گیری تصادفی با جایگزینی هر نمونه بیش از یک‌بار می‌تواند انتخاب شود. اغلب روش‌های آماری از نوع دوم یعنی با جایگذاری استفاده می‌کنند. نمونه‌گیری منظم نوع دیگری از نمونه‌گیری است که در آن تمام نمونه‌ها بدون هر گونه ترتیبی قرار گرفته و سپس با یک نظم معین نمونه مورد نظر انتخاب می‌شود. برای مثال مشخصات دانشجویان را بدون هیچ‌گونه ترتیبی قرار می‌دهیم و سپس برای نمونه‌گیری از هر ۱۰ دانشجو یکی را انتخاب می‌کنیم. بدین ترتیب پس از نمونه‌گیری اول ۱۰ دانشجوی اول در انتخاب بعدی شانسی برای گزینش ندارند. نوع دیگر نمونه‌گیری طبقه‌ای است. از این روش هنگامی استفاده می‌شود که مایلیم نمونه‌هایی از هر طبقه در نتیجه نهایی وجود داشته باشند. برای مثال، فرض کنید مایلیم از میان دانشجویان یک دانشگاه افرادی را انتخاب کنیم. از آن‌جا که می‌خواهیم در نتیجه نمونه‌گیری ما هم خانم‌ها و هم آقایان شرکت داشته باشند، می‌توان نمونه‌گیری طبقه‌ای را برای دختران و پسران به‌صورت مستقل انجام داد. این روش تضادی با اصل تصادفی بودن ندارد؛ زیرا ابتدا نمونه‌ها به طبقه‌های مختلف تقسیم و سپس هر نمونه به‌صورت تصادفی از طبقه‌ها برگزیده می‌شود. نمونه‌گیری خوشه‌ای نوع دیگری از روش‌های نمونه‌گیری است که در آن به جای نمونه، گروه‌ها (خوشه‌ها) به‌صورت تصادفی انتخاب می‌شوند. این روش هنگامی مناسب است که انتخاب نمونه مشکل یا غیر ممکن باشد. اگر عمل نمونه‌گیری با استفاده از خوشه فقط یک بار انجام شود، به آن نمونه‌گیری خوشه‌ای تک مرحله‌ای گفته می‌شود. چنان‌چه این عمل بیش از یک بار صورت پذیرد، روش را نمونه‌گیری خوشه‌ای چند مرحله‌ای می‌نامند. در این روش می‌توانید پس از نمونه‌گیری خوشه‌ها، نمونه‌ها را از خوشه‌های انتخاب شده گزینش کنید. برای اطلاعات بیشتر در مورد روش‌های مختلف نمونه‌گیری به (شارون، ۲۰۰۹) مراجعه کنید.

۸.۱ کاهش مقادیر یک متغیر توضیحی

مقادیر ثبت‌شده برای یک متغیر توضیحی، معمولا، محدوده‌ای را تشکیل می‌دهند که زیرمجموعه‌ای از دامنه متغیر مزبور است. برای مثال، متغیر سن را در نظر بگیرید. فرض کنید دامنه آن از ۱ تا ۱۵۰ باشد و مقادیر موجود در داده‌ها از ۱۰ شروع و به ۱۳۰ ختم شوند. در واقع حداکثر تعداد اعداد متمایز موجود در

[۱] With replacement

[۲] Without replacement

پایگاه داده برای متغیر توضیحی سن، حدود ۱۲۱ عدد خواهد بود. اغلب، هدف اصلی تکنیک‌های کاهش مقادیر متغیرهای توضیحی، کاهش این تعداد است. در این مثال، شما می‌توانید با تعیین محدوده‌های خاص این ۱۲۱ عدد متمایز را در ۵ گروه سنی کودک، نوجوان، جوان، میانسال و پیر قرار دهید.

۹.۱ انواع داده‌ها و خصوصیات آن‌ها

به‌طور حتم نوع داده‌ها می‌تواند ما را در انتخاب تکنیک‌های تحلیل کمک کند. صرف نظر از این‌که داده‌ها دارای چه ساختاری هستند، به دو گروه کمی[۱] و کیفی[۲] تقسیم‌بندی می‌شوند. در ادامه به توضیح هر کدام می‌پردازیم.

۱.۹.۱ متغیرهای کمی

هرگاه مقادیر متغیر توضیحی و داده مورد نظر را بتوان شمارش یا اندازه‌گیری کرد و سپس آن را به‌صورت عدد بیان نمود، یک متغیر کمی خواهیم داشت که به آن متغیر عددی نیز می‌گویند. در این نوع از داده‌ها ما دارای دو خاصیت ترتیب و فاصله هستیم. این متغیرها می‌توانند پیوسته یا گسسته باشند. متغیرهای پیوسته، متغیرهایی هستند که می‌توانند کلیه مقادیر حقیقی دامنه خود را اختیار کنند. وزن و قد، مثال‌هایی از این نوع می‌باشند. همه عملیات جبری بر روی آن‌ها قابل اعمل هستند. نوعی از این داده‌ها مقدارشان با زمان تغییر نمی‌کنند که با نام داده‌های ایستا[۳] شناخته می‌شوند. در مقابل داده‌های پویا[۴] قرار دارند که مقدارشان با زمان به‌روز می‌شوند. برای تحلیل داده‌های ایستا اغلب از روش‌های داده‌کاوی استفاده می‌شود و برای داده‌های پویا روش‌های محدود و خاصی وجود دارند.

۲.۹.۱ متغیرهای کیفی

متغیرهایی مانند جنسیت اشخاص، محل تولد، آدرس و رنگ چشم را متغیرهای کیفی می‌نامند. حاصل متغیرهای کیفی را نمی‌توان با عدد نشان داد، بلکه بر اساس خاصیتی که مورد نظر است، داده‌ها در طبقات و دسته‌های مختلفی قرار می‌گیرند. گاهی این متغیرها را داده‌های طبقه‌بندی شده و گسسته نیز می‌نامند. فراموش نکنید در متغیرهای کمی گسسته، ترتیب و فاصله معنی دارند، در حالی که این دو خصوصیت برای متغیرهای کیفی گسسته قابل تعریف نیست یا حداقل خصوصیت ترتیب را می‌توان به سختی و برای برخی از موارد خاص تعریف نمود. رابطه‌ای که در این نوع از داده‌ها وجود دارد، مساوی یا نامساوی بودن است. یعنی دو مقدار در این نوع از داده‌ها یا مساوی هستند یا دارای مقادیر یکسان نیستند. برای مثال محل تولد یک شخص یا مساوی محل تولد شخص دیگری است یا خیر. چنان‌چه دسته‌هایی که متغیر در آن‌ها قرار می‌گیرند، دارای یک نوع ترتیب طبیعی باشند، آن متغیر کیفی را متغیر

---

[۱] Quantitative

[۲] Qualitative

[۳] Static data

[۴] Dynamic data

کیفی ترتیبی[۱] گویند. مجموعه مقادیر برای مدارک تحصیلی (دیپلم، کاردانی، کارشناسی، کارشناسی ارشد و دکترا) نمونه بارزی از این نوع است. در داده‌های نوع ترتیبی همان‌طور که از نام آن مشخص است، ترتیب رعایت می‌شود ولی فاصله میان آن‌ها معیار دقیقی نیست. رتبه‌بندی دانشجویان یک کلاس مثال مناسب دیگری است. در این مثال، فاصله رتبه دوم با سوم به‌طور دقیق برابر با فاصله رتبه دهم با یازدهم نیست، هر چند این دو فقط در یک رتبه متفاوتند. در داده‌های کیفی ترتیبی به جز رابطه مساوی بودن، رابطه کوچک‌تر و بزرگ‌تر نیز معنی پیدا می‌کند. به وسیله این روابط می‌توان داده‌ها را دسته‌بندی نمود. دسته‌بندی و گروه‌بندی این نوع از داده‌ها، رابطه نزدیکی با نظریه منطق فازی و دیدگاه نزدیکی با زبان محاوره دارد.

چنان‌چه دسته‌هایی که متغیر در آن‌ها قرار می‌گیرد، دارای هیچ‌گونه ترتیب طبیعی نباشند، هر یک از آن متغیرها را متغیر صوری یا اسمی[۲] کیفی می‌نامند. در محاسبات و تحلیل داده‌هایی که متغیرهای کیفی در آن دخالت دارند، گاهی اوقات ساده‌تر به نظر می‌رسد که در ابتدا متغیرها کدگذاری شوند و سپس از مقادیر معرفی‌شده برای کدها به جای به کار بردن متغیرهای اصلی، استفاده نمود. به عبارت دیگر با نسبت دادن اعداد دلخواه و مناسب به داده‌های کیفی، این داده‌ها به صورت کمی نشان داده می‌شوند. برای این داده‌ها نیز فاصله میان آن‌ها بی‌معنی است. برای کدگذاری داده‌ها، اگر کاربر از ماهیت داده‌ها دانش کافی در اختیار داشته باشد، می‌تواند این کار را به نحوی انجام دهد که در پرس و جوهای پایگاه داده، راحت‌تر بتوان عبارات را بیان کرد یا هر یک از کدها دارای بار معنایی مناسبی باشند.

۱۰۱ درون‌یابی اسپلاین

ریشه تاریخی توابع اسپلاین، مربوط به کار نقشه‌کش‌ها می‌باشد که اغلب نیازمند ترسیم یک منحنی با شیب ملایم بین نقاط روی نقشه بودند. این فرآیند صیقل دادن نامیده می‌شد و با تعدادی وسایل خاص مانند پیستونه، که از پلاستیک و تعدادی منحنی از پیش تعیین‌شده با انحنای متفاوت برای نقشه‌کشی ساخته شده بود، انجام می‌گرفت. نوارهای چوبی بلندی هم مورد استفاده قرار می‌گرفتند. نوارها توسط وزنه‌های روی میز نقشه‌کشی، در امتداد نقاط کنترل قرار می‌گرفتند. از سال ۱۸۹۱ مرسوم بوده که وزنه‌ها، دوک و نوارهای چوبی، اسپلاین نامیده می‌شدند.

طبیعت کشسان نوارها باعث می‌شد، در حالی‌که نقاط اتصال ثابت هستند، بتوانند اندکی خم شوند. در حقیقت چوب یک معادله دیفرانسیل حل می‌کرد تا انرژی کشسانی می‌نیمم گردد. عبارت اخیر به‌عنوان تابع ساده‌ای از خمیدگی شناخته می‌شود. ریاضیات نظری این منحنی‌ها بیش‌تر مرهون پژوهشگران قدیمی به‌خصوص ایزاک شوئنبرگ در دهه‌های ۴۰ و ۵۰ قرن ۱۹ می‌باشد.

دیگر شخصیت‌های مهم موثر در توسعه قدیمی این موضوع (قبل از ۱۹۴۶)، گارت بیرخوف، دبور، آلبرگ، نیلسون، گارابدیان، جانسون، لاندیس، ویتنی، والش و هلادی می‌باشند. اولین کتابی که یک بیان ساختاریافته از نظریه اسپلاین‌ها ارائه داد، کتاب آلبرگ، نیلسون و والش (۱۹۶۷) بود.

---

[۱] Ordinal

[۲] Nominal

کاربرد اسپلاین‌ها در ترسیم‌ها و به‌طور فزاینده‌ای در روش‌های عددی می‌باشد. برای مثال، توابع اسپلاین می‌توانند به‌عنوان توابع آزمایشی روش ریلی ریتز گالرکین برای مسائل مقدار مرزی در معادلات دیفرانسیل معمولی و جزئی، مورد استفاده قرار گیرند. اخیرا از آن‌ها در مبحث پردازش سیگنال نیز استفاده می‌شود.

۱.۱۰.۱ اسپلاین‌ها

تعریف ۱.۱۰.۱. فرض کنید $\triangle = \{a = t_0, t_1, \ldots, t_n = b\}$ که در آن $t_0 < t_1 < \cdots < t_n$ یک افراز $[a, b]$ است. این نقاط گره نامیده می‌شوند. فرض کنید یک عدد صحیح $k \geq 0$ از قبل مشخص‌شده باشد. یک تابع اسپلاین از درجه k و با گره‌های $t_0, t_1, \ldots, t_n$، تابعی است مانند $S : [a, b] \to \mathbb{R}$ که دارای ویژگی‌های زیر است:

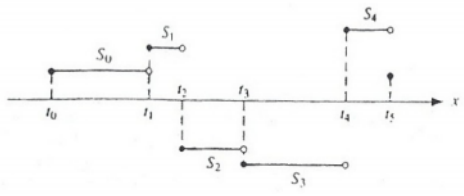
۱. S در هر زیر بازه $[t_{i-1}, t_i]$ یک چندجمله‌ای با درجه k است.

۲. $S \in C^{k-1}[a, b]$ (یعنی مشتقات S تا مرتبه $k-1$ام، به‌خصوص در گره‌ها، موجود و پیوسته‌اند). از این‌رو S یک چندجمله‌ای تکه‌ای پیوسته از درجه حداکثر k است که دارای مشتق‌های پیوسته همه مرتبه‌ها تا $k-1$ می‌باشد.

اسپلاین‌های درجه صفر، تکه‌ای ثابت هستند. یک اسپلاین درجه صفر به‌طور صریح به‌صورت زیر ارائه می‌شود:

$$S(x) = \begin{cases} S_0(x) = c_0 & x \in [t_0, t_1] \\ S_1(x) = c_1 & x \in [t_1, t_2] \\ \vdots & \\ S_{n-1}(x) = c_{n-1} & x \in [t_{n-1}, t_n] \end{cases}$$

بازه‌های $[t_{i-1}, t_i]$ هیچ اشتراکی با یکدیگر ندارند و بنابراین هیچ ابهامی در تعریف یک چنین تابعی در گره‌ها وجود ندارد. یک اسپلاین درجه صفر با ۶ گره در شکل (۱.۳) نشان داده شده است. یک اسپلاین

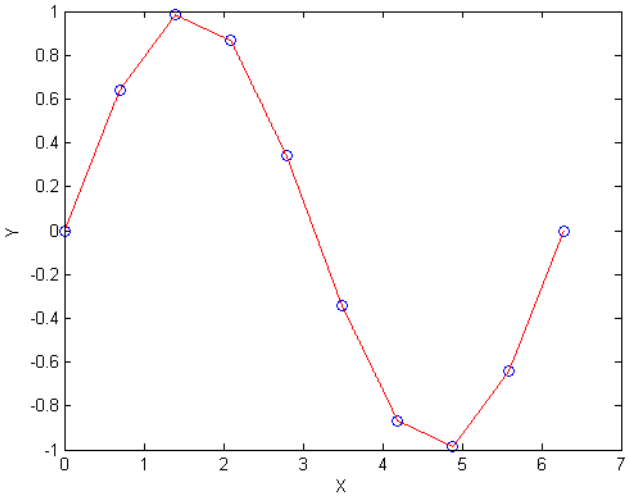


شکل ۴.۱: یک اسپلاین درجه صفر

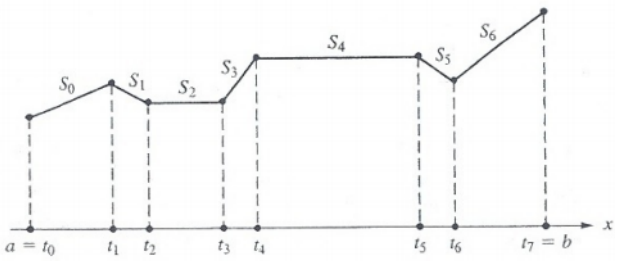
درجه یک به‌طور صریح به‌صورت زیر نوشته می‌شود:

$$S(x) = \begin{cases} S_0(x) = a_0 x + b & x \in [t_0, t_1] \\ S_1(x) = a_1 x + b_1 & x \in [t_1, t_2] \\ \vdots & \\ S_{n-1}(x) = a_{n-1} x + b_{n-1} & x \in [t_{n-1}, t_n] \end{cases}$$

اگر گره‌های t_i و ضرایب a_i و b_i همگی معلوم باشند، آن‌گاه مقدار S در x ابتدا با مشخص کردن زیربازه $[t_i, t_{i+1})$ که x را در بر دارد، به‌دست می‌آید. تابع اسپلاین می‌تواند بر روی تمام خط حقیقی تعریف شود. برای راحتی می‌توانیم از عبارت $a_0 x + b$ بر روی بازه $(-\infty, t_1)$ و از عبارت $a_{n-1} x + b_{n-1}$ بر روی بازه $[t_{n-1}, +\infty)$ استفاده کنیم. تابع S پیوسته است، بنابراین چندجمله‌ای‌های تکه‌ای در گره‌ها به هم می‌پیوندند؛ یعنی $S_i(t_i + 1) = S_{i+1}(t_i + 1)$. در شکل‌های ۵.۱ و ۶.۱ اسپلاین‌های درجه یک، به ترتیب با ۱۰ و ۸ گره نمایش داده شده‌اند.



شکل ۵.۱: اسپلاین خطی تکه‌ای با ۱۰ گره



شکل ۶.۱: اسپلاین خطی تکه‌ای با ۸ گره

تعریف ۱۰۱.۲. منظور از اسپلاین مکعبی[1] روی افراز $\triangle = \{a = t_0 < t_1 < \ldots, < t_n = b\}$، تابع

[1] Cubic Spline

$S : [a,b] \to \mathbb{R}$ با خواص زیر است:

۱. مشتق مرتبه دوم S روی بازه $[a,b]$ پیوسته است.

۲. اسپلاین مکعبی از یک سری چندجمله‌ای‌های درجه سوم تشکیل یافته که هر دو چندجمله‌ای مجاور P_{i-1} و P_i (یعنی دو چندجمله‌ای که بر روی بازه‌های $[t_{i-1}, t_i]$ و $[t_i, t_i+1]$ تعریف شده‌اند) در گره‌های t_i خود و مشتقات اول و دوم‌شان مساوی هستند. بنابراین

$$\begin{cases} P_{i-1}(t_i) = P_i(t_i) \\ P'_{i-1}(t_i^-) = P'_i(t_i^+) \\ P''_{i-1}(t_i^-) = P''_i(t_i^+) \end{cases} \qquad i = 1, \ldots, n-1 \qquad (1.1)$$

حال $n+1$ عدد $\{y_0, y_1, \ldots, y_n\}$ را در نظر بگیرید. منظور از حل مساله درون‌یابی اسپلاینی در زوج‌های (t_i, y_i)، $(0 < i < n)$، یافتن اسپلاین مکعبی $S : [a,b] \to \mathbb{R}$ است که

$$S(t_i) = y_i \qquad i = 1, \ldots, n-1 \qquad (2.1)$$

برای حل این مساله باید چندجمله‌ای P_i توصیف‌شده در (۱.۱) را بیابیم که هر کدام ۴ ضریب برای تعیین شدن دارند. پس جمعا $4n$ مجهول داریم. این در حالی است که تعداد شرایط گفته شده در (۱.۱) برابر $3(n-1)$ و تعداد شرایط در (۲.۱) برابر $n+1$ است که مجموعا $4n-2$ شرط داریم. پس برای آن‌که مساله را بتوان به‌طور منحصر به فرد حل کرد، به ۲ شرط دیگر نیاز داریم. سه شرط زیر در این راه می‌توانند کمک کنند:

۱. $S''(a) = S''(b)$

۲. S متناوب باشد، یعنی برای $k = 0, 1, 2$، $S^k(a) = S^k(b)$.

۳. $S'(a)$ و $S'(b)$ مساوی مقادیر خاصی فرض شوند.

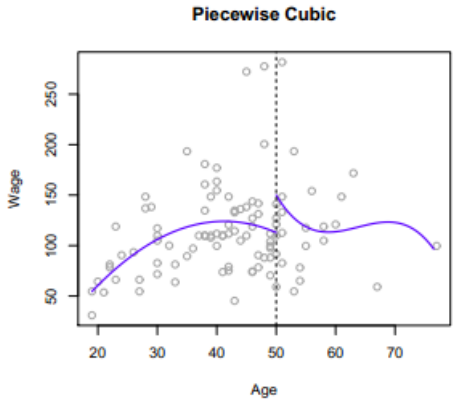
ملاحظه ۳.۱۰.۱. توجه شود، در شرط ۲ که در ظاهر تعداد قیدها $4n+1$ می‌باشد، رابطه $S(a) = S(b)$ به معنی $y_0 = y_n$ است که اگر این شرط از ابتدا برای yها برقرار باشد، آن‌گاه شرط $S(a) = S(b)$ چیز جدیدی را نتیجه نمی‌دهد و هنوز $4n$ قید داریم، و اگر شرط $y_0 = y_n$ از ابتدا برقرار باشد، آن‌گاه مساله اسپلاین فاقد جواب است.

تحت هر یک از این شرایط، اسپلاین مکعبی منحصر به فردی وجود دارد که در این شرایط صدق می‌کند.

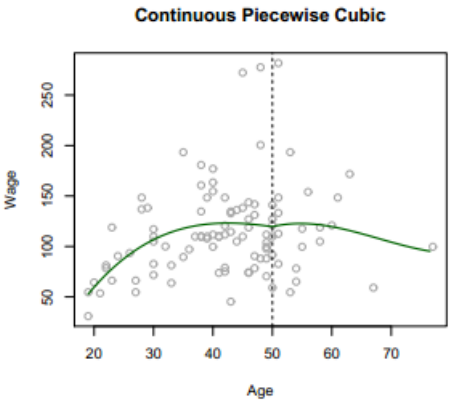
تعریف ۴.۱۰.۱. تابع $h : [a,b] \to \mathbb{R}$ بر بازه $[a,b]$ مطلقا پیوسته است، هرگاه برای هر $\varepsilon > 0$ بتوان $\delta > 0$ یافت که برای هر تعداد $\{(c_i, d_i)\}_{i=1}^n$ از بازه‌های باز غیرمتقاع در $[a,b]$، اگر $\sum_{i=1}^n |c_i - d_i| < \delta$ آن‌گاه

$$\sum_{i=1}^n |h(c_i) - h(d_i)| < \varepsilon.$$

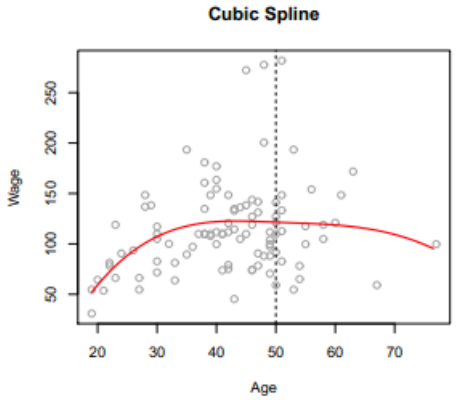
برای مطالعه بیش‌تر به فریدمن (۱۹۹۱) مراجعه کنید.



شکل ۷.۱: اسپلاین مکعبی تکه‌ای



شکل ۸.۱: اسپلاین مکعبی پیوسته



شکل ۹.۱: اسپلاین مکعبی

۱۱.۱ نمادهای O و o

نمادهای O و o، رایج‌ترین نمادها برای توصیف مرتبه مجانبی کمیت‌های غیرتصادفی هستند. تعریف این دو نماد، در ادامه، برگرفته از ون در وارت (۱۹۹۸) می‌باشند.

تعریف ۱.۱۱.۱. دو دنباله از اعداد ثابت $\{a_n\}$ و $\{b_n\}$ را در نظر بگیرید. گوییم $b_n = o(a_n)$، هرگاه

$$\lim_{n\to\infty} |b_n/a_n| = ۰.$$

تعریف ۲.۱۱.۱. دو دنباله از اعداد ثابت $\{a_n\}$ و $\{b_n\}$ را در نظر بگیرید. گوییم $b_n = O(a_n)$، هرگاه $\varepsilon > ۰$، $K(\varepsilon) > ۰$ و عدد صحیح مثبت $N(\varepsilon)$ وجود داشته باشند، به طوری که برای $n \geq N(\varepsilon)$،

$$|b_n| < K(\varepsilon)|a_n|.$$

با استفاده از این دو تعریف، برای هر ثابت حقیقی c، ترتیب‌های $O(a_n)$ و $o(a_n)$ به ترتیب معادل با $ca_nO(۱)$ و $ca_no(۱)$ هستند.

فصل ۲

روش‌های رده‌بندی

در این فصل به معرفی روش‌های معروف و پرطرفدار رده‌بندی، در دو حوزه آمار و یادگیری ماشین، می‌پردازیم. برای این کار ابتدا رده‌بندی و مجموعه داده‌های آموزشی و آزمون را معرفی می‌کنیم. مطالب این فصل عمدتا برگرفته از هستی و همکاران (۲۰۰۹) می‌باشد.

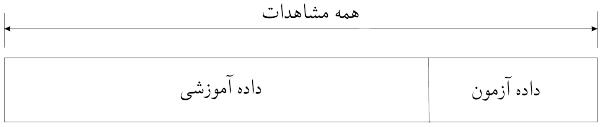
۱.۲ رده‌بندی

پایگاه داده‌ها منبع بسیار غنی از اطلاعات پنهان است که می‌توان به کمک کشف این اطلاعات، تصمیمات هوشمندی را اتخاذ نمود. در این میان، رده‌بندی و رگرسیون، دو شکل از تحلیل داده‌ها محسوب می‌شوند که می‌توان به کمک آن‌ها مدلی جهت توصیف داده‌ها استخراج کرد یا برای داده‌های آتی جهتی متصور شد. رده‌بندی، یک روش گروه‌بندی با نظارت است که در آن گروه‌ها از قبل معلوم هستند و تفاوت اصلی آن با خوشه‌بندی (گروه‌بندی بدون نظارت)، وجود متغیری به نام متغیر پاسخ است که مشخص می‌کند هر مشاهده به کدام گروه (رده) تعلق دارد و هدف از آن تخصیص مشاهدات به گروها و یافتن رابطه یا قاعده‌ای برای تخصیص مشاهدات جدید به گروه‌های موجود است. در نهایت، این رابطه توسط یک ساختار به‌عنوان مدل نمایش داده می‌شود که با کمک آن قادر خواهیم بود نمونه‌ها را به یکی از چندین رده تعریف‌شده منتسب یا مقدار تعیین‌شده‌ای را برای متغیر پاسخ تعیین کنیم. رده‌بندی در بسیاری از علوم کاربرد دارد و تحقیقات گسترده‌ای برای معرفی روش‌های خوب رده‌بندی صورت گرفته است. به عنوان چند نمونه می‌توان به بایکل و لوینا (۲۰۰۴)، هال و همکاران (۲۰۰۸)، هال و همکاران (۲۰۰۹) و لی و همکاران (۲۰۰۴) اشاره کرد.

به منظور معرفی روش‌های مختلف رده‌بندی، ابتدا باید مجموعه داده‌های آزمایشی و آزمون را معرفی کنیم.

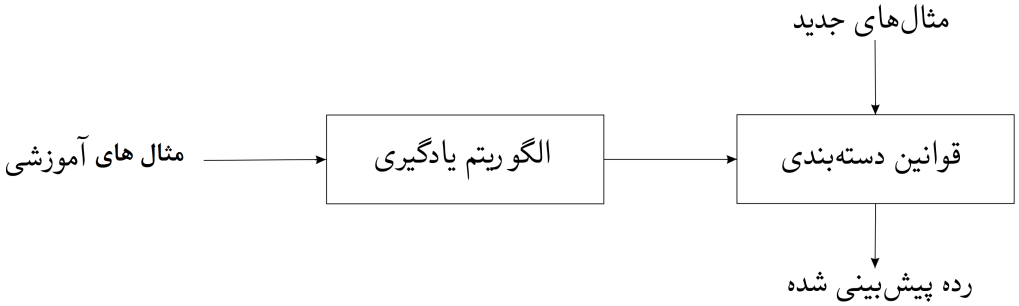
۲.۲ مجموعه داده‌های آموزشی و آزمون

یک روش برای اندازه‌گیری قدرت پیشگویی مدل، آزمودن آن بر روی مجموعه‌ای از داده‌ها است که در برازش آن مدل مورد استفاده قرار نگرفته باشد. در یادگیری ماشین و داده‌کاوی به چنین مجموعه‌ای، مجموعه داده‌های آزمون[۱] و به مجموعه داده‌ای که برای برازش مدل استفاده شده است، مجموعه داده‌های آموزشی[۲] می‌گویند. در شکل ۱.۲ دو مجموعه داده آموزشی و آزمون نشان داده شده‌اند.



شکل ۱.۲: داده‌های آموزشی و آزمون

دقت رده‌بندی یک مدل را می‌توان با میانگین توان دوم خطاهای[۳] (MSE) آن، که بر روی مجموعه آزمون محاسبه شده است، اندازه گیری کرد. به‌طور کلی، از آن‌جا که داده‌های آزمون در برازش مدل استفاده نشده‌اند، این MSE نسبت به مقدار مشابه حاصل از مجموعه آموزشی، بزرگتر خواهد بود. شکل ۲.۲ نحوه برازش و آزمون یک مدل با داده‌های آموزشی و آزمون (مثال‌های جدید) را نشان می‌دهد.



شکل ۲.۲: روند پردازش از داده‌های آموزشی تا داده‌های آزمون

۱.۲.۲ معایب استفاده از مجموعه داده‌های آموزشی و آزمون

دسته‌بندی مشاهدات به دو گروه داده‌های آموزشی و آزمون و استفاده از آن‌ها برای انتخاب مدل رده‌بندی و در انتها استفاده از مدل برای رده‌بندی مشاهدات جدید، با معایبی همراه است که می‌توان به صورت

[۱] Testing data
[۲] Training data
[۳] Mean Squared Errors

زیر فهرست کرد:

الف - در مواردی که حجم داده‌ها کم است، ممکن است در نظر گرفتن بخشی از داده‌ها به‌عنوان مجموعه آزمون مناسب نباشد و مدل منتخب حاصل از این داده‌ها، قابل اطمینان نباشد.

ب - چون این رهیافت یک آزمایش آموزش-آزمون تکی است، اگر تقسیم داده‌ها به دو مجموعه آموزشی و آزمون به‌طور مناسب انجام نشده باشد، آن‌گاه استنباط‌ها می‌توانند گمراه کننده باشند.

توجه داشته باشید که برای داده‌های با بعد بالا (که معمولا تعداد نمونه در مقابل تعداد متغیرها خیلی کوچک است)، این دو مشکل می‌توانند جدی باشند. راه‌حل مرتفع ساختن این دو عیب، استفاده از صورت‌های پیچیده‌تری از مجموعه‌های آموزشی و آزمون مانند LOO[۱]، LKO[۲] و k-fold است.

۲.۲.۲ اعتبارسنجی متقابل

اعتبارسنجی متقابل[۳] که گاهی تخمین گردشی نیز نامیده شده است، یک روش ارزیابی است که مشخص می‌کند نتایج یک تحلیل آماری بر روی یک مجموعه داده تا چه اندازه قابل تعمیم و مستقل از داده‌های آموزشی است. هر آماردانی می‌داند که معیارهای برازش مدل، راهنماهای خوبی برای پاسخ به این سئوال که یک مدل با چه دقتی پیشگویی خواهد کرد، نیستند. به‌عنوان مثال، مقدار بزرگ ضریب تعیین یک مدل، R^2، لزوما به معنی خوب بودن آن مدل نیست، زیرا

۱. با افزودن پارامترهای بیشتر به یک مدل، مقدار R^2 افزایش می‌یابد؛ در حالی که می‌دانیم این امر موجب بیش‌برازش در مدل می‌شود که برای پیشگویی خوب نیست.

۲. در یک رگرسیون چندجمله‌ای، با افزودن جملات با مرتبه بالاتر می‌توان مدلی با برازش بهتر به داده‌ها به‌دست آورد. این عمل باعث کاهش اریبی اما افزایش واریانس می‌شود و برای پیشگویی می‌تواند مضر باشد.

برای پاسخ مناسب به میزان قدرت پیشگویی مدل، روش اعتبارسنجی متقابل اولین گزینه است. البته، این روش مشکلات و انتقادهای خود را داراست. بزرگترین انتقادی که به این روش وارد است، سرعت کند و پیچیدگی محاسباتی است.

از اعتبارسنجی متقابل برای اهداف متنوعی در استنباط آماری و به‌ویژه انتخاب مدل، استفاده می‌شود. به عنوان نمونه می‌توان به موارد زیر اشاره کرد:

الف - سنجش نیکویی برازش[۱] مدل‌ها و انتخاب بهترین مدل.

ب - ارزیابی پایداری[۲] برآوردهای پارامترها.

ج - سنجش دقت رده‌بندی الگوریتم‌های رده‌بندی.

---

[۱] Leave-one-out

[۲] Leave-k-out

[۳] Cross validation

[۱] Goodness of fit

[۲] Stability

د – انتخاب پارامترهای میزان‌ساز[۳] در مدل‌های آماری مختلف، مانند مقدار k در روش k-NN، پارامترهای هسته در روش ماشین بردار پشتیبان، و پارامتر پهنای نوار[۴] در روش هسته[۵].

در عمل، انتخاب مدل بر اساس معیارهای اعتبارسنجی متقابل خیلی بهتر از انتخاب مدل مبتنی بر آزمون‌های آماری است و تقریبا یک اندازه نااریب از MSE واقعی برای داده‌های جدید، ارائه می‌کند. در ادامه، به تعریف این روش و چند نسخه متفاوت آن، می‌پردازیم.

۳.۲.۲ روش‌های اعتبارسنجی

فرض کنید مدلی با یک یا چند پارامتر نامعلوم داریم و یک مجموعه داده که مدل می‌تواند با آن برازش یابد (مجموعه داده آموزشی) در دسترس است. فرآیند برازش، پارامترهای مدل را تا جایی بهینه می‌کند که مدل با داده‌های آموزش تا حد امکان برازش یابد. سپس با اعمال یک نمونه مستقل از داده‌های ارزیابی، با حجم برابر با داده‌های آموزشی، به‌طور کلی به این نتیجه می‌رسیم که مدل به همان خوبی که برای داده‌های آموزشی است (برازش‌یافته)، برای داده‌های ارزیابی نیست. این حالت بیش‌برازش نامیده می‌شود و به‌ویژه زمانی که حجم مجموعه داده‌های آموزشی کم باشد یا زمانی که تعداد پارامترهای مدل زیاد باشد، احتمال وقوع آن وجود دارد. روش ارزیابی متقابل، یک روش ارزیابی (اعتبار) مدل برای بررسی چگونگی تعمیم نتایج یک تحلیل آماری به یک مجموعه داده مستقل می‌باشد.

در روش اعتبار سنجی متقابل، نمونه اصلی، به شیوه‌ای، به k زیرنمونه (نمونه فرعی) با حجم مساوی تقسیم می‌شود. از k نمونه فرعی، یک زیرنمونه به‌عنوان داده ارزیابی برای آزمایش مدل بیرون نگه داشته می‌شود و $k-1$ زیرنمونه باقی‌مانده به‌عنوان داده‌های آموزشی به‌کار می‌روند. سپس فرآیند برازش مدل، k بار (برای هر داده آموزشی حاصل) تکرار می‌شود، به‌گونه‌ای که هر کدام از k زیرنمونه دقیقا یک بار به‌عنوان مجموعه داده ارزیابی به‌کار می‌رود. سپس k تا نتیجه به‌دست آمده از دسته‌ها را، برای ارزیابی مدل، می‌توان با هم ترکیب کرد. مثلا میانگین توان دوم خطاهای حاصل از پیش‌گویی[۶] (MSPE) داده‌های آزمون یک معیار معمول است (شائو، ۲۰۱۲).

برای تقسیم کردن مجموعه داده اصلی به داده‌های آموزشی و آزمون، روش‌های مختلفی پیشنهاد شده‌اند که معمول‌ترین آن‌ها LOOCV[۷]، LKOCV[۸] و اعتبارسنجی k دسته‌ای[۹] هستند. در روش LOOCV، تعداد دسته‌ها برابر مشاهدات است، یعنی $k=n$. فرض کنید n پاسخ مستقل $y_1,\dots,y_n$ مشاهده شده باشند. در این روش، در مرحله i ام، مشاهده i ام یعنی y_i حذف می‌شود یا به‌عبارت دیگر مجموعه (تکی) آزمون را تشکیل می‌دهد. سپس مدل با بقیه مشاهدات برازش داده شده و پاسخ y_i از روی مدل پیشگویی می‌شود. سپس خطای $e_i^* = y_i - \hat{y}_i$ برای مشاهده حذف‌شده محاسبه می‌گردد. به این

---

[۳] Tuning parameters

[۴] Bandwidth

[۵] Kernel

[۶] Mean Squared Prediction Errors

[۷] Leave-one-out cross validation

[۸] Leave-k-out cross validation

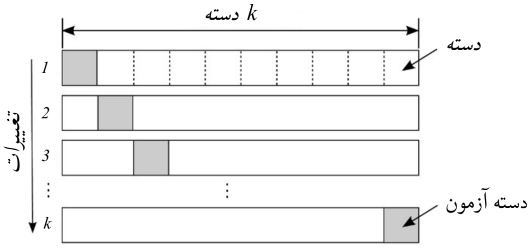
[۹] k-fold cross validation

خطا گاهی خطای پیشگویی نیز می‌گویند. این عمل برای همه داده‌های آزمون، $i = ۱, \ldots, n$، تکرار و خطاهای پیشگویی $e^*_۱, \ldots, e^*_n$ محاسبه می‌شوند. سپس معیار MSPE به صورت زیر به دست می‌آید:

$$MSPE = \sum_{i=۱}^{n} e^{*۲}_i.$$

این روش، استفاده خیلی کاراتری از داده‌های موجود است. زیرا در هر مرحله تنها یک مشاهده حذف می‌شود. البته این روش، جز برای مدل‌های خطی، می‌تواند خیلی زمان‌بر باشد.

در روش LKOCV، در هر بار تقسیم‌بندی داده‌ها k مشاهده (به‌جای یک مشاهده) حذف می‌شود یا به عبارت دیگر مجموعه آزمون را تشکیل می‌دهد و بقیه مراحل مشابه LOOCV انجام می‌شود. یکی از مزیت‌های این رهیافت، محاسبات کمتر آن است. البته در مورد نحوه انتخاب k باید دقت و توجه کافی داشت. معمول‌ترین مقدار برای k برابر ۱۰ است. شکل ۳.۲ نحوه انتخاب مجموعه‌های آزمون را در روش LKOCV نشان می‌دهد.



شکل ۳.۲: نحوه انتخاب مجموعه‌های آزمون در روش LKOCV

یکی دیگر از روش‌های اعتبارسنجی متقابل، اعتبارسنجی متقابل k دسته‌ای است که در آن نمونه اصلی به‌طور تصادفی به k زیرنمونه تقسیم می‌شود و در هر مرحله، یکی از آن‌ها به عنوان مجموعه آزمون در نظر گرفته می‌شود. در این روش نیز انتخاب k برابر ۱۰ متداول است اما در حالت کلی k پارامتر غیر ثابتی است. در این روش، دسته‌ها به‌گونه‌ای انتخاب می‌شوند که میانگین مقدار پاسخ، به‌طور تقریبی، در همه دسته‌ها یکسان باشد. برای جزئیات بیش‌تر به رابرت و کسلا (۲۰۱۰) مراجعه کنید.

۴.۲.۲ اجرای اعتبارسنجی متقابل در نرم‌افزار R

در محیط R، بسته‌ها و توابع متفاوتی وجود دارند که در رده‌های مشخصی از مدل‌های آماری، اعتبارسنجی متقابل را اجرا می‌کنند. در این‌جا به سه مورد از آن‌ها اشاره می‌کنیم:

۱. تابع Validate در بسته Design برای مدل‌های خطی و لجستیک، روش اعتبارسنجی متقابل را اجرا می‌کند. این تابع، اعتبارسنجی مبتنی بر روش خودگردان‌سازی[۱] را نیز فراهم آورده است.

۲. بسته DAAG: این بسته دارای سه تابع cv.lm، CVlm و CVbinary است که روش اعتبارسنجی متقابل با بازنمونه‌گیری تصادفی را در مدل‌های به ترتیب رگرسیون ساده، چندگانه و لجستیک اجرا می‌کند.

[۱] Bootstrap

۳. بسته boot: تابع cv.glm روش اعتبارسنجی متقابل با بازنمونه‌گیری تصادفی k تایی را برای مدل‌های خطی تعمیم‌یافته[۲] (GLMs) (نلدر و ودربرن، ۱۹۷۲) اجرا می‌کند.

۳.۲ روش‌های رده‌بندی

ساخت یک مدل رده‌بندی، یک فرآیند دومرحله‌ای است. در مرحله اول با کمک مجموعه داده‌های آموزشی که برچسب رده تمام نمونه‌های آن مشخص است، مدل ساخته می‌شود. این مرحله به نام مرحله یادگیری شناخته می‌شود. در مرحله دوم، با کمک مجموعه داده‌های آزمون که در آن معمولا برچسب رده‌ها نامعلوم است، مدل به‌دست آمده ارزیابی می‌شود. در واقع ارزشیابی مدل با توجه به این‌که رده چه تعداد از نمونه داده‌های آزمایشی درست تخمین زده شده است، محاسبه می‌شود (تیب شیرانی، ۲۰۰۳). روش‌های رده‌بندی دارای طیف وسیعی است که آن‌ها را می‌توان در دو گروه روش‌های رده‌بندی آماری و یادگیری ماشین دسته‌بندی کرد. نام برخی از این روش‌ها در زیر فهرست شده‌اند و مواردی که با ستاره مشخص شده‌اند را در ادامه به اختصار توضیح می‌دهیم.

- روش‌های رده‌بندی آماری
  ۱. درخت‌های تصمیم (داهان و همکاران، ۲۰۱۴)
  ۲. رده‌بندی بیزی (هستی و همکاران، ۲۰۰۱)
  ۳. رده‌بندی بر اساس مدل‌های رگرسیونی*
  ۴. MARS (فریدمن، ۱۹۹۱)
  ۵. رده‌بندی مرکز-مبنا*
- روش‌های رده‌بندی یادگیری ماشین
  ۱. ماشین بردار پشتیبان (وپنیک، ۱۹۹۵؛ آبی، ۲۰۰۵)
  ۲. رده‌بندی بر اساس تشابه*
  ۳. شبکه‌های عصبی[۳] (کراون و شاولیک، ۱۹۹۷)
  ۴. الگوریتم ژنتیک (ملانی، ۱۹۹۶)
  ۵. منطق فازی[۱] (مندل، ۲۰۰۱)

۱.۳.۲ رده‌بندی بر اساس تشابه

همه روش‌های رده‌بندی برای تخمین رده یک نمونه آزمایشی، ابتدا مدلی را با کمک داده‌ها برازش می‌کنند و با استفاده از مدل، برچسب رده نمونه خواسته‌شده را تخمین می‌زنند. اما تصور کنید بدون ساختن

[۲] Generalized Linear Models
[۳] Artificial neural networks
[۱] Fuzzy

مدل، بتوان رده هر داده آزمایشی را با یافتن مشابه‌ترین نمونه‌های آزمایشی، تعیین کرد. این روش‌ها موسوم به یادگیرنده‌های تنبل[۲] هستند (دابنی، ۲۰۰۵). کلمه تنبل به این جهت انتخاب شده است که روش مذکور تا ورود داده آزمایشی صبر می‌کند و به ساخت مدل برای انجام رده‌بندی نمی‌پردازد. این دسته از الگوریتم‌ها به تکنیک‌های کارایی، جهت ذخیره‌سازی و بازیابی نیاز دارند. در مقابل، این روش‌ها برای داده‌هایی که از یک مدل پیچیده پیروی می‌کنند، عملکرد خوبی دارند (هستی و همکاران، ۲۰۰۱)؛ در این موارد، سایر راهکارها برای ساختن مدل با مشکلاتی روبرو می‌شوند (آلتمن، ۱۹۹۲). یکی از معروف‌ترین روش‌های رده‌بندی بر اساس تشابه، روش k–نزدیک‌ترین همسایگی[۳] (k-NN) نام دارد که در ادامه توضیحاتی پیرامون این روش بیان خواهیم نمود.

روش k-NN

با افزایش قدرت محاسباتی کامپیوترها، روش k-NN نیز محبوب شد. یکی از کاربردهای رایج این روش تشخیص الگو است. برای یک داده آزمایشی، الگوریتم k-NN به دنبال k نمونه از نزدیک‌ترین نمونه‌ها می‌گردد (k نمونه مشابه). نزدیکی دو نمونه با محاسبه تشابه یا فاصله میان این دو نمونه اندازه‌گیری می‌شود. هر نمونه می‌تواند از انواع داده‌ها تشکیل شده باشد که باید تشابه میان آن‌ها بررسی شود. همه روش‌های محاسبه تشابه یا فاصله مربوط به خوشه‌بندی را می‌توان برای الگوریتم k-NN نیز استفاده کرد. پس از یافتن k داده مشابه با نمونه آزمایشی، برچسب رده با بیش‌ترین فراوانی، رده داده آزمایشی انتخاب می‌شود. چنان‌چه مقدار یک برای k در نظر گرفته شود، یعنی روش 1-NN، در این صورت رده نزدیک‌ترین داده به نمونه آزمایشی به عنوان رده تخمینی ارائه می‌شود. اما به‌دلیل وجود داده‌های نوفه و خارج از محدوده، مقدار یک برای k عدد مناسبی نیست. می‌توان مقدار مناسب را به‌صورت تجربی به‌دست آورد. برای مثال، می‌توان با عدد یک شروع و برای مجموعه آزمایشی نرخ خطا را محاسبه کرد. با افزایش مقدار k، این عمل را تکرار می‌کنیم. مقداری از k که باعث حداقل نرخ خطا می‌شود، انتخاب مناسبی است. در کاربردها، معمولا، انتخاب مقادیر پنج و ده برای k نتایج خوبی را به دنبال دارند.

این الگوریتم هم‌چنین می‌تواند برای داده‌هایی که برچسب رده آن‌ها از نوع پیوسته (عددی) است، نیز استفاده شود. در این صورت پس از یافتن k همسایه، میانگین مقادیر حاصل از رده این k نمونه به‌عنوان برچسب رده نمونه آزمایشی برگزیده می‌شود. هرگاه مقدار یا مقادیری از متغیرهای توضیحی در مجموعه داده‌های اصلی یا آزمایشی گمشده باشند در محاسبه تشابه، کم‌ترین و در محاسبه فاصله، بیش‌ترین فاصله برای این متغیرهای توضیحی در نظر گرفته می‌شود. فرض کنید مقادیر یک متغیر توضیحی عددی به فاصله بین صفر تا یک نگاشت می‌شوند. این نگاشت با یک نرمال‌سازی ساده قابل انجام است. برای محاسبه فاصله میان دو نمونه و برای این متغیر توضیحی، اگر هر دو مقدار گمشده باشند فاصله برابر یک و تشابه برابر صفر تعریف می‌شود. اما چنان‌چه یکی از این مقادیر گمشده و دیگری دارای ارزشی برابر

[۲] Lazy learner

[۳] k-Nearest Neighbor

با v باشد، فاصله و تشابه به ترتیب از فرمول‌های زیر به‌دست می‌آیند:

$$Dis(A_i, A_j) = Max(|۱ - v|, |۰ - v|)$$
$$Sim(A_i, A_j) = Min(|۱ - v|, |۰ - v|).$$

برای متغیرهای توضیحی غیر عددی، کافی است حداقل یکی از آن‌ها گمشده باشد تا فاصله برابر با یک و تشابه برابر با صفر در نظر گرفته شود.

با انتساب وزن به هر یک از متغیرهای توضیحی، می‌توان درصد مشارکت این متغیرها در محاسبه تشابه یا فاصله میان نمونه‌ها را کم‌تر یا بیش‌تر کرد. بدین ترتیب متغیرهای توضیحی نامربوط یا داده‌های نوفه، تاثیر کم‌تری در فرآیند خواهند داشت. به دلیل این‌که الگوریتم k-NN برای یافتن برچسب رده داده‌های آموزشی باید همه داده‌ها را پیمایش کند، این عمل در داده‌هایی با حجم بسیار بالا می‌تواند به‌شدت از کارایی الگوریتم بکاهد. تمهیداتی وجود دارند که پیچیدگی الگوریتم را بهبود می‌بخشند. به عنوان مثال، ذخیره داده‌های اولیه در یک ساختار درختی، می‌تواند پیچیدگی جست و جو و پیمایش را کاهش دهد. روشی دیگر جهت بهبود الگوریتم اولیه، محاسبه فاصله یا تشابه میان زیرمجموعه‌ای از متغیرهای توضیحی به جای فضای کل است.

تشریح روش k-NN را با نمایش مثالی از نحوه اجرای الگوریتم به پایان می‌رسانیم. جدول ۱.۲ حاوی مشخصات ۶ نقطه است که در دو رده A و B قرار دارند. با تنظیم عدد ۳ برای مقدار k می‌خواهیم

جدول ۱.۲: مشخصات دکارتی ۶ نقطه همراه با رده آن‌ها

نقطه	x	y	رده	فاصله تا نقطه p
۱	۲	۵	A	۳٫۰۰
۲	۶	۳	B	۴٫۱۲
۳	۴	۴	A	۲٫۸۳
۴	۱	۲	B	۱٫۰۰
۵	۱	۶	B	۴٫۱۲
۶	۳	۲	A	۱٫۰۰

رده نقطه $p = (x, y) = (۲, ۲)$ را با روش k-NN به‌دست آوریم. در ستون آخر جدول ۱.۲، فاصله اقلیدسی نقطه p با هر یک از نمونه‌ها گزارش شده است. از آن‌جا که مقدار k برابر با ۳ است، الگوریتم ۳ نمونه نزدیک به نقطه p را انتخاب می‌کند. انتخاب الگوریتم، نمونه‌های سوم، چهارم و ششم می‌باشد. در میان این سه نمونه، دو برچسب A و یک برچسب B قرار دارند. بنابراین، رده نقطه p برابر با A تخمین زده می‌شود. جزییات بیشتر در مورد این روش را می‌توانید در هستی و همکاران (۲۰۰۱) و هستی و همکاران (۲۰۰۹) بیابید.

۲.۳.۲ رده‌بندی مرکز-مبنا

فرض کنید برای $i, j = ۱, \ldots, n$، مشاهدات p - بعدی

$$X_i = (X_{i۱}, \ldots, X_{ip})^T, \quad Y_j = (Y_{j۱}, \ldots, Y_{jp})^T$$

n مشاهده از بردار متغیرهای توضیحی متعلق به دو جامعه با رده‌های Π_X و Π_Y باشند و بخواهیم رده مشاهده جدید $Z = (Z_1, \ldots, Z_p)$ را، که یکی از این دو جامعه است، پیش‌گویی کنیم. در روش رده‌بندی بر مبنای فاصله، رده Z را به جامعه Π_X نسبت می‌دهیم اگر فاصله Z از Π_X نسبت به فاصله آن از Π_Y کم‌تر باشد. این تصمیم را می‌توان به صورت زیر بیان کرد: مشاهده Z به رده Π_X تعلق دارد اگر

$$\sum_{k=1}^{p} (d(Z_k, Y_{jk}) - d(Z_k, X_{ik})) > 0 \qquad (۱.۲)$$

که در آن $d(Z_k, X_{ik})$ اندازه فاصله k امین مولفه بردار $Z = (Z_1, \ldots, Z_p)$ از مجموعه مشاهدات $X_i = (X_{i1}, \ldots, X_{ip})$ برای k امین مولفه از بردار متغیرها یعنی $X = (X_1, \ldots, X_p)^T$ است و $d(Z_k, Y_{jk})$ به طور مشابه محاسبه می‌شود. در غیر این صورت به رده Π_Y تعلق دارد. به عنوان مثال، در رده‌بندی مرکز-مبنا و متوسط فاصله، Z به رده Π_X تعلق دارد اگر

$$\|Z - \overline{Y}\| > \|Z - \overline{X}\|$$

یا به‌طور معادل

$$\frac{1}{n} \sum_{j} \|Z - Y_j\|^2 > \frac{1}{n} \sum_{i} \|Z - X_i\|^2$$

که در آن‌ها $\|.\|$، نرم اقلیدسی برای سنجش فاصله است. رده‌بندی بر اساس قاعده (۱.۲) معادل این است که در روش رده‌بندی مرکز-مبنا، فاصله را به صورت زیر تعریف کنیم:

$$d(Z_k, X_{ik}) = (Z_k - \overline{X}_k)^2.$$

در واقع برای روش رده‌بندی مرکز-مبنا، قاعده عضویت مشاهده جدید Z به رده Π_X در رابطه (۱.۲) به صورت

$$\sum_{k=1}^{p} (\overline{Y}_k - \overline{X}_k)(\overline{X}_k + \overline{Y}_k - 2Z_k) > 0$$

بازنویسی می‌شود (هال و همکاران، ۲۰۰۹).

۳.۳.۲ رگرسیون لجستیک

رگرسیون لجستیک یکی از تکنیک‌های کاربردی برای تحلیل داده‌های رده‌بندی‌شده است که از آن برای بیان رابطه بین متغیرهای توضیحی با یک متغیر پاسخ از نوع دوسطحی[1] استفاده می‌شود. به عنوان نمونه، اگر نتیجه آزمایشی را به صورت موفقیت و شکست تعریف کنیم، در این حالت متغیر پاسخ دیگر پیوسته نبوده، و به صورت رده‌بندی‌شده خواهد بود. در این حالت برای رده‌بندی مشاهدات جدید، می‌توان از مدل رگرسیون لجستیک استفاده کرد که عضوی از رده GLMs محسوب می‌شود. تاکید می‌کنیم که در این مورد، تعداد رده‌های متغیر پاسخ دو تا (برد و باخت) است.

[1] Binary

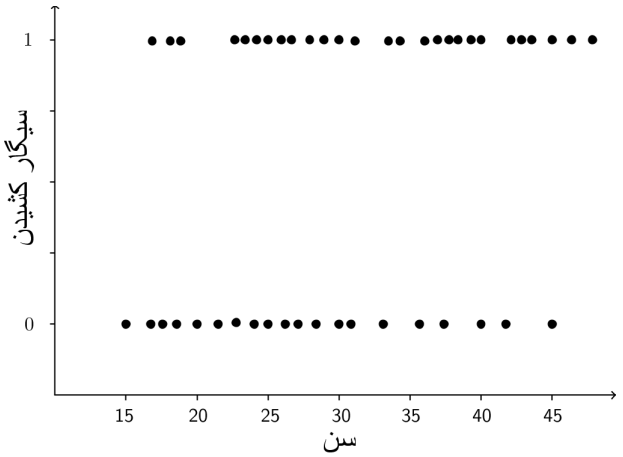
در مسائل رگرسیونی برای بیان رابطه بین متغیر پاسخ با متغیرهای توضیحی، معمولا، از کمیت کلیدی امید ریاضی شرطی متغیر پاسخ به شرط مفروض بودن متغیرهای توضیحی استفاده می‌شود که آن را با نماد $E(Y|X)$ نشان می‌دهند. به این میانگین شرطی، تابع رگرسیون نیز گفته می‌شود. میانگین شرطی در مسائل رگرسیونی به‌صورت یک تابع (خطی یا غیرخطی) از متغیرهای توضیحی است. به عنوان مثال، در مدل رگرسیون خطی ساده، این تابع به شکل

$$E(Y|X=x) = \beta_0 + \beta_1 x \qquad (۲.۲)$$

تعریف می‌شود که برد آن مجموعه اعداد حقیقی، $\Re$، است. اگر متغیر پاسخ از نوع کیفی دوسطحی با مقادیر ۰ و ۱ باشد، آن‌گاه

$$E(Y|X=x) = P(Y=1|X=x).$$

یعنی میانگین شرطی معادل احتمال مشاهده موفقیت $(Y=1)$، به شرط $X=x$ است. بنابراین، برد میانگین شرطی در این حالت، بازه $(0,1)$ است. از آن‌جا که مقادیر مشاهده‌شده متغیر پاسخ ۰ یا ۱ هستند، اگر تنها یک متغیر توضیحی در نظر بگیریم، نمودار پراکنش متغیر پاسخ در برابر متغیر توضیحی، تفسیرپذیر نخواهد بود. برای مثال، اگر متغیر پاسخ سیگاری بودن یا نبودن افراد و متغیر توضیحی سن آن‌ها باشد، نمودار پراکنش آن‌ها، برای یک نمونه مشاهده‌شده خاص، به صورت شکل ۴.۲ خواهد بود (هاسمر و همکاران، ۲۰۰۴).

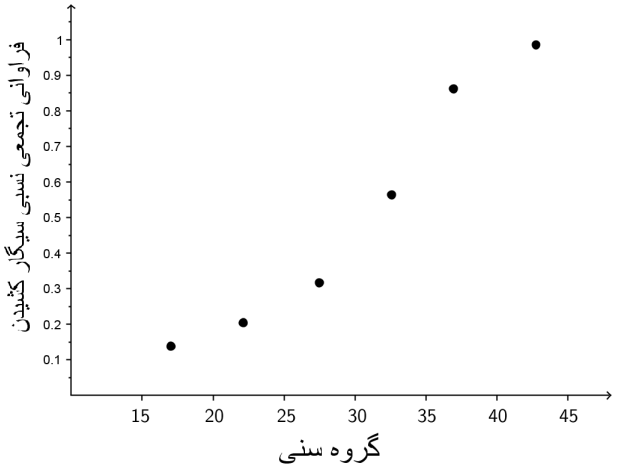


شکل ۴.۲: نمودار پراکنش پاسخ وضعیت سیگاری بودن در مقابل متغیر سن

برای رفع مشکل تفسیرپذیری، می‌توان با تبدیل متغیر کمی سن به متغیر کیفی گروه سنی، این مشکل را رفع کرد. نمودار پراکنش متغیر پاسخ در برابر متغیر گروه سنی در شکل ۵.۲ آمده است که شکل S را در ذهن تداعی می‌کند.

از آن‌جا که توابع توزیع تجمعی متغیرهای تصادفی، تقریباً S شکل هستند، مدل‌سازی متغیر پاسخ می‌تواند مطابق صورت تابعی یکی از این توابع تجمعی باشد. در رگرسیون لجستیک برای این تابع توزیع تجمعی، از توزیع لجستیک استفاده شده است (ایزنمن، ۲۰۰۸).

فرض کنید T معرف تعداد مشاهدات، N معرف تعداد متغیرهای توضیحی، Y بردار متغیر پاسخ، Q ماتریس طرح متغیرهای توضیحی با $T \times (N+1)$ درایه و بردار پارامترها، β، برداری ستونی با



شکل ۵.۲: نمودار پراکنش پاسخ وضعیت سیگاری بودن در مقابل متغیر گروه‌بندی‌شده سن

$N+1$ مولفه باشند. صورت ماتریسی این کمیت‌ها به شکل زیر است:

$$Q=\begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,N} \\ 1 & x_{2,1} & \cdots & x_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T,1} & \cdots & x_{T,N} \end{bmatrix}_{T\times(N+1)} \qquad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}_{(N+1)\times 1} \qquad Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}_{T\times 1}.$$

اگر $X=(X_1,X_2,\cdots,X_N)^T$ و $P(Y=1|X=x)=\pi(x)$، مدل رگرسیون لجستیک چندمتغیره به‌صورت زیر تعریف می‌شود:

$$\pi(x)=P(Y=1|X=x)=\frac{e^{g(x)}}{1+e^{g(x)}}$$

که در آن $g(x)$ تبدیل لجیت بوده و به‌صورت زیر محاسبه می‌شود:

$$g(x)=\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_Nx_N. \qquad (۳.۲)$$

برای یک مشاهده جدید مثل x^*، پیش‌گویی حاصل از مدل برازش‌شده، $\hat{\pi}(x^*)$، مقداری است بین ۰ و ۱. اگر هدف رده‌بندی متغیر پاسخ متناظر با این مشاهده باشد، معمولا با تعریف یک مقدار آستانه (مثلا عدد ۰/۵) رده مشاهده پیش‌گویی می‌شود. مثلا اگر $\hat{\pi}(x^*)>0/5$ آن‌گاه رده برابر ۱ و در غیر این‌صورت برابر ۰ است. برای انجام این مرحله، ابتدا باید پارامترهای مدل برآورد شوند.

۴.۳.۲ برآورد پارامترها

در رگرسیون لجستیک، پارامترها، معمولا، به روش درستنمایی ماکسیمم برآورد می‌شوند. اما چون صورت بسته‌ای برای برآوردهای درستنمایی ماکسیمم این مدل وجود ندارد، از روش‌های عددی برای برآورد آن‌ها استفاده می‌شود.

رابطه (۳.۲) را برای هر یک از مشاهدات، می‌توان به صورت زیر نوشت:

$$g_i(x) = \ln\left(\frac{\pi_i(x)}{۱ - \pi_i(x)}\right) = \sum_{k=۰}^{N} x_{i,k}\beta_k \ , \quad i = ۱, \ldots, T \tag{۴.۲}$$

که در آن برای هر $i = ۱, \ldots, T$، $X_{i,۰} = ۱$. چون متغیر پاسخ از نوع کیفی دوسطحی است، بنابراین متغیر پاسخ دارای توزیع برنولی خواهد بود. یعنی

$$Y_i \sim \text{Bin}\left(۱, \pi_i(x)\right), \quad i = ۱, \ldots, T.$$

برای محاسبه برآوردهای درستنمایی ماکسیمم، باید تابع درستنمایی را محاسبه و ماکسیمم کنیم. تابع درستنمایی مدل به‌صورت زیر تعریف می‌شود:

$$L(\beta|y) = \prod_{i=۱}^{T} \pi_i(x)^{y_i} \left(۱ - \pi_i(x)\right)^{۱-y_i}.$$

با در نظر گرفتن $\pi = \pi(x)$، داریم

$$L(\beta|y) = \prod_{i=۱}^{T} \left(\frac{\pi_i}{۱ - \pi_i}\right)^{y_i} (۱ - \pi_i). \tag{۵.۲}$$

بنا به رابطه (۴.۲)

$$\left(\frac{\pi_i}{۱ - \pi_i}\right) = \mathrm{e}^{\sum_{k=۰}^{N} x_{i,k}\beta_k} \tag{۶.۲}$$

که پس از ساده کردن، نتیجه می‌شود

$$\pi_i = \frac{\mathrm{e}^{\sum_{k=۰}^{N} x_{i,k}\beta_k}}{۱ + \mathrm{e}^{\sum_{k=۰}^{N} x_{i,k}\beta_k}}. \tag{۷.۲}$$

با جایگذاری رابطه (۶.۲) و (۷.۲) در (۵.۲)، داریم

$$\begin{aligned} L(\beta|y) &= \prod_{i=۱}^{T} \left(\mathrm{e}^{\sum_{k=۰}^{N} x_{i,k}\beta_k}\right)^{y_i} \left(۱ - \frac{\mathrm{e}^{\sum_{k=۰}^{N} x_{i,k}\beta_k}}{۱ + \mathrm{e}^{\sum_{k=۰}^{N} x_{i,k}\beta_k}}\right) \\ &= \prod_{i=۱}^{T} \left(\mathrm{e}^{y_i \sum_{k=۰}^{N} x_{i,k}\beta_k}\right) \left(۱ + \mathrm{e}^{\sum_{k=۰}^{N} x_{i,k}\beta_k}\right)^{-۱}. \end{aligned} \tag{۸.۲}$$

رابطه (۸.۲)، هسته تابع درستنمایی برای ماکسیمم شدن است. با توجه به این‌که تابع لگاریتم یک تابع یکنوا است و ماکسیمم کردن تابع درستنمایی معادل با ماکسیمم کردن لگاریتم آن می‌باشد، پس می‌توان نوشت

$$\ell(\beta) = \sum_{i=۱}^{T} y_i \left(\sum_{k=۰}^{N} x_{i,k}\beta_k\right) - \ln(۱ + \mathrm{e}^{\sum_{k=۰}^{N} x_{i,k}\beta_k}). \tag{۹.۲}$$

با مشتق گرفتن از (۹.۲) نسبت به β_k داریم

$$\begin{aligned}\frac{\partial \ell(\beta)}{\partial \beta_k} &= \sum_{i=1}^{T} y_i x_{i,k} - \frac{1}{1+\mathrm{e}^{\sum_{k=0}^{N} x_{i,k}\beta_k}} \frac{\partial}{\partial \beta_k}\left(1+\mathrm{e}^{\sum_{k=0}^{N} x_{i,k}\beta_k}\right) \\ &= \sum_{i=1}^{T} y_i x_{i,k} - \frac{\mathrm{e}^{\sum_{k=0}^{N} x_{i,k}\beta_k}}{1+\mathrm{e}^{\sum_{k=0}^{N} x_{i,k}\beta_k}} x_{i,k} \\ &= \sum_{i=1}^{T} y_i x_{i,k} - \pi_i . x_{i,k}. \end{aligned} \tag{۱۰.۲}$$

برآورد درستنمایی ماکسیمم بردار β با حل دستگاه معادلات

$$\frac{\partial \ell(\beta)}{\partial \beta_k} = 0, k \neq k' = 0, \ldots, N$$

نتیجه می‌شود. لازم به ذکر است که حل این دستگاه نیازمند روش‌های عددی است.

حال مشتق دوم رابطه (۸.۲) را بررسی می‌کنیم. برای این‌که نقاط به‌دست آمده ماکسیمم کننده باشند، باید ماتریس مشتقات دوم، معین منفی باشد. یعنی همه عناصر روی قطر اصلی آن، کوچک‌تر از صفر باشند. مشتق دوم رابطه (۸.۲) به صورت زیر به‌دست می‌آید:

$$\begin{aligned}\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_{k'}} &= \frac{\partial}{\partial \beta_{k'}} \sum_{i=1}^{T} y_i x_{i,k} - x_{i,k}\pi_i \\ &= \frac{\partial}{\partial \beta_{k'}} \sum_{i=1}^{T} -x_{i,k}\pi_i \\ &= -\sum_{i=1}^{T} x_{i,k} \frac{\partial}{\partial \beta_{k'}} \left(\frac{\mathrm{e}^{\sum_{k=0}^{N} x_{i,k}\beta_k}}{1+\mathrm{e}^{\sum_{k=0}^{N} x_{i,k}\beta_k}}\right). \end{aligned} \tag{۱۱.۲}$$

با در نظر گرفتن $u(x) = \sum_{k=0}^{N} x_{i,k}\beta_k$، برای $i = 1, \ldots, T$، داریم

$$\begin{aligned}\frac{d}{dx}\frac{\mathrm{e}^{u(x)}}{1+\mathrm{e}^{u(x)}} &= \frac{(1+\mathrm{e}^{u(x)})\mathrm{e}^{u(x)}\frac{d}{dx}u(x) - \mathrm{e}^{u(x)}\mathrm{e}^{u(x)}\frac{d}{dx}u(x)}{\left(1+\mathrm{e}^{u(x)}\right)^2} \\ &= \frac{\mathrm{e}^{u(x)}\frac{d}{dx}u(x)}{\left(1+\mathrm{e}^{u(x)}\right)^2} \\ &= \frac{\mathrm{e}^{u(x)}}{1+\mathrm{e}^{u(x)}}\frac{1}{1+\mathrm{e}^{u(x)}}\frac{d}{dx}u(x). \end{aligned}$$

در نتیجه، رابطه (۱۱.۲) را می‌توان به‌صورت

$$l''(\beta) = \frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_{k'}} = -\sum_{i=1}^{T} x_{i,k}\pi_i(1-\pi_i)x_{i,k'} < 0$$

نوشت. بنابراین نقاط به‌دست آمده از رابطه (۱۰.۲) ماکسیمم‌کننده لگاریتم تابع درستنمایی مذکور هستند.

اگر رابطه (۱۰.۲) را با $l'(\beta)$ نشان دهیم و $\beta^{(0)}$ مقادیر اولیه دلخواه برای بردار β باشد، آنگاه در روش تکراری نیوتن–رافسون، مقدار جدید β برابر می‌شود با

$$\beta^{(1)} = \beta^{(0)} + \left[-l''(\beta^{(0)}) \right]^{-1} l'(\beta^{(0)}). \qquad (۱۲.۲)$$

برای بازنویسی و فهم آسان‌تر رابطه (۱۲.۲)، فرض کنید μ یک بردار ستونی با T مولفه

$$\mu_i = \mathrm{E}(Y_i|x_i) = \pi_i$$

باشد. بنابراین، مشتق اول لگاریتم تابع درستنمایی را می‌توان به صورت ماتریسی نوشت:

$$l'(\beta) = Q^T(y - \mu)$$

که یک بردار ستونی با $N+1$ مولفه $\frac{\partial \ell(\beta)}{\partial \beta_k}$ می‌باشد. اگر W یک ماتریس قطری با مولفه‌های $\pi_i(1-\pi_i)$ باشد، مشتق دوم لگاریتم تابع درستنمایی را نیز می‌توان به صورت ماتریسی به شکل زیر نوشت:

$$l''(\beta) = -Q^T W Q,$$

که یک ماتریس مربعی $(N+1) \times (N+1)$ با مولفه‌های $\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_{k'}}$ می‌باشد. بنابراین رابطه (۱۲.۲) را به صورت زیر می‌توان بازنویسی کرد:

$$\beta^{(1)} = \beta^{(0)} + \left[Q^T W Q\right]^{-1} Q^T(y - \mu). \qquad (۱۳.۲)$$

رابطه (۱۳.۲) تا زمانی‌که مولفه‌های بردار β به همگرایی برسند، تکرار می‌شود.

فصل ۳

تشخیص اثرات اصلی و متقابل موثر در داده‌های با بعد بالا

مدل‌های با تعداد متغیرهای توضیحی کم، همواره دلخواه استفاده‌کنندگان آمار بوده‌اند؛ چرا که این‌گونه مدل‌ها روابط بین متغیرها را به صورت ساده‌ای بیان کرده و قابلیت تفسیر ساده‌تری نیز دارند. امروزه هزینه‌های محاسباتی بسیار زیاد، روش‌های سنتی انتخاب متغیر را برای داده‌های با بعد بالا با چالش جدی مواجه ساخته است و باعث شده است که این گونه روش‌ها از نظر محاسباتی به صرفه نباشند. از جمله این روش‌ها می‌توان به روش رگرسیون گام به گام اشاره کرد. این مشکلات زمانی‌که بعد داده‌ها افزایش می‌یابد، دو چندان نیز می‌شوند. بنابراین برای استفاده بهینه از روش‌های معمول آماری، مثل رگرسیون و رده‌بندی، لازم است ابتدا تعداد متغیرهای توضیحی تقلیل یابد. در همین راستا، روش غربال‌گری مستقل مطمئن (فن و لیو، ۲۰۰۸) و یک روش دو مرحله‌ای مبتنی بر ضریب همبستگی تعمیم‌یافته که توسط هال و ژو (۲۰۱۴) معرفی شده است، برای کاهش بعد، را در این فصل تشریح خواهیم کرد.

۱.۳ غربال‌گری مستقل مطمئن

در مسائل رگرسیونی و رده‌بندی برای داده‌های با بعد بالا، معمولا بعد داده‌ها را کاهش داده، سپس به برازش مدل و تحلیل آن پرداخته می‌شود. روش "غربال‌گری مستقل مطمئن" (SIS) یک روش توانمند برای انتخاب متغیر در بعد بالا است که آن را به اختصار شرح می‌دهیم. فرض کنید $\mathbf{X}$ ماتریس طرح $n \times p$ با ستون‌های استانداردشده، $X_i = (X_{i1}, \ldots, X_{ip})$ بردار p - بعدی متغیرهای توضیحی و Y_i متغیر پاسخ برای مشاهده iام، $i = 1, \ldots, n$، باشند. ماتریس طرح را می‌توان به صورت زیر هم نمایش

داد:

$$\mathbf{X} = \begin{bmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{np} \end{bmatrix} = (X_1^T, X_2^T, \dots, X_n^T)^T.$$

در فصل اول اشاره کردیم که برای داده‌های با بعد بالا، وقتی تعداد متغیرهای توضیحی بسیار بزرگ‌تر از تعداد نمونه‌ها است، یعنی $p > n$، دو مشکل عمده وجود دارد:

- ماتریس داده‌ها مستطیلی شکل بوده و تعداد ستون‌ها (متغیرها) بیش از تعداد سطرها (مشاهدات) است. در این حالت ماتریس X^TX وارون‌پذیر نیست و برآورد ضرایب رگرسیونی با مشکل مواجه می‌شود.
- توزیع $Z = \hat{\Sigma}^{-\frac{1}{2}}$ می‌تواند دم‌سنگین باشد که Σ، ماتریس کواریانس متغیرهای توضیحی است. در نتیجه دقت بردار ضرایب رگرسیونی برآوردشده، با چالش جدی روبرو می‌شود.

روش غربال‌گری مطمئن به‌عنوان یک روش سریع و کارآمد، فرآیند کاهش بعد را از ابعاد خیلی بالای p یعنی $exp\{O(n^{\xi})\}$ برای مقادیری از $\xi > 0$، به بعد d $(d \sim o(n))$ انجام می‌دهد (فن و لیو، ۲۰۰۸). این روش بر اساس همبستگی متغیرهای توضیحی با متغیر پاسخ بنا شده است و متغیرهای با اهمیت کم (ضریب همبستگی کوچک) را حذف می‌کند. در این‌جا روش غربال‌گری مستقل مطمئن را برای حالتی که متغیر پاسخ Y از نوع کمی پیوسته است، با استفاده از رگرسیون مولفه به مولفه[1] معرفی می‌کنیم. در این شرایط باید همه متغیرهای ورودی، قبل از اعمال هر فرآیندی همچون برازش مدل یا برآورد پارامتر، استانداردشده باشند.

فرض کنید در یک مدل مفروض شامل همه متغیرهای توضیحی، تعداد متغیرهای موثر برابر s باشد، یعنی $s = |M_*|$ که در آن $M_* = \{1 \le i \le p : \ \beta_i \neq 0\}$ مجموعه ضرایب ناصفر مدل مفروض است. سایر $p - s$ متغیر توضیحی نیز می‌توانند از طریق متغیرهای توضیحی موجود در مدل، با متغیر پاسخ همبسته باشند. برادر p - بعدی $W = (w_1, \dots, w_p)^T$ بر حسب یک مدل رگرسیونی به صورت زیر تعریف می‌شود:

$$W = X^TY$$

که در آن، ستون‌های ماتریس $X_{n\times p}$ استانداردشده‌اند. بنابراین W یک بردار حقیقی مقدار از همبستگی‌های حاشیه‌ای متغیرهای توضیحی با متغیر پاسخ است. برای هر مقدار $\gamma \in (0, 1)$ بردار p - بعدی W را به‌صورت صعودی مرتب کرده و متغیرهای توضیحی مهم برای برازش مدل به کمک مجموعه زیر به دست می‌آیند:

$$M_\gamma = \{1 \le i \le p : |W| \text{ تعداد از بزرگ‌ترین مقادیر } [\gamma n]\}$$

که در آن $[\gamma n]$ مقدار صحیح γn است. این یک راه ساده برای کاهش تعداد متغیرهای توضیحی مدل کامل با p متغیر توضیحی به زیر مدل M_γ با تعداد $d = [\gamma n] < n$ متغیر توضیحی است. به کمک

[1] Componentwise regression

ضریب همبستگی، متغیرهای با اهمیت را، با توجه به همبستگی حاشیه‌ای آن‌ها با متغیر پاسخ، رتبه‌بندی می‌کنیم و متغیرهایی که همبستگی حاشیه‌ای ضعیفی با متغیر پاسخ دارند، حذف می‌شوند. این رهیافت، قاعده عملکرد روش SIS است. در این روش، هر متغیر به صورت مستقل از متغیرهای توضیحی دیگر به‌کار گرفته می‌شود تا مشخص شود که مفید (موثر) است یا خیر. این مفهوم گسترده‌تری از غربال‌گری همبستگی بوده و برای مدل‌های خطی تعمیم‌یافته، مسائل رده‌بندی و غیره نیز به‌کار گرفته می‌شود. لازم به ذکر است هزینه محاسباتی روش SIS متناسب با اندازه ماتریس $n \times p$ می‌باشد. توجه داشته باشید که مدل‌های خطی با تعداد متغیرهای توضیحی بیشتر از مشاهدات ($p > n$)، قابل شناسایی و برازش نیستند. بنابراین برای پیاده‌سازی روش SIS، باید $d = [\gamma n]$ را با احتیاط انتخاب کنیم. به‌عنوان مثال، ممکن است d را طوری انتخاب کنیم که از حجم مشاهدات، n، کمتر باشد. یا بسته به حجم نمونه، به صورت $\frac{n}{logn}$ انتخاب شود. اگر چه این روش به‌منظور کاهش بعد از ابعاد بالا p به ابعادی کمتر از حجم نمونه ارائه شده است، اما استفاده از این روش برای برازش مدلی با $d \geq n$ متغیر توضیحی ($\gamma \geq ۱$) ممکن است. واضح است که مقدار بزرگ d به این معنی است که با احتمال زیاد مدل نهایی M_γ شامل همه متغیرهای توضیحی مدل واقعی M_* است.

روش SIS، یک نوع روش آستانه-سخت[۱] است (فن و لیو، ۲۰۰۸). این روش برای ماتریس طرح متعامد، به‌خوبی قابل درک و اجرا است، اما برای ماتریس‌های طرح کلی، هیچ پشتیبان نظری برای عملکرد مناسب آن وجود ندارد. البته این ایده در اغلب کاربردها استفاده می‌شود. برای مطالعه بیش‌تر به کارهای بیکل (۲۰۰۸)، فن و سونگ (۲۰۰۹) و هال و همکاران (۲۰۰۸) مراجعه کنید.

فن و لیو (۲۰۰۸)، تحت قضیه‌ای، نشان دادند که با احتمال متمایل به ۱، روش SIS متغیرهای موثر را شناسایی و انتخاب می‌کند. در واقع چیزی که در این مورد مهم است، شناسایی شرایطی است که تحت آن‌ها، وقتی $n \to \infty$، به ازای تعدادی از مقادیر γ

$$P(M_* \subset M_\gamma) \to ۱.$$

شرایط لازم برای برقراری این نتیجه در قضیه مذکور مطرح شده‌اند.

۱.۱.۳ منطق روش غربال‌گری مستقل مطمئن

برای درک بهتر منطق یادگیری همبستگی، غربال‌گر رگرسیون ریج آستانه‌ای تکراری[۲] (ITRRS) را معرفی می‌کنیم که یک تعمیم از روش کاهش بعد SIS است، اما برای پیاده‌سازی عملی تنها یادگیری همبستگی لازم است. هم‌چنین، روش ITRRS ابزار تکنیکی خوبی برای درک ما از ویژگی غربال‌گری مستقل مطمئن با ضریب همبستگی و دیگر روش‌ها، ایجاد می‌کند. وقتی تعداد متغیرهای توضیحی بیش‌تر از مشاهدات است، معمولا برآوردهای کمترین توان‌های دوم $\hat{\beta}_{LS} = (X^T X)^+ X^T Y$ دارای خطای بزرگی است، که در آن $(X^T X)^+$ معکوس تعمیم‌یافته (مور - پنرز) می‌باشد (روگر و پنرز ، ۱۹۹۵). در این

[۱] Hard-thresholded

[۲] Iteratively Thresholded Ridge Regression Screener

موارد، یک رهیافت جانشین برای کاهش واریانس برآوردگرها، استفاده از مدل رگرسیون ریج (ستیغی[۱]) است که با افزودن یک پارامتر میزان‌ساز $\lambda > 0$ معادله کمترین توان‌های دوم را حل می‌کند. بردار p – بعدی برآوردگر $W^\lambda = (W_1^\lambda, \ldots, W_p^\lambda)^T$، به صورت زیر تعریف می‌شود:

$$W^\lambda = (X^T X + \lambda I_p)^{-1} X^T Y.$$

هم‌چنین ماتریس کوواریانس این برآوردگر عبارتست از

$$Var(W^\lambda) = \sigma^2 (X^T X + \lambda I_p)^{-T} X^T X (X^T X + \lambda I_p)^{-1} \qquad (۱.۳)$$

که در آن σ^2 واریانس جمله خطا در مدل رگرسیونی است. اگر $\lambda \to 0$، آن‌گاه

$$W^\lambda \to \hat{\beta}_{LS}$$

و برآوردهای مقیاس‌بندی شده رگرسیون ریج با برآوردهای حاصل از رگرسیون معمولی، در حالت مجانبی یعنی زمانی که $\lambda \to \infty$، یکسان است؛ یعنی $\lambda W^\lambda \to W$. همچنین با توجه به عبارت (۱.۳) برای کاهش واریانس برآوردگرهای W^λ، باید یک مقدار غیر صفر بزرگ برای پارامتر λ انتخاب کنیم. از طرفی، چون برآوردهای رگرسیون معمولی یک حالت خاص از برآوردهای رگرسیون ریج هستند، رتبه‌بندی قدر مطلق مولفه‌های W^λ با رتبه‌بندی λW^λ یکسان است. برای هر $\delta \in (0, 1)$، p مولفه W^λ را به صورت صعودی مرتب کرده و یک زیرمدل با $[\delta p]$ متغیر اول $|W^\lambda|$ را به شکل زیر تعریف می‌کنیم:

$$M_{\lambda,\delta}^1 = \{1 \le i \le p : |W_i^\lambda| \text{ تعداد از بزرگ‌ترین مقادیر } [\delta p]\}.$$

با این روش p متغیر توضیحی را کاهش می‌دهیم. با توجه به قضیه ۲ فن و لیو (۲۰۰۸) با انتخاب مناسب λ و δ، با احتمال نزدیک به یک، زیر مدل $M_{\lambda,\delta}^1$ شامل مدل واقعی M_γ بوده و اندازه آن کم‌تر از p و از مرتبه $o(n^\theta)$، برای بعضی از $\theta > 0$، می‌باشد. این ویژگی‌های روش ITRRS سبب می‌شود که به دنبال کاهش تعداد بیش‌تری از متغیرهای توضیحی باشیم. بنابراین روش ITRRS را به صورت زیر بیان می‌کنیم:

۱. زیرمدل $M_{\lambda,\delta}^1$ با اندازه $[\delta p]$ را به‌دست می‌آوریم.

۲. با روشی مشابه، زیرمدل $M_{\lambda,\delta}^2$ را با مجموعه متغیرهای توضیحی حاصل از زیر مدل $M_{\lambda,\delta}^1$ با اندازه $[\delta^2 p]$ به‌دست می‌آوریم، به‌طوری که

$$M_{\lambda,\delta}^2 \subset M_{\lambda,\delta}^1.$$

۳. با تکرار مراحل، سرانجام زیرمدل $M_{\lambda,\delta}^k$ با اندازه $d = [\delta^k p] < n$ را می‌یابیم به‌طوری که $[\delta^{k-1} p] \ge n$.

[۱] Ridge regression

این روش متفاوت از روش آستانه رگرسیون ریج بوده و زیرمدل‌ها و برآورد پارامترهای آن در طول تکرارها تغییر می‌کنند. تنها در حالت $\lambda = \infty$ رتبه‌دهی متغیرها در طول تکرارها عوض نمی‌شود. با توجه به این روش می‌توان نتیجه گرفت، روش SIS یک حالت خاص از روش ITRRS است. زیرا رگرسیون مولفه به مولفه حالت خاصی از رگرسیون ریج، به ازای $\lambda = \infty$ است.

۲.۱.۳ روش غربال‌گری مستقل مطمئن در رده‌بندی داده‌های کیفی

روش SIS به آسانی می‌تواند برای کاهش فضای متغیرها در مسائل رده‌بندی به کار رود. فرض کنید n_0 و n_1 تعداد مشاهداتی از متغیر توضیحی X_j باشند که رده آن‌ها به ترتیب ۰ و ۱ است. هم‌چنین فرض کنید $\bar{X}_{j,0}$ و $\bar{X}_{j,1}$ به ترتیب میانگین متغیر توضیحی X_j در رده‌های صفر و یک، و $se(X_j)$ انحراف معیار متغیر توضیحی X_j باشد. تحت این شرایط، برآورد j امین مولفه بردار p - بعدی W در رگرسیون مولفه به مولفه به صورت زیر محاسبه می‌شود:

$$W_j = \frac{(n_1 \bar{X}_{j,1} - n_0 \bar{X}_{j,0})}{se(X_j)}.$$

بقیه مراحل برای انتخاب متغیرهای موثر، مشابه قبل صورت می‌پذیرد.

۳.۱.۳ مشکلات روش غربال‌گری مستقل مطمئن

ایده اصلی روش SIS تنها با به‌کارگیری رگرسیون مولفه به مولفه است. سه مسئله مهم که ممکن است در مواجهه با این رویکرد اتفاق بیفتند، عبارتند از

۱- برخی از متغیرهای بی‌اهمیت که همبستگی بالایی با متغیرهای با اهمیت دارند می‌توانند زودتر از دیگر متغیرهای با اهمیت (مثل متغیر پنهان)، که همبستگی ضعیفی با متغیر پاسخ دارند، به‌وسیله روش SIS انتخاب شوند.

۲- ممکن است یک متغیر مهم که به صورت تکی و حاشیه‌ای با متغیر پاسخ ناهمبسته است، از طریق سایر متغیرها با متغیر پاسخ همبسته باشد، که به وسیله روش SIS انتخاب نشده و وارد مدل نهایی نشود.

۳- وجود هم‌خطی بین متغیرهای توضیحی می‌تواند انتخاب متغیرهای موثر را خدشه‌دار کند.

با انگیزه و هدف حل سه مشکل ذکرشده در بالا، یک حالت تعمیم‌یافته از روش SIS به‌نام روش غربال‌گری مستقل مطمئن تکراری[۱] (ISIS) توسط فن و لیو (۲۰۰۸) معرفی شد. این روش به ما اجازه می‌دهد علاوه بر اطلاعات حاشیه‌ای متغیرها، به‌طور کامل از اطلاعات مشترک آن‌ها در انتخاب متغیرها نیز استفاده کنیم.

[۱] Iterative Sure Independence Screening

۴.۱.۳ روش غربال‌گری مستقل مطمئن تکراری

در بخش ۱.۳ تاکید کردیم اگر ماتریس طرح داده‌ها متعامد باشد، روش SIS به‌عنوان یک روش انتخاب متغیر قابل اعتماد بوده و نتایج آن مطلوب خواهد بود. اما در داده‌های با بعد بالا ممکن است متغیرهای توضیحی با هم وابستگی خطی (تقریبی یا کامل) داشته باشند که در این صورت با مشکل هم‌خطی مواجه می‌شویم. بنابراین ممکن است با اعمال روش SIS برخی از متغیرهای توضیحی مهم را از دست بدهیم. برای رفع این مشکل و به منظور افزایش کارایی انتخاب متغیر، فن و لیو (۲۰۰۸) روش SIS را به صورت مکرر به‌کار گرفتند که آن را روش ISIS نامیدند. ماهیت این روش انجام مکرر غربال‌گری متغیرها در مقیاس بزرگ با در نظر گرفتن یک معیار دقت انتخاب متغیر، مانند ضریب همبستگی، است. الگوریتم روش ISIS به شرح زیر است:

الف- یک زیرمجموعه k_1 تایی از متغیرها، مثل $A_1 = \{X_{i1}, \ldots, X_{i_{k_1}}\}$، را با استفاده از یک روش غربال‌گری مبتنی بر انتخاب مدل مانند روش SIS به‌دست می‌آوریم. سپس مدل را با این مجموعه از متغیرهای توضیحی بر روی متغیر پاسخ برازش می‌دهیم.

ب- مانده‌های مرحله الف را به‌عنوان متغیر پاسخ جدید در نظر گرفته و مرحله الف را برای $p - k_1$ متغیر باقی‌مانده به‌کار می‌بریم. در نتیجه زیرمجموعه جدید با k_2 متغیر توضیحی جدا از A_1 به‌صورت $A_2 = \{X_{j1}, \ldots, X_{j_{k_2}}\}$ حاصل می‌شود. باید توجه داشته باشیم که برازش مانده‌ها از مرحله قبل بر روی متغیرهای توضیحی باقی‌مانده، می‌تواند به‌طور قابل توجهی اولویت انتخاب آن دسته از متغیرهای بی‌اهمیتی که همبستگی بالایی با متغیر پاسخ از طریق متغیرهای توضیحی $\{X_{i1}, \ldots, X_{i_{k_1}}\}$ دارند را تضعیف کند. زیرا مانده‌ها با این متغیرهای توضیحی ناهمبسته‌اند و این امر موجب رفع مشکل هم‌خطی می‌شود. این عمل، هم‌چنین، باعث می‌شود متغیرهای توضیحی مهم از دست‌رفته در مرحله الف، احتمالا، حفظ شوند. این کار تا زمانی ادامه می‌یابد که l زیرمجموعه جدا از هم $A_1, \ldots, A_l$ حاصل شوند تا مجموعه $A = \cup_{i=1}^{l} A_i$ با اندازه d را تشکیل دهند. در عمل l را می‌توان طوری انتخاب کرد که $|A| < n$ باشد. در نهایت، مجموعه A شامل همه متغیرهای موثر است که با روش ISIS به‌دست آمده‌اند و برای ساخت مدل‌های رده‌بندی استفاده می‌شوند.

۲.۳ کاهش بعد مبتنی بر روش‌های جریمه‌ای

متداول‌ترین روش برآورد پارامترها در رگرسیون، روش‌های کم‌ترین توان‌های دوم معمولی است. یکی از برجسته‌ترین ویژگی‌های این روش، محاسبات نسبتا ساده و سر راست آن است. روش‌های کم‌ترین توان‌های دوم با برخی کاستی‌ها نیز روبرو هستند. برای مثال، در توزیع‌های دم‌سنگین نقاط دور افتاده می‌توانند برآوردگرهای کم‌ترین توان‌های دوم معمولی ضرایب رگرسیونی را با اریبی مواجه سازند. این بدین معناست که قدر مطلق برآوردهای کم‌ترین توان‌های دوم خیلی بزرگ می‌شود و لذا این برآوردگرها

بسیار ناپایدار[۱] می‌شوند. یعنی از نمونه‌ای به نمونه دیگر به‌طور قابل توجهی تغییر می‌کنند. برآوردگرهای کم‌ترین توان‌های دوم معمولی اغلب نااریب، ولی با واریانس زیاد می‌باشند. در برخی مواقع می‌توان با انقباض یا صفر کردن بعضی از ضرایب برآوردگرها، افزایش کمی در اریبی ایجاد کرد و به این وسیله واریانس را کاهش داد و در نتیجه پیش‌گویی را بهبود بخشید. تعداد متغیرهای زیاد می‌تواند منجر به افزایش واریانس پیش‌گویی، $\hat{Y}$، شود و لذا دقت پیش‌گویی‌ها کاهش می‌یابد. با استفاده از روش کم‌ترین توان‌های دوم (تاوانیده) جریمه‌شده[۲] (PLS)، می‌توان این مشکلات را تا حد قابل توجهی کاهش داد. در این روش‌ها به مجموع توان دوم خطاها که باید می‌نیمم شود، یک تابع جریمه (تاوان)، به عنوان جریمه پیچیدگی مدل، اضافه می‌شود و به دنبال می‌نیمم کردن نتیجه آن هستند. به عبارت دیگر، PLS به صورت زیر تعریف می‌شود:

$$\ell(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|)$$

که در آن تابع $p_{\lambda_j}(\cdot)$ یک تابع جریمه (تاوان) است که دارای پارامتر میزان‌ساز λ_j می‌باشد. این پارامتر معمولا λ در نظر گرفته می‌شود و می‌توان آن را به کمک روش‌هایی مانند اعتبارسنجی متقابل (برایمن، ۱۹۹۶؛ تیب‌شیرانی، ۱۹۹۶) انتخاب کرد. بسته به این‌که از چه نوع تابع جریمه استفاده شود، روش‌های مختلفی پیشنهاد شده‌اند که از جمله می‌توان به روش‌های لاسو (تیب‌شیرانی، ۱۹۹۶)، SCAD[۳] (فن و لی، ۲۰۰۱؛ فن و پنگ، ۲۰۰۴)، لاسوی تطبیقی (ژو، ۲۰۰۶) و انتخاب‌گر دانتزیگ[۴] (DS) (کاندس و تائو، ۲۰۰۷) اشاره کرد. روش رگرسیون ریج نیز عضوی از این روش‌ها محسوب می‌شود. برای جزئیات بیش‌تر نیز می‌توانید به ولبرگ (۲۰۰۶)، بیکل و لی (۲۰۰۶)، یان و لین (۲۰۰۶)، رئو و یو (۲۰۰۶) و لیو و فن (۲۰۰۹) مراجعه کنید.

روش SIS را می‌توان با این روش‌های کاهش بعد، برای کاهش بیشتر، ترکیب کرد. در این صورت می‌توان روش‌های ترکیبی مثل SIS-SCAD یا SIS-DS را ایجاد کرد. در ادامه دو روش لاسو و SCAD را به اختصار توضیح می‌دهیم.

۱.۲.۳ روش لاسو

ایده جریمه کردن به منظور کاهش تعداد متغیرهای توضیحی در مدل‌بندی‌های آماری می‌تواند بسیار مفید باشد. انواع روش‌های برآورد ضرایب رگرسیونی مبتنی بر جریمه را می‌توان به دو گروه کلی تقسیم کرد: گروه اول، روش‌های مبتنی بر جریمه کردن اندازه مدل است مانند معیار آکائیک[۵] (AIC) (هیروتاگو، ۱۹۷۴)، معیار c_p (مالو، ۱۹۷۳) و معیار اطلاع بیزی[۶] (BIC) (شوارتز، ۱۹۷۸). گروه دوم، روش‌های

[۱] Unstable
[۲] Penalized Least Squares
[۳] Smoothly Clipped Absoulte Deviation
[۴] Dantzig Selector
[۵] Akaike Information Criterion
[۶] Bayesian Information Criterion

جریمه کردن تک تک متغیرهای توضیحی است که به روش‌های PLS معروف هستند. رایج‌ترین روش در بین روش‌های رگرسیونی جریمه‌شده، روش لاسو است که در مجموع PLS از یک تابع جریمه با نرم L_1 استفاده می‌کند.

مدل رگرسیونی چندگانه زیر را در نظر بگیرید:

$$Y = X\beta + \epsilon$$

که در آن $X_{n\times p}$ ماتریس طرح، $Y_{n\times 1}$ بردار مشاهدات n بعدی، $\beta_{p\times 1}$ بردار ضرایب p بعدی و $\epsilon_{n\times 1}$ بردار خطاها می‌باشد که مولفه‌های آن مستقل و دارای توزیع نرمال با میانگین صفر و واریانس ثابت σ^2 هستند، یعنی $\epsilon_{n\times 1}\sim N_n(0,\sigma^2 I_n)$. در این صورت مجموع PLS به صورت

$$\ell(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

خواهد بود و برآورد کم‌ترین توان‌های دوم جریمه‌شده لاسو به شکل

$$\hat{\beta}_{Lasso} = argmin_\beta\ \ell(\beta)$$

تعریف می‌شود. به پارامتر λ پارامتر انقباضی[1] نیز گفته می‌شود که اندازه انقباض به صفر را کنترل می‌کند. یکی از مزایای مدل رگرسیونی لاسو، با توجه به شکل آن، انقباض ضرایب و مساوی صفر قرار دادن ضرایبی است که تاثیر ناچیزی بر متغیر پاسخ دارند. تعداد ضرایبی که برابر صفر قرار داده می‌شوند به پارامتر λ بستگی دارد. جزئیات بیش‌تر در تیبشیرانی (۱۹۹۶)، بیکل و همکاران (۲۰۰۸)، بونا و همکاران (۲۰۰۷) و مایر و همکاران (۲۰۰۸) قابل مشاهده است.

۲.۲.۳ روش SCAD

فن و لی (۲۰۰۱) استفاده از تابع جریمه SCAD را در مجموع PLS پیشنهاد دادند که توسط فن (۱۹۹۷) معرفی شده است. این تابع، توسط مشتقش به شکل زیر تعریف می‌شود:

$$p'_\lambda(\theta) = \lambda\{I(\theta\le\lambda) + \frac{(a\lambda-\theta)}{(a-1)\lambda}I(\theta\ge\lambda)\}$$

که در آن $a > 2$ و $\theta > 0$. به ازای این مقادیر، داریم

$$\hat{\theta}_{SCAD} = \begin{cases} sign(z)(|z|-\lambda) & for \quad |z|\le 2\lambda \\ \dfrac{\{(a-1)z - sign(z)a\lambda\}}{(a-2)} & for \quad 2\lambda < |z| \le a\lambda \\ z & for \quad |z|\ge a\lambda \end{cases}$$

علاوه بر این، فن و لی (۲۰۰۱) مقدار $a = 3.7$ را پیشنهاد کردند.

[1] Shrinkage

۳.۲.۳ ویژگی‌های یک تابع جریمه خوب

فن و لی (۲۰۰۱) سه شرط را برای یک تابع جریمه خوب فهرست کردند:

۱. نااریبی: به منظور جلوگیری از اریبی، مقدار برآوردگر باید به مقدار پارامتر نامعلوم نزدیک باشد، به‌ویژه وقتی که ضرایب غیر صفر θ_j بزرگ هستند. این شرط باعث کاهش اریبی مدل می‌شود.

۲. تنکی: برآوردگر حاصل باید یک قانون آستانه‌گذاری داشته باشد که به صورت خودکار پارامترهای برآوردشده کوچک را صفر در نظر بگیرد، تا به این وسیله از پیچیدگی مدل کاسته شود.

۳. پیوستگی: برآوردگر حاصل باید پیوسته باشد تا پیش‌گویی‌های مدل پایدار باشند.

تمامی توابع جریمه که نام برده شدند، به غیر از تابع جریمه SCAD، هم‌زمان هر سه شرط بالا را ندارند. برای مثال توابع جریمه L_p به ازای $0 \leq p \leq 1$، شرط پیوستگی را ندارند (فن و لی، ۲۰۰۱). توابع جریمه L_p، برای $p \leq 1$، علی رغم آن‌که منجر به تولید جواب‌های تنک می‌شوند، در مقابل جواب‌های به‌دست آمده توسط آن‌ها نااریب نیستند. مثالی از این نوع روش لاسو است. توابع جریمه L_p، برای $p \geq 1$، نیز شرط تنکی را برقرار نمی‌کنند. مثلا تابع جریمه L_2 (روش رگرسیون ریج) علی رغم آن‌که واریانس را از طریق منقبض کردن جواب‌ها کاهش می‌دهد، اما دارای خاصیت تنکی نیست. بنابراین هیچ یک از توابع جریمه L_p به ازای مقادیر مختلف p هر سه شرط بالا را به صورت هم‌زمان برآورده نمی‌سازند. البته این مطلب به این معنا نیست که چنین توابعی وجود ندارند. برای مثال، تابع جریمه SCAD که توسط فن (۱۹۹۷) ارائه شد، هر سه شرط بالا را دارد. این تابع جریمه نسبت به روش‌های سنتی هزینه محاسباتی کم‌تری دارد و جواب‌های پیوسته ایجاد می‌کند. در نتیجه باعث پایداری در مدل می‌شود. افزون بر این، تابع جریمه SCAD با کاهش تغییرپذیری در پیش‌گویی‌های مدل، شباهت‌هایی نیز به تابع جریمه L_1 دارد و از طرفی نااریب است که جنبه برتری آن نسبت به رگرسیون لاسو است. جزئیات بیش‌تر در ونگ و همکاران (۲۰۰۷)، کیم و همکاران (۲۰۰۸) و کیم و کان (۲۰۰۹) قابل رویت است.

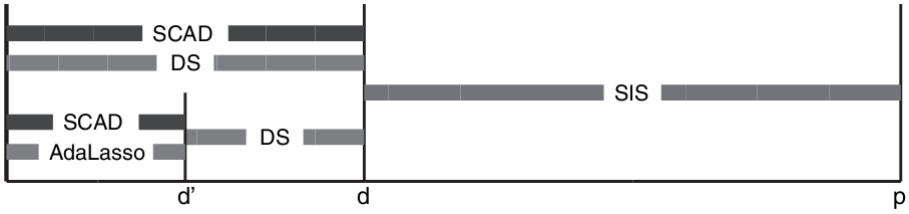
۴.۲.۳ ترکیب روش SIS با روش‌های جریمه‌ای

همان‌طور که اشاره کردیم، روش SIS را می‌توان با سایر روش‌های کاهش بعد ذکرشده ترکیب کرد. هدف از این کار دستیابی به کاهش بعد بیش‌تر است. این مساله وقتی مهم می‌شود که بعد داده‌ها خیلی بالا[1] باشد. در این حالت، فن و لیو (۲۰۰۸) پیشنهاد کردند ابتدا روش SIS را برای کاهش بعد p به یک بعد نسبتا بزرگ مثل d (که می‌تواند کمتر از n باشد) انجام دهیم. سپس از یک روش انتخاب متغیر با بعد کوچک‌تر[2] مثل لاسو، SCAD، DS، یا لاسوی تطبیقی برای کاهش بیشتر به بعدی کوچک مثل $d' < d$ بهره ببریم. در این صورت روش‌های ترکیبی را می‌توان با مثلا SIS-SCAD یا SIS-DS و غیره نمایش داد.

[1] Ultrahigh dimension

[2] Lower dimensional reduction

در برخی موارد، ممکن است بخواهیم حتی بیشتر از d' کاهش بعد داشته باشیم. در این صورت می‌توان با ترکیب روش‌های بیشتر به صورت دنباله‌ای، مثل SIS-DS-SCAD، این کار را انجام داد. شکل ۱.۳ نحوه کاهش بعد توسط ترکیب روش‌های SIS و PLS را به صورت شماتیک نمایش می‌دهد.



شکل ۱.۳: روش‌های انتخاب متغیر برای داده‌های با بعد خیلی بالا

۳.۳ روش ضریب همبستگی تعمیم‌یافته

در این بخش، رهیافت رتبه‌دهی مبتنی بر ضریب همبستگی تعمیم‌یافته تجربی بین متغیرهای توضیحی و متغیر پاسخ را معرفی کرده و به شرح مختصر آن می‌پردازیم. این تکنیک بر اساس پیش‌گویی (مدل پیش‌گو) نیست و می‌تواند متغیرهایی که موثرند اما به صراحت بخشی از مدل نیستند، را نیز شناسایی کند. در ادامه، عملکرد این روش را با دو مثال شبیه‌سازی و داده‌های واقعی (در فصل چهارم) ارزیابی کرده و اعتبار آن را با یک روش اعتبارسنجی متقابل، نشان می‌دهیم. این روش می‌تواند به عنوان یک گام و نه یک جایگزین، برای روش‌های انتخاب متغیر مبتنی بر PLS، مثل لاسو، که به عنوان روش‌های انتخاب متغیرهای موثر در داده‌های با بعد بالا به‌کار می‌روند، مورد استفاده قرار گیرد. انواع روش‌های انتخاب متغیر مبتنی بر مدل‌های خطی، برای این روش، قابل استفاده هستند.

اگر رابطه بین متغیر پاسخ Y و متغیرهای توضیحی X خطی باشد، آنگاه استفاده از روش SIS یا ISIS که بر اساس ضریب همبستگی معمولی[۱] عمل می‌کنند، گزینه مناسبی است و می‌تواند خیلی موثر باشد. در حالتی که این رابطه خطی نباشد، می‌توان انتظار داشت که استفاده از روش‌های مبتنی بر ضریب همبستگی معمولی، باعث نادیده گرفتن متغیرهای توضیحی موثر قابل توجهی شود؛ زیرا، در این حالت، متغیرهای توضیحی که رابطه غیرخطی با متغیر پاسخ دارند، بر اساس ضریب همبستگی معمولی معنادار نمی‌شوند. در این موارد، استفاده از روش‌های غیرخطی می‌تواند نتیجه را بهبود بخشد. ضریب همبستگی تعمیم‌یافته از جمله کمیت‌هایی است که برای این منظور پیشنهاد شده است.

۱.۳.۳ ضریب همبستگی تعمیم‌یافته

از میان روش‌های رتبه‌دهی متغیرها، روش‌های مبتنی بر همبستگی دارای تکنیک‌های ساده و مباحث نظری ظریف بوده و در عمل نیز به‌طور گسترده مورد استفاده قرار می‌گیرند (سایز و همکاران، ۲۰۰۷؛

[۱] Conventional correlation coefficient

گویان و الی‌سیف، ۲۰۰۳). رتبه‌دهی مبتنی بر همبستگی می‌تواند متغیرهای توضیحی را مستقل از هم و به‌صورت تکی رتبه‌دهی کند؛ به عنوان نمونه می‌توان به هال و مایلر (۲۰۰۹) و فن و لیو (۲۰۰۸) اشاره کرد. از طرف دیگر این نوع رتبه‌دهی می‌تواند به صورت چندمتغیره (گروهی) نیز عمل کند به‌طوری که یک زیرمجموعه از متغیرهای توضیحی را هم‌زمان رتبه‌دهی کند؛ به عنوان نمونه‌ای از این حالت نیز می‌توان هال (۲۰۰۳) را نام برد. به غیر از روش‌های انتخاب متغیر مبتنی بر همبستگی، همه روش‌های انتخاب متغیر، مانند SIS و لاسو، به منابع محاسباتی زیادی نیاز ندارند. اما آن‌ها اثرات متقابل متغیرها را در نظر نمی‌گیرند که ممکن است در تحقیقات بعضی از زمینه‌های علوم، تجربه محقق حاکی از لزوم اثرات متقابل باشد. هال و ژو (۲۰۱۴) برای حل این مشکل، با استفاده از ضریب همبستگی تعمیم‌یافته، یک الگوریتم جست و جوی ساده معرفی کردند. این روش هزینه محاسباتی زیادی ندارد و به کمک آن اثرات متقابل مهم نیز قابل شناسایی هستند و در هر دو حالت متغیرهای پاسخ پیوسته و گسسته قابل استفاده است.

برای ورود به بحث، ابتدا فرض کنید $\mathcal{H}$ یک فضای برداری است که تمام توابع خطی را نیز شامل می‌شود که در این پایان‌نامه فقط از تقریب‌های اسپلاین و توابع خطی استفاده شده است.

متغیر پاسخ پیوسته

یک مجموعه داده با متغیر پاسخ از نوع پیوسته را در نظر بگیرید. فرض کنید داده‌های مستقل و هم‌توزیع $(X_1, Y_1), \ldots, (X_n, Y_n)$ مشاهده شده باشند، که در آن متغیرهای توضیحی

$$X_i = (X_{i1}, \ldots, X_{ip})$$

بردارهای p بعدی و Y_i ها متغیرهای پاسخ هستند. معمولا، رتبه‌دهی بر اساس ضریب همبستگی خطی بین متغیرهای توضیحی با متغیر پاسخ، زمانی به‌کار گرفته می‌شود که متغیر پاسخ از نوع کمی باشد. استفاده از ضریب همبستگی تعمیم‌یافته برای وابستگی‌های غیرخطی بین متغیرهای توضیحی و متغیر پاسخ توسط هال و مایلر (۲۰۰۹) پیشنهاد شد. اندازه‌های نظری و تجربی ضریب همبستگی تعمیم‌یافته بین متغیر پاسخ Y_i و مولفه j ام بردار متغیرهای توضیحی X_i، X_{ij}، به صورت زیر تعریف می‌شوند:

$$\rho_j = \sup_{h \in \mathcal{H}} \frac{cov\{h(X_{1j}), Y_1\}}{\sqrt{var\{h(X_{1j})\} var(Y_1)}} \tag{۲.۳}$$

و

$$\hat{\rho}_j = \sup_{h \in \mathcal{H}} \frac{\sum_i \{h(X_{ij}) - \bar{h}_j\}(Y_i - \bar{Y})}{\sqrt{\sum_i \{h(X_{ij})^2 - \bar{h}_j^2\} . \sum_i (Y_i - \bar{Y})^2}} \tag{۳.۳}$$

که در آن $\bar{Y} = \frac{1}{n} \sum_i Y_i$ و $\bar{h}_j = \frac{1}{n} \sum_i h(X_{ij})$. در مخرج کسر عبارت‌های (۲.۳) و (۳.۳)، واضح است که مولفه‌های $var(Y_1)$ و $\sum_i (Y_i - \bar{Y})^2$ به j وابسته نیستند. بنابراین می‌توان آن‌ها را بدون تاثیر بر نتیجه رتبه‌بندی متغیرها حذف کرد.

برای رتبه‌بندی متغیرهای ۱ تا p، بر اساس اندازه ضریب همبستگی تعمیم‌یافته آن‌ها با متغیر پاسخ، کافی است $\hat{\rho}_j$ ها را به‌صورت زیر مرتب کنیم:

$$\hat{\rho}_{\hat{J}_1} \geq \hat{\rho}_{\hat{J}_2} \geq \cdots \geq \hat{\rho}_{\hat{J}_p}. \tag{۴.۳}$$

برای سهولت در نمادگذاری، رابطه (۴.۳) را می‌توان به صورت زیر نوشت:

$$\hat{J}_1 \succeq \hat{J}_2 \succeq \cdots \succeq \hat{J}_p.$$

اگر $\mathcal{H}$ فقط به عناصر خطی محدود شود، آن‌گاه کمیت‌های تعریف‌شده در (۲.۳) و (۳.۳)، همان ضریب همبستگی معمولی خواهند بود.

متغیر پاسخ دوسطحی

یک مجموعه داده با متغیر پاسخ از نوع کیفی گسسته را در نظر بگیرید، که در آن متغیر پاسخ با دو رده صفر و یک بیان شده است. در این حالت، رتبه‌دهی مبتنی بر رابطه بین Y_i و X_{ij} بر حسب یک مدل رگرسیون لجستیک انجام می‌شود (هال و مایلر، ۲۰۰۹). به عبارتی، در این حالت فرض می‌کنیم متغیر پاسخ Y_i دارای توزیع برنولی، با احتمال‌های موفقیت و شکست به صورت زیر است:

$$P(Y_i = 0|h, X_{ij}) = \frac{1}{1 + \exp h(X_{ij})}$$

و

$$P(Y_i = 1|h, X_{ij}) = 1 - P(Y_i = 0|h, X_{ij}) = \frac{\exp h(X_{ij})}{1 + \exp h(X_{ij})}$$

که در آن $h \in \mathcal{H}$. برای رتبه‌دهی متغیرهای توضیحی بر اساس همبستگی، در شرایطی که متغیر پاسخ از نوع کیفی باشد، از تابع درست‌نمایی استفاده می‌شود. تابع درست‌نمایی برای Y_i، با داشتن $h(\cdot)$ و X_{ij} به صورت زیر تعریف می‌شود:

$$L_{ij}(y_i|h, X_{ij}) = \left(\frac{t_{ij}}{1 + t_{ij}}\right)^{y_i} \left(\frac{1}{1 + t_{ij}}\right)^{1-y_i} = \frac{t_{ij}^{y_i}}{1 + t_{ij}}$$

که در آن $t_{ij} = \exp h(X_{ij})$. برای سادگی محاسبات، قرینه لگاریتم تابع درست‌نمایی را محاسبه می‌کنیم. داریم

$$\begin{aligned}
\ell_{ij}(h) &= -\log L_{ij}(y_i|h, X_{ij}) = -\log\left[\frac{t_{ij}^{y_i}}{1 + t_{ij}}\right] \\
&= -y_i \log(t_{ij}) + \log(1 + t_{ij}) \\
&= -y_i h(X_{ij}) + \log\{(1 + \exp h(X_{ij})\}.
\end{aligned}$$

به‌طور مشابه، برای $X_{1j}, \ldots, X_{nj}$ داریم

$$\ell_j(h) = \frac{1}{n} \sum_i \ell_{ij}(h).$$

به‌عبارت دیگر، متناظر با ماتریس X، ابتدا ماتریس‌های $L(h)$ و $-\log L(h)$ را به شکل‌های زیر محاسبه می‌کنیم:

$$L(h) = \begin{pmatrix} L_{11} & \ldots & L_{1p} \\ \vdots & \ddots & \vdots \\ L_{n1} & \ldots & L_{np} \end{pmatrix}$$

$$-\log L(h) = \begin{pmatrix} \ell_{11} & \ldots & \ell_{1p} \\ \vdots & \ddots & \vdots \\ \ell_{n1} & \ldots & \ell_{np} \end{pmatrix} \tag{۵.۳}$$

و سپس متوسط هر یک از ستون‌های (۵.۳) مقدار $\ell_j(h)$ خواهند بود. با تعریف

$$\hat{\ell}_j(h) = \inf_{h \in \mathcal{H}} \ell_j(h) \tag{۶.۳}$$

مقادیر $\hat{\ell}_j(h)$ نقش $\hat{\rho}_j$ را در رابطه (۳.۳) بازی می‌کنند. در نتیجه رتبه‌بندی متغیرهای توضیحی به صورت زیر انجام می‌شود:

$$\hat{\ell}_{\hat{J}_1} \leq \hat{\ell}_{\hat{J}_2} \leq \cdots \leq \hat{\ell}_{\hat{J}_p}. \tag{۷.۳}$$

مشابه قبل، رتبه‌بندی (۷.۳) را به صورت زیر نمایش می‌دهیم:

$$\hat{J}_1 \succeq \hat{J}_2 \succeq \cdots \succeq \hat{J}_p.$$

اگر فضای برداری $\mathcal{H}$ تنها شامل توابع خطی باشد، مدل رگرسیون لجستیک کلاسیک نتیجه می‌شود. یعنی مدلی که متغیر توضیحی در آن به‌جای $h(X_{ij})$ خود X_{ij} است. اما برازش مدل رگرسیون لجستیک با در نظر گرفتن فضای برداری کلی $\mathcal{H}$، می‌تواند چالش‌برانگیز باشد. برای مواجهه با این مساله، در این پایان‌نامه، از توابع اسپلاین برای تقریب (ناپارامتری) $h(X_{ij})$ استفاده کرده‌ایم. کاربرد ماکسیمم درستنمایی کناری برای هر متغیر توضیحی برای رتبه‌دهی متغیرها در هال و میلر (۲۰۰۹) و فن و همکاران (۲۰۰۹) بحث شده است.

۴.۳ الگوریتم دومرحله‌ای برای انتخاب توام اثرات اصلی و متقابل

اشاره کردیم که در برخی از موقعیت‌های کاربردی واقعی، اثرات متقابل می‌توانند نقش مهمی در تحلیل مدل‌های آماری داشته باشند، در حالی‌که روش‌هایی نظیر SIS تنها به انتخاب اثرات اصلی می‌پردازند. در این بخش، با توجه به روش انتخاب متغیر مبتنی بر ضریب همبستگی تعمیم‌یافته (ماکسیمم تابع

درستنمایی)، یک الگوریتم دومرحله‌ای برای انتخاب همزمان اثرات اصلی و متقابل معرفی می‌کنیم. در ادامه، هر کدام از مراحل را به‌طور جداگانه تشریح می‌کنیم.

- رتبه‌دهی اثرات اصلی

پس از محاسبه مقادیر ضریب همبستگی تعمیم‌یافته در (۳.۳) برای متغیر پاسخ پیوسته یا ماکسیمم تابع درستنمایی در (۶.۳) برای متغیر پاسخ کیفی با دو سطح صفر و یک، ترتیب زیر را برای $p_1 = p$ مولفه از بردار متغیرهای توضیحی، به دست می‌آوریم:

$$\hat{J}_{11} \succeq \hat{J}_{12} \succeq \cdots \succeq \hat{J}_{1p}$$

که در آن اندیس اول، یعنی ۱، به مرحله اول الگوریتم رتبه‌دهی تاکید می‌کند.

- رتبه‌دهی اثرات اصلی به همراه اثرات متقابل دوتایی

اثرات متقابل به صورت $X_{i\hat{j}_{1k_1}} X_{i\hat{j}_{1k_2}}$ برای $1 \leq k_1 < k_2 \leq p$ تعریف می‌شوند. تعداد اثرات متقابل به همراه p اثر تکی برابر است با $\frac{1}{2}p(p+1) = \frac{1}{2}p(p-1)+p$. برای داده‌های با بعد بالا p عدد بزرگی است و استفاده از $\frac{1}{2}p(p+1)$ تعداد متغیر نیاز به منابع محاسباتی حجیم و هزینه بالایی دارد. بنابراین علاقه‌مند به کاهش این تعداد به p_2 هستیم. اصولا در این کاهش، تعداد اثرات متقابل حذف‌شده نسبت به تعداد اثرات اصلی حذف‌شده، بیش‌تر است. با این که مقدار مناسب p_2 مشخص نیست و ممکن است این مقدار از مساله‌ای به مساله‌ی دیگر متفاوت باشد، با این حال اگر منابع محاسباتی اجازه رتبه‌دهی همه p اثر اصلی در مرحله یک را بدهد، به احتمال زیاد بتوان p اثر اصلی را با $\frac{1}{2}\sqrt{p}\sqrt{(p+1)}$ اثر متقابل رتبه‌دهی کرد. در نتیجه یک رتبه‌بندی با $p_2 = p + \frac{1}{2}\sqrt{p}\sqrt{(p+1)}$ اثر خواهیم داشت. باید توجه داشت محاسبات مربوط به رتبه‌دهی اثرات متقابل دوتایی خیلی بیش‌تر از محاسبات اثرات اصلی می‌باشد. به طور کلی، وقتی تعداد اثرات را به p_2 کاهش می‌دهیم، هم‌چنان توجه خود را در رتبه‌دهی انجام شده، به اثرات اصلی و متقابل کاملا موثر محدود می‌کنیم. در این مرحله برای رتبه‌دهی اثرات، بعد از محاسبه اثرات متقابل، مشابه مرحله اول اثرات متقابل را به همراه اثرات اصلی رتبه‌بندی می‌کنیم. نتیجه این رتبه‌بندی به صورت زیر قابل نمایش است:

$$\hat{J}_{21} \succeq \hat{J}_{22} \succeq \cdots \succeq \hat{J}_{2p_2}. \qquad (۸.۳)$$

رتبه‌بندی (۸.۳) بر اساس برآورد ضریب همبستگی تعمیم‌یافته (ماکسیمم تابع درستنمایی) بین Y_i و هر یک از اثرات اصلی و متقابل آن‌ها انجام شده است. این رتبه‌بندی میزان مشارکت تک تک متغیرها و اثرات متقابل دوتایی آن‌ها را بر متغیر پاسخ نشان می‌دهد.

لازم به تذکر است که این الگوریتم دومرحله‌ای را می‌توان برای رتبه‌بندی‌های گروهی متغیرهای توضیحی و انتخاب اثرات متقابل بین اثرات اصلی مهم درون گروه‌ها، تعمیم داد. مشکل عمده بر سر این تعمیم، تعریف معیار مناسب برای تشخیص گروه‌ها و اندازه آن‌هاست. برای مشاهده سایر جنبه‌های این الگوریتم به هال و ژو (۲۰۱۴) مراجعه کنید.

۵.۳ ویژگی‌های نظری الگوریتم دومرحله‌ای

فرض کنید متغیر پاسخ Y_i تنها به $r \leq p$ متغیر توضیحی اول بردار $X_i = (X_{i1}, \ldots, X_{ip})$ وابسته است و فقط بین دو متغیر X_{ij_1} و X_{ij_2} اثر متقابل وجود دارد. این وابستگی بین متغیر پاسخ با متغیرهای توضیحی بر حسب یک مدل جمعی با اثرات متقابل به شکل زیر قابل بیان است:

$$Y_i = \sum_{j=1}^{r} g_j(X_{ij}) + \gamma_1(X_{ij_1})\gamma_2(X_{ij_2}) + \epsilon_i, \quad 1 \leq i \leq n \tag{۹.۳}$$

که در آن $r = r(n) \leq p$ که چنان‌چه $n \to \infty$ می‌تواند به بی‌نهایت میل کند. هم‌چنین اگر $n \to \infty$، آن‌گاه $p = p(n) \to \infty$ و $g_1, \ldots, g_r$، γ_1 و γ_2 همگی تابع هستند. برای اثبات این‌که الگوریتم معرفی‌شده با احتمال متمایل به ۱، اثرات با اهمیت را انتخاب می‌کند، نیاز به شرایطی است که ابتدا آن‌ها را مطرح می‌کنیم و سپس قضیه اصلی را بیان و اثبات می‌کنیم.

۱.۵.۳ شرایط نظم

برای درک واضح‌تر مدل (۹.۳)، فرض می‌کنیم

(۱) برای هر ثابت r_0، $1 \leq r_0 \leq r - 2$، به ازای هر n، توابع γ_1, γ_2 و $g_1, \ldots, g_{r_0}$ غیرتباهیده بوده و به n نیز وابسته نیستند.

(۲) $j_1 = r_0 + 1$ و $j_2 = r_0 + 2$

(۳) توابع g_j برای $r_0 + 3 \leq j \leq r$، تابعی از n بوده و $sup\ |g_j| \leq c_1 n^{-a}$، که در آن $a, c_1 > 0$.

(۴) X_{ij} ها و ϵ_i ها همه مستقل‌اند.

(۵) $r = O(n^a)$

پذیره استقلال در قسمت (۴)، مساله را مشکل‌تر می‌کند. زیرا نشان می‌دهد که اهمیت تاثیر متغیرهای اصلی X_{ij_1} و X_{ij_2}، به صورت تکی، که از طریق $\gamma_1(X_{ij_1})\gamma_2(X_{ij_2})$ به‌طور هم‌زمان به‌کار گرفته می‌شوند، قابل تشخیص نیست. روش‌های مبتنی بر همبستگی، چه همبستگی خطی یا فراتر از آن همبستگی تعمیم‌یافته، عبارت $\gamma_1(X_{ij_1})\gamma_2(X_{ij_2})$ را همانند خیلی از روش‌ها یا تکنیک‌های مبتنی بر برازش یک مدل خطی، نادیده می‌گیرند. بنابراین روش پیشنهادی بر رصد کردن X_{ij_1} و X_{ij_2} از طریق تاثیر آن‌ها بر قسمت جمعی مدل (۹.۳) متکی است.

شرط $sup\ |g_j| \leq c_1 n^{-a}$ در قسمت (۳) به این معنی است که این اثرات به سختی قابل شناسایی‌اند و بنابراین مساله انتخاب متغیر را با چالش بیشتری مواجه می‌کنند. با این حال امید است که اثرات کوچک معنی‌دار برای متغیرهای با اندیس‌های j_1 و j_2 رتبه به اندازه کافی بالا در مرحله اول الگوریتم داشته باشند تا جفت اندیس‌های (j_1, j_2) رتبه بالایی در مرحله دوم از آن خود کنند.

بیش از یک اثر متقابل را می‌توان در سمت راست (۹.۳) در نظر گرفت. با این حال، از آن‌جایی که این کار تعمیم ساده‌ای از مدل با یک اثر متقابل است، بنابراین برای سادگی نمادگذاری فقط یک اثر

متقابل در نظر گرفته شد. پذیره‌های (۳) و (۵) با هم این اطمینان را حاصل می‌کنند که سری (۹.۳) به‌طور یکنواخت کراندار است.

با توجه به پذیره‌های (۱) و (۳)، متغیرهای توضیحی با بیشترین تاثیر روی متغیر پاسخ Y_i، به‌صورت جدا از اثرات متقابل به مدل اضافه می‌شوند $(X_{i1}, \ldots, X_{i,r_0+2})$؛ دیگر متغیرهای توضیحی $X_{i,r_0+3}, \ldots, X_{ip}$ به صورت تکی اثرات نسبتا جزئی دارند. برای افزایش چالش مساله انتخاب متغیرهای موثر، اجازه می‌دهیم در توزیع متغیرهای توضیحی X_{ij} و توابع g_j، n به‌طور فزاینده و نامنظمی افزایش یابد. هم‌چنین فضای برداری $\mathcal{H}$ می‌تواند همه چندجمله‌ای‌های با درجه $d \geq 1$ باشد. از آن‌جایی که همبستگی تعمیم‌یافته (۲.۳) و (۳.۳) نسبت به تغییرات مکانی و مقیاسی برای $h \in \mathcal{H}$ ثابت هستند، توجه خود را به $h \in \mathcal{H}_0$ محدود می‌کنیم که $\mathcal{H}_0$ شامل مجموعه همه h هایی است که $E\{h(N)^2\} = 1$، به‌طوری که N دارای توزیع نرمال استاندارد است.

برای عدد صحیح $K \geq 1$، $\mathcal{F}_K$ رده توزیع‌های F با متغیر تصادفی V و $\mathcal{G}_K$ رده توابع g طوری تعریف می‌شوند که $E_F|g(V)|^K$ و $E_F|h(V)|^K$ به‌طور یکنواخت به ازای $F \in \mathcal{F}$، $g \in \mathcal{G}$ و $h \in \mathcal{H}_0$، کراندار باشند. همچنین ثابت $C(K) > 0$ وجود دارد که برای هر $F \in \mathcal{F}_K$، $g \in \mathcal{G}_K$ و هر $k = 1, \ldots, K$ داریم

$$\frac{E_F\big|h(V) - E_F h(V)\big|^{2k}}{\{\mathrm{var}_F h(V)\}^k} \leq C(K)$$

و

$$\frac{E_F\big|g(V) - E_F g(V)\big|^{2k}}{\{\mathrm{var}_F g(V)\}^k} \leq C(K). \qquad (۱۰.۳)$$

در این‌جا E_F و V_F نماد امید ریاضی و واریانس برای متغیر تصادفی V با تابع توزیع F است. برای مثال، (۱۰.۳) برقرار است اگر $\mathcal{F}_K$ رده تمام توزیع‌هایی باشد که تابع چگالی آن‌ها، در یک فاصله فشرده ناتهی $\mathcal{I}$، از بالا در مثلا C_1 کران‌دار است و اگر $\mathcal{G}_K = \mathcal{G}_K(C_2, C_3)$ مجموعه همه ضرایب ثابت توابع پیوسته g باشد که در شرط‌های $|g| \leq C_2$ و $var(g(U)) \geq C_3$ صدق می‌کنند، به‌طوری که $C_1, C_2, C_3 > 0$ و U روی بازه $\mathcal{I}$ دارای توزیع یکنواخت است.

فرض کنید V_j با متغیر توضیحی X_{ij} هم‌توزیع باشد. هم‌چنین فرض کنید نماد $a_n \asymp b_n$ به این معنی است که $\frac{a_n}{b_n}$، وقتی اعداد مثبت a_n و b_n به بی‌نهایت میل کنند و p و p' توابعی بر حسب n باشند، بین صفر و بی‌نهایت قرار می‌گیرد. در این صورت پذیره‌های زیر را در نظر می‌گیریم:

۱. مرحله دوم الگوریتم، همه جفت‌های متمایز را با p' متغیر اصلی برتر رتبه‌بندی‌شده در مرحله یک تجمیع می‌کند که $p' \asymp p$ و سپس $p_2 = p + \frac{1}{2}p'(p'-1)$ متغیر حاصل را رتبه‌بندی می‌کند.

۲. هر توزیعی برای V_j در رده $\mathcal{F}_K$ است و هر تابع g_j و γ_j در $\mathcal{G}_K$ هستند.

۳. برای هر $j = 1, \ldots, r_0$ داریم

$$sup_{h \in \mathcal{H}_0} corr\{g_j(V_j), h(V_j)\} \geq C_4.$$

هم‌چنین

$$sup_{h\in\mathcal{H}_0} corr\{\gamma_1(V_{j_1})\gamma_2(V_{j_2}), h(V_{j_1}V_{j_2})\} \geq C_4.$$

برای حداقل $p - C_5 p'$ مقدار از $j \in \{1, \ldots, p\}$

$$sup_{h\in H_0} corr\{g_j(V_j), h(V_j)\} \leq n^{-a_2}$$

و برای $j = r_0 + 1, r_0 + 2$، که $a \leq a_1 < a_2 < \frac{1}{2}$ و a در پذیره قسمت (۳) صدق می‌کند،

$$sup_{h\in\mathcal{H}_0} corr\{g_j(V_j), h(V_j)\} \geq n^{-a_1}.$$

در ضمن $C_4 > 0$ و $0 < C_5 < 1$.

۴. $p = O(n^{C_6})$

۵. $E|\epsilon|^{C_7} < \infty$ و هر دوی C_7 و K در تعاریف $\mathcal{F}_K$ و $\mathcal{G}_K$، بسته به انتخاب C_6، به اندازه کافی بزرگ انتخاب می‌شوند.

شرط ۱ زمانی مناسب است که منابع محاسباتی فرد را به انجام محاسبات با مرتبه $O(p)$ محدود کند. شرط ۳ به ما اطمینان می‌دهد که متغیرهایی که با تصمیم حاصل از ۱ حذف شده‌اند، و تنها $O(\sqrt{p})$ متغیر برتر حفظ شده‌اند، با احتمال بالا، متغیرهای کلیدی که شامل اثرات متقابل در عبارت (۹.۳) هستند را خارج نمی‌کنند. در شرط ۴ این اطمینان حاصل می‌شود که p بیشتر از یک تابع چندجمله‌ای از n نیست. شرط ۳ مهم‌ترین شرط این بخش است. زیرا تضمین می‌کند که، در اغلب موارد، متغیرهای با اندیس‌های $r_0 + 1$ و $r_0 + 2$ از مرحله یک به مرحله دو رتبه‌های بالایی می‌گیرند.

با توجه به شرایط نظم بیان‌شده، می‌توان قضیه زیر را مطرح کرد که نشان می‌دهد با انتخاب $\sqrt{p}$ اثر اصلی اول مرحله یک الگوریتم رتبه‌بندی برای ساخت اثرات متقابل دوتایی و رتبه‌بندی توام اثرات اصلی به‌همراه اثرات متقابل دوتایی با احتمال ۱، اثرات اصلی و متقابل دوتایی موثر را حفظ کرده و مدل واقعی را شامل می‌شود. قضیه را برای متغیر پاسخ پیوسته بیان و اثبات می‌کنیم اما می‌توان نتایج مشابهی را برای متغیر پاسخ دوسطحی با استفاده از تحلیل‌های همبستگی مبتنی بر رگرسیون لجستیک، به دست آورد.

۲.۵.۳ نتیجه نظری اصلی

قضیه ۱.۵.۳. فرض کنید مدل (۹.۳) در شرایط نظم بیان‌شده صدق کند و در تمام مراحل رتبه‌بندی، از الگوریتم رتبه‌دهی بر مبنای همبستگی تعمیم‌یافته استفاده شود. در این صورت با احتمال نزدیک به ۱، وقتی $n \to \infty$،

- در رتبه‌بندی مرحله اول الگوریتم رتبه‌دهی بر مبنای همبستگی تعمیم‌یافته، دنباله $\hat{J}_{11}, \ldots, \hat{J}_{1,r_0}$ یک جایگشت از $1, \ldots, r_0$ است و $r_0 + 1$ و $r_0 + 2$ در بین $\hat{J}_{1,r_0+1}, \ldots, \hat{J}_{1p'}$ قرار دارند.
- در رتبه‌بندی مرحله دوم الگوریتم رتبه‌دهی بر مبنای همبستگی تعمیم‌یافته، دنباله $\hat{J}_{21}, \ldots, \hat{J}_{2,r_0+1}$ یک جایگشت از $1, \ldots, r_0$ و هر اندیس از دو متغیری است که شامل j_1 و j_2 باشد.

برهان. برهان قضیه در هال و ژو (۲۰۱۴) آمده است. □

۶.۳ مطالعه شبیه‌سازی

در بخش قبل از دیدگاه نظری اعتبار روش پیشنهادی برای انتخاب اثرات اصلی و متقابل ارائه شد. در این بخش، قصد داریم این اعتبار را با دو مثال شبیه‌سازی، بررسی کنیم. برای ارزیابی اعتبار روش انتخاب متغیر دومرحله‌ای پیشنهادی، از داده‌های شبیه‌سازی‌شده بر اساس دو مدل رده‌بندی استفاده می‌کنیم. دو مدل رده‌بندی زیر را در نظر بگیرید:

$$\log \frac{p(Y_i = ۱|X_i)}{p(Y_i = ۰|X_i)} = \beta_۰ + \sum_{j=۱}^{۳} \frac{۴-j}{۳}(X_{ij} + X_{i,j+۳} + X_{ij}X_{i,j+۳}) \qquad (۱۱.۳)$$

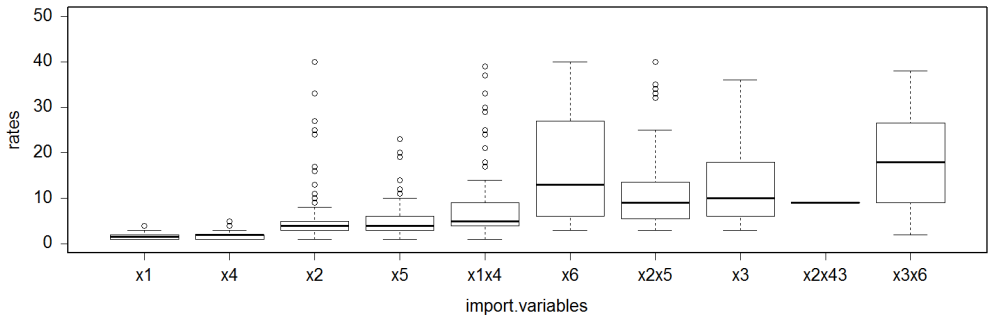
و

$$\log \frac{P(Y_i = ۱|X_i)}{P(Y_i = ۰|X_i)} = \beta_۰ + \sum_{j=۱}^{۳} \frac{۴-j}{۳}\{X_{ij} + X_{i,j+۳} + \sin(X_{ij})\exp(X_{i,j+۳})\}. \qquad (۱۲.۳)$$

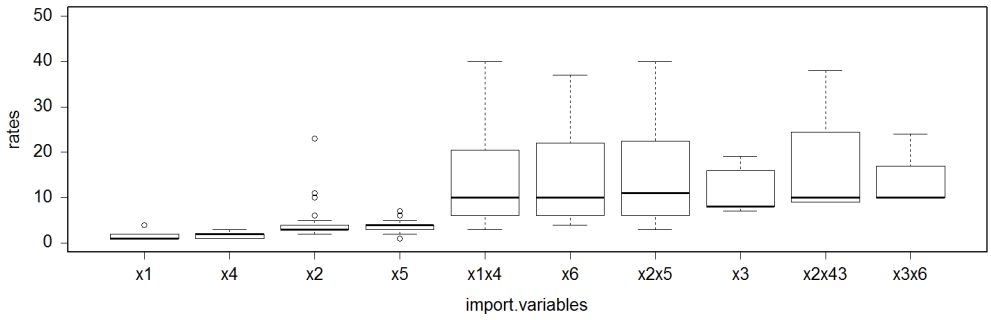
در این دو مدل، متغیر پاسخ $Y_i \in \{۰, ۱\}$ و دارای توزیع برنولی با احتمال موفقیت $P(Y_i = ۱|X_i)$ است. برای $۱ \le j \le ۳$، هر جفت از متغیرهای $(X_{ij}, X_{ij+۳})$ با همبستگی ۰٫۸۵ را از توزیع نرمال ۶ متغیره با ماتریس کواریانس

$$\Sigma = \begin{pmatrix} ۱ & ۰ & ۰ & ۰٫۸۵ & ۰ & ۰ \\ ۰ & ۱ & ۰ & ۰ & ۰٫۸۵ & ۰ \\ ۰ & ۰ & ۱ & ۰ & ۰ & ۰٫۸۵ \\ ۰٫۸۵ & ۰ & ۰ & ۱ & ۰ & ۰ \\ ۰ & ۰٫۸۵ & ۰ & ۰ & ۱ & ۰ \\ ۰ & ۰ & ۰٫۸۵ & ۰ & ۰ & ۱ \end{pmatrix}$$

تولید کردیم. سایر متغیرهای توضیحی $X_{i۷}, \ldots, X_{ip}$ را، برای $p = ۱۰۰۰$، از توزیع نرمال استاندارد تولید کردیم. عرض از مبدا را نیز برای هر دو مدل برابر $\beta_۰ = -۲٫۵$ انتخاب کردیم. برای ارزیابی عملکرد الگوریتم دومرحله‌ای انتخاب متغیر، در توانایی شناسایی اثرات اصلی و متقابل غیر صفر واقعی، از هر دو مدل رده‌بندی (۱۱.۳) و (۱۲.۳) ۱۰۰ مجموعه داده به حجم ۲۰۰ تولید کردیم. سپس با استفاده از روش انتخاب متغیر پیشنهادی برای هر ۱۰۰ مجموعه داده تولیدشده، اثرات اصلی و متقابل را رتبه‌بندی کرده و میانه رتبه‌های هر یک از متغیرهای توضیحی به‌همراه اثرات متقابل آن‌ها را محاسبه کردیم. نتایج در دو شکل ۲.۳ و ۳.۳ گزارش شده‌اند. هر دو شکل نشان می‌دهد که روش انتخاب متغیر مذکور به خوبی و با دقت بالا، اثرات غیر صفر واقعی را در اغلب موارد شناسایی و انتخاب کرده است. حتی اثرات متقابل معنی‌دار به خوبی توسط الگوریتم دیده شده‌اند. این عملکرد در حضور ۱۰۰۰ متغیر توضیحی اصلی و اثرات متقابل حاصل از اثرات اصلی منتخب در مرحله اول، قابل تقدیر است.



شکل ۲.۳: میانه رتبه‌های ۱۰ اثر اول در مدل (۱۱.۳) توسط روش پیشنهادی برای رده‌بندی متغیر پاسخ به ازای ۱۰۰ تکرار



شکل ۳.۳: میانه رتبه‌های ۱۰ اثر اول در مدل (۱۲.۳) توسط روش پیشنهادی برای رده‌بندی متغیر پاسخ به ازای ۱۰۰ تکرار

فصل ۴

ارزیابی عملکرد روش مبتنی بر همبستگی تعمیم‌یافته: مطالعات موردی

در مدل‌بندی بسیاری از مجموعه داده‌های واقعی، مانند داده‌های ریزآرایه، متغیر پاسخ Y اغلب از نوع کیفی بوده و به صورت صفر و یک می‌باشد و تعداد مولفه‌های بردار متغیرهای توضیحی $X_1, \ldots, X_p$، از مشاهدات موجود بزرگ‌تر بوده و داده‌ها با بعد بالا محسوب می‌شوند. برای مثال، فن و لیو (۲۰۰۸) برای رده‌بندی ژن‌های متفاوت، از همبستگی معمولی برای انتخاب اثرات اصلی، به کمک روش SIS، استفاده کردند. هم‌چنین هال و مایلر (۲۰۰۹) برای رده‌بندی این نوع داده‌ها از همبستگی تعمیم‌یافته، بر اساس مدل رگرسیون لجستیک، کمک گرفتند.

در این فصل، برای ارزیابی عملکرد روش‌های پیشنهادی در فصل سوم، و سنجش تاثیر ورود اثرات متقابل در مسایل رده‌بندی در مقایسه با استفاده از فقط اثرات اصلی، از سه مجموعه داده واقعی با بعدهای متوسط (دو مجموعه) و (خیلی) بالا (یک مجموعه) استفاده کرده‌ایم. این سه مجموعه داده عبارتند از

- مجموعه داده سرطان خون با نام Leukemia که یک مجموعه داده ریزآرایه است با ۷۱۲۹ ژن به عنوان متغیرهای توضیحی که برای ۷۲ فرد اندازه‌گیری شده‌اند. این داده‌ها توسط گالوب و همکاران (۱۹۹۹) گردآوری شده‌اند.
- مجموعه داده تصویربرداری قلبی با نام Heart که این مجموعه در پایگاه داده گروه یادگیری ماشین دانشگاه کالیفرنیا اروین[۱] (UCI) موجود و قابل دسترس است. این داده‌ها شامل ۲۶۷ تصویر ($n = ۲۶۷$) از قلب است که ۵۵ مورد آن‌ها طبیعی و ۲۱۲ تصویر دیگر غیر طبیعی هستند. برای هر تصویر نیز ۴۴ مشخصه ($p = ۴۴$) ثبت شده‌اند (کورگان و همکاران، ۲۰۰۱).
- مجموعه داده ساختار یونی اتمسفر با نام Ionoshpere که این مجموعه داده نیز در پایگاه داده گروه یادگیری ماشین UCI موجود و قابل دسترس است. این داده‌ها شامل ۳۱۵ بازتاب رادار

[۱] University of California Irvine

($n = ۳۱۵$) برای شناسایی شواهدی از ساختار یونی اتمسفر هستند که ۲۲۵ بازتاب خوب و ۱۲۶ بازتاب بد ثبت شده‌اند. برای هر کدام از این بازتاب‌ها نیز ۳۳ مشخصه ($p = ۳۳$) به عنوان متغیرهای توضیحی اندازه‌گیری شده‌اند (سیگیلیتو و همکاران، ۱۹۸۹).

از این مجموعه داده‌ها برای بررسی دقت رده‌بندی پس از انتخاب اثرات مهم استفاده کرده‌ایم. برای این منظور، ابتدا با روش‌های انتخاب متغیر مطرح‌شده فرآیند انتخاب اثرات موثر مهم را انجام داده و سپس با مجموعه متغیرهای توضیحی نتیجه‌شده، به رده‌بندی یک مجموعه داده آزمون در نظر گرفته‌شده با روش‌های k-نزدیک‌ترین همسایگی و مرکز-مبنا پرداخته‌ایم. برای مطالعه نتایج سایر روش‌های رده‌بندی از جمله رگرسیون لجستیک، تحلیل ممیزی خطی و تحلیلی ممیزی درجه دوم با مجموعه متغیرهای توضیحی کاهش داده شده با روش SIS می‌توان به محمدی و همکاران (۱۳۹۴) مراجعه کرد. هم‌چنین اگر مجموعه داده مورد مطالعه بیش از دو رده داشته باشد مانع از استفاده روش پیشنهادی نمی‌شود، کافی است تابع درست‌نمایی را براساس توزیع چند جمله‌ای متغیر پاسخ تشکیل دهیم. نتایج به‌دست آمده نشان می‌دهند که رهیافت انتخاب اثرات موثر اصلی و متقابل دوتایی، می‌تواند رده‌بندی دقیق‌تری نسبت به روش SIS، بر اساس مدل رگرسیون لجستیک و انتخاب فقط اثرات اصلی، ارائه دهد.

۱.۴ داده‌های سرطان خون

مجموعه داده Leukemia با ۷۲ مشاهده و ۷۱۲۹ ژن، به عنوان متغیرهای توضیحی، به صورت دو مجموعه آزمون و آموزشی توسط گالوب و همکاران (۱۹۹۹) جمع‌آوری شده است. سرطان خون، گروهی از انواع سرطان می‌باشد که با نام‌های مختلفی مانند چنگار خون، لوسمی و لوکمیا از آن‌ها یاد می‌شود. این نوع سرطان معمولا از مغز استخوان شروع می‌شود و باعث شکل‌گیری تعداد زیادی گلبول سفید غیر طبیعی می‌گردد. این گلبول‌های سفید خونی به صورت کامل تشکیل نشده‌اند و به آن‌ها بلاست یا سلول‌های لوکمی یا سرطان خون گفته می‌شود. از نشانه‌های این بیماری می‌توان به مشکلات خون‌ریزی، کبود شدگی، تب و افزایش عفونت اشاره کرد. این نشانه‌ها به دلیل نبود سلول‌های خونی نرمال ایجاد می‌شوند. تشخیص این بیماری با استفاده از آزمایش خون و بیوپسی مغز استخوان[۱] صورت می‌گیرد (باین، ۲۰۰۳). در حالی که عوامل اصلی و واقعی بیماری سرطان خون ناشناخته‌اند، دانشمندان و پزشکان بر این باورند که ترکیبی از عوامل ژنتیکی و محیطی در پیدایش این بیماری نقش دارند. چهار نوع اصلی سرطان خون عبارتند از

۱. لوسمی حاد لنفوئیدی[۲] (ALL)

۲. لوسمی حاد میلوئیدی[۳] (AML)

۳. لوسمی مزمن لنفوئیدی[۴] (CLL)

[۱] Bone marrow biopsy

[۲] Acute Lymphoblastic Leukemia

[۳] Acute Myelogenous Leukemia

[۴] Chronic Lymphocytic Leukaemia

۴. لوسمی مزمن میلوئیدی[۱] (CML)

برای مطالعه بیش‌تر در زمینه داده‌های بیان ژنی و ریزآرایه، مانند این داده‌ها، می‌توان به هانگ و پن (۲۰۰۳)، استوری و تیب‌شیرانی (۲۰۰۳)، سگال و همکاران (۲۰۰۳) و سینگ و همکاران (۲۰۰۲) مراجعه کنید.

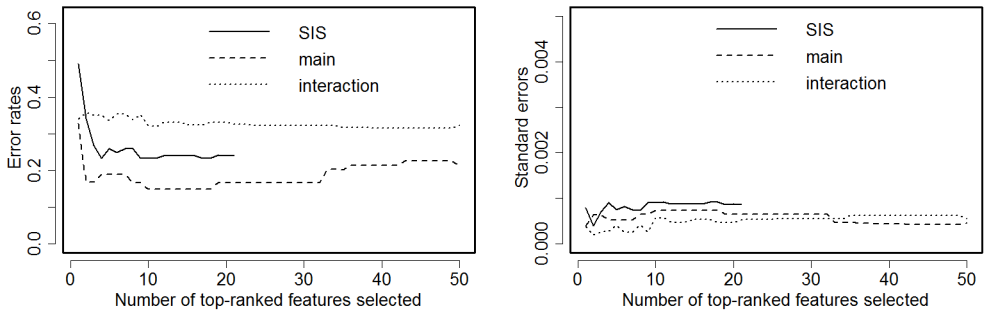
برای تحلیل این داده‌ها، ابتدا بعد آن‌ها را با دو روش انتخاب متغیر SIS و ضریب همبستگی معمولی کاهش داده، سپس از مجموعه متغیرهای توضیحی حاصل برای رده‌بندی متغیر پاسخ دوسطحی، مبتنی بر دو روش رده‌بندی مرکز-مبنا و k-نزدیکترین همسایگی، در داده‌های آزمون استفاده می‌کنیم. پاسخ در این مثال وجود یا عدم وجود سرطان خون در افراد تحت مطالعه است. مرحله آخر نیز مقایسه نتایج رده‌بندی‌های به‌دست آمده از دو روش انتخاب متغیر است. در این مجموعه داده، فقط ۷۱۲۹ ژن بررسی‌شده در ۷۲ مورد را در دو گروه ALL و AML، بررسی می‌کنیم. به منظور ارزیابی عملکرد انتخاب متغیر و رده‌بندی، از روش خودگردان‌سازی استاندارد (نمونه‌گیری تصادفی با جایگذاری) برای داده‌های آموزشی استفاده می‌کنیم تا مجموعه داده‌های آموزشی جدید، با حجم مشابه مجموعه داده آموزشی اصلی، بسازیم. این عمل به این خاطر انجام می‌شود تا بتوان با ترکیب نتایج حاصل از چندین داده آموزشی (برابر تعداد نمونه‌های خودگردان)، استنباط قابل اعتمادی را گزارش کنیم.

برای هر مجموعه داده آموزشی تولیدشده با روش خودگردان‌سازی، فرآیند انتخاب متغیر و سپس رده‌بندی انجام می‌شود. سپس با مجموعه داده آزمون از پیش معین‌شده، که بخشی از نمونه‌های خودگردان تولیدشده را شامل می‌شوند، خطای رده‌بندی و انحراف استاندارد آن را به کمک آماره‌های خودگردان محاسبه می‌کنیم. تعداد نمونه خودگردان در این مثال برابر ۲۰۰ در نظر گرفته شده است. در این ۲۰۰ تکرار، از دو روش ذکرشده برای به‌دست آوردن ۵۰ متغیر مهم اول استفاده می‌کنیم و با آن‌ها دو روش رده‌بندی را اجرا می‌کنیم. در روش پیشنهادی مبتنی بر ضریب همبستگی معمولی، همه

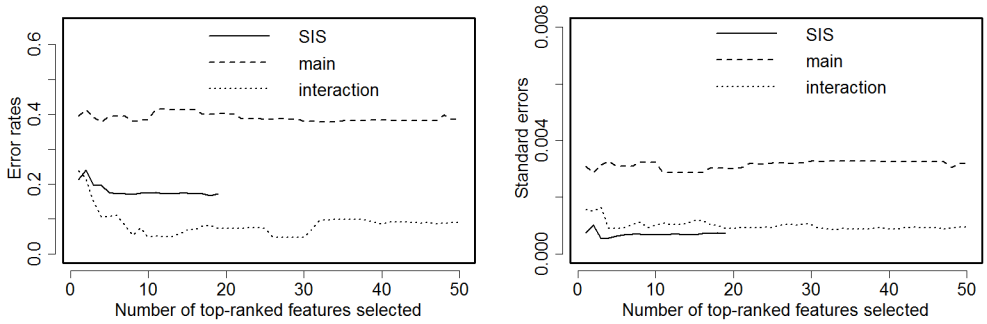
$$p + \frac{۱}{۲}\sqrt{p}(\sqrt{p} - ۱)$$

اثرات اصلی و متقابل حضور دارند. نتایج خطای رده‌بندی با روش k-NN در شکل‌های ۱.۴ و ۲.۴ نمایش داده شده‌اند. شکل ۱.۴ مربوط به حالتی است که $k = ۱$ در نظر گرفته شده است. در این حالت روش پیشنهادی خطای رده‌بندی کمتری نسبت به روش SIS دارد، اما حضور اثرات متقابل (interaction) برای رده‌بندی نتیجه را بدتر کرده است. از طرفی افزایش تعداد اثرات برای رده‌بندی با روش SIS و اثرات متقابل، کاهش نرخ خطای رده‌بندی را در پی داشته است، اما در روش پیشنهادی با اثرات اصلی (main) افزودن اثرات بیشتر از ۳۰، نرخ خطا را افزایش داده است. با توجه به نمودارهای سمت راست شکل ۱.۴ می‌توان فهمید که دقت خطای رده‌بندی 1-NN در هر سه روش کاهش بعد، تقریبا مشابه است.

[۱] Chronic Myelogenous Leukemia

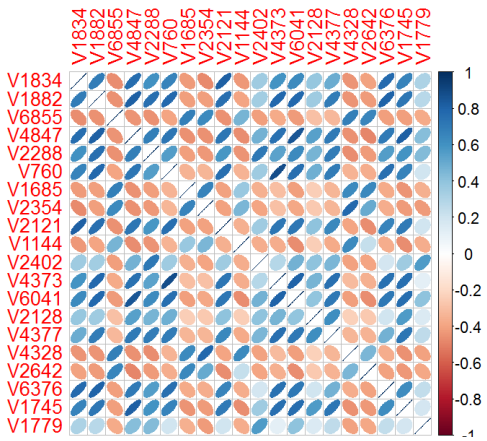


شکل ۱.۴: نمودارهای نرخ خطای رده‌بندی 1-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با سه روش کاهش بعد برای مجموعه داده سرطان خون



شکل ۲.۴: نمودارهای نرخ خطای رده‌بندی 10-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با سه روش کاهش بعد برای مجموعه داده سرطان خون

برخلاف انتظاری که در مورد کاهش خطای رده‌بندی در حضور اثرات متقابل داشتیم، استفاده از فقط اثرات اصلی دارای کم‌ترین نرخ خطای رده‌بندی با روش 1-NN است. دلیل آن هم می‌تواند به خاطر استفاده از تنها یک نقطه نزدیک برای رده‌بندی در روش 1-NN باشد. به همین دلیل، تمام مراحل تحلیل را با روش رده‌بندی 10-NN اجرا کردیم. نتیجه در شکل ۲.۴ گزارش شده است که نشان می‌دهد روش انتخاب متغیر با اثرات متقابل در این حالت دارای کم‌ترین نرخ خطای رده‌بندی است. هم‌چنین دقت خطای رده‌بندی با استفاده از متغیرهای مهم حاصل از روش SIS، مشابه رهیافت دومرحله‌ای با اثرات متقابل است. با این رده‌بندی، روش پیشنهادی با تنها اثرات اصلی عملکرد رضایت‌بخشی ندارد. این مثالی از اهمیت در نظر گرفتن اثرات متقابل برای هدف رده‌بندی است. شکل ۳.۴ نتیجه همبستگی خطی بین ۲۰ اثر موثر اول توسط روش پیشنهادی دو مرحله‌ای در مجموعه داده سرطان خون می‌باشد که نشان دهنده مساله همبستگی خطی بین متغیرهای توضیحی است.



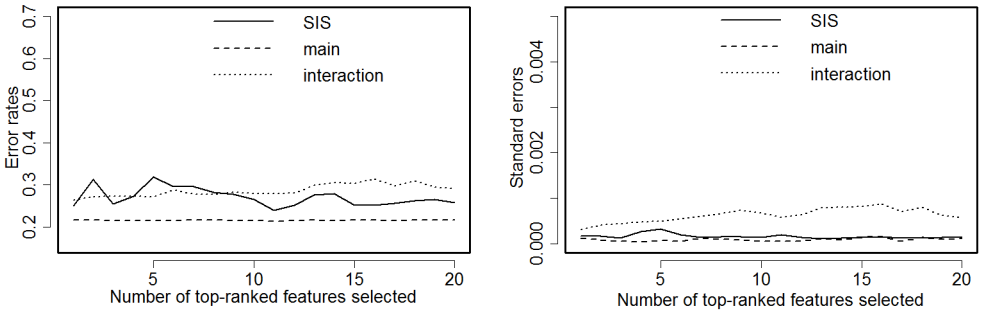
شکل ۳.۴: نمودار همبستگی خطی بین ۲۰ اثر موثر اول در مجموعه داده سرطان خون

۱.۱.۴ داده‌های تصاویر قلبی

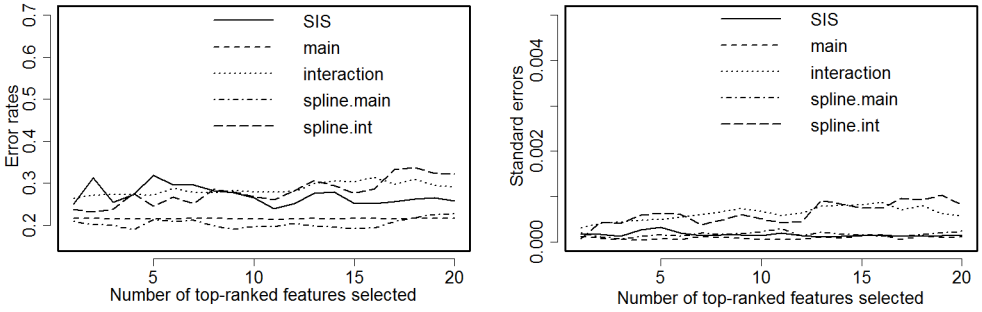
مجموعه داده تصاویر قلبی (Heart) یک مجموعه داده کلاسیک است که به دفعات برای ارزیابی روش‌های رده‌بندی از آن استفاده شده است. این مجموعه داده در پایگاه داده یادگیری ماشین دانشگاه کالفرنیا اروین موجود است و از طریق وب‌گاه www.archive.ics.uci.edu/ml قابل دستیابی است. این مجموعه داده شامل ۲۶۷ تصویر از قلب است که ۵۵ مورد آن‌ها تصاویر طبیعی و ۲۱۲ مورد دیگر غیر طبیعی هستند. در هر تصویر نیز ۴۴ مشخصه به عنوان متغیرهای توضیحی اندازه‌گیری شده‌اند. این مجموعه داده یک مجموعه با بعد متوسط تلقی می‌شود. مجموعه داده Heart از همان ابتدا به صورت دو مجموعه آموزشی و آزمون جمع‌آوری شده است. برای مجموعه آموزشی، ۴۰۰ نمونه خودگردان تولید کردیم و روش‌های انتخاب متغیر و رده‌بندی k-NN را به کار گرفتیم. در روش پیشنهادی دومرحله‌ای، ابتدا از ضریب همبستگی معمولی استفاده کردیم. نتیجه در شکل ۴.۴ نمایش داده شده است. با توجه به نتایج به‌دست آمده، می‌توان گفت اثرات متقابل دوتایی منتخب در روش پیشنهادی مبتنی بر همبستگی معمولی، تاثیر بیش‌تری روی متغیر پاسخ نسبت به برخی از اثرات اصلی دارد که می‌تواند در برخی موارد، مثل پزشکی، ناشی از یک پزشک مشخص باشد.

برای این داده‌ها نیز افزایش تعداد متغیرهای با اهمیت، موجب کاهش نرخ خطای رده‌بندی در روش SIS شده است، در حالی‌که برای روش مبتنی بر همبستگی معمولی تغییر محسوسی در نرخ خطا دیده نمی‌شود. اما به‌طور کلی، روش انتخاب متغیر با اثرات متقابل دارای کم‌ترین نرخ خطای رده‌بندی است و اثرات متقابل باعث بهبود قابل توجهی در کاهش نرخ خطای رده‌بندی 1-NN شده‌اند. دقت خطای رده‌بندی 1-NN نیز در هر سه روش کاهش بعد، مشابه است.

برای ارزیابی عملکرد رهیافت ضریب همبستگی تعمیم‌یافته، تحلیل‌های بالا را با نتایج رده‌بندی بر اساس کاهش بعد داده‌ها مبتنی بر روش ضریب همبستگی تعمیم‌یافته، هم با اثرات اصلی (spline.main) و هم اثرات اصلی و متقابل (spline.int)، تکرار کردیم. شکل ۵.۴ نتیجه این تحلیل را نشان می‌دهد. مشهود است که هر سه روش انتخاب متغیر spline.int، interaction و main دارای خطای رده‌بندی مشابه هستند، اما نسبت به روش انتخاب متغیر SIS، به‌طور قابل توجهی دارای خطای رده‌بندی کمتری

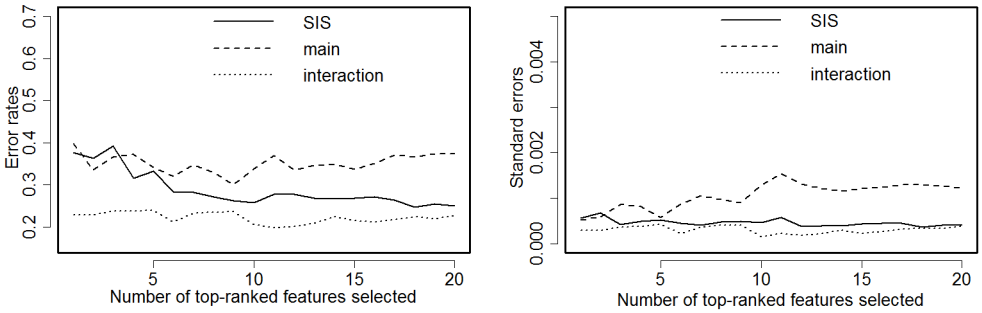


شکل ۴.۴: نمودارهای نرخ خطای رده‌بندی 1-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با سه روش کاهش بعد برای مجموعه داده تصاویر قلبی



شکل ۵.۴: نمودارهای نرخ خطای رده‌بندی 1-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با پنج روش کاهش بعد برای مجموعه داده تصاویر قلبی

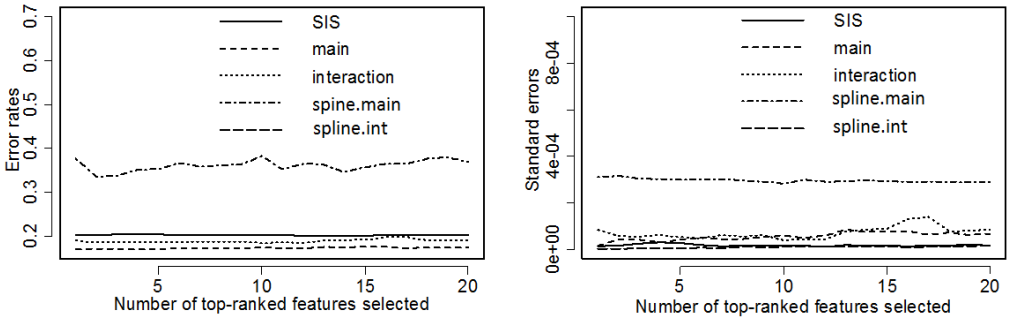
هستند. هم‌چنین دقت خطای رده‌بندی 1-NN در هر پنج روش کاهش بعد به‌کار رفته، مشابه است.



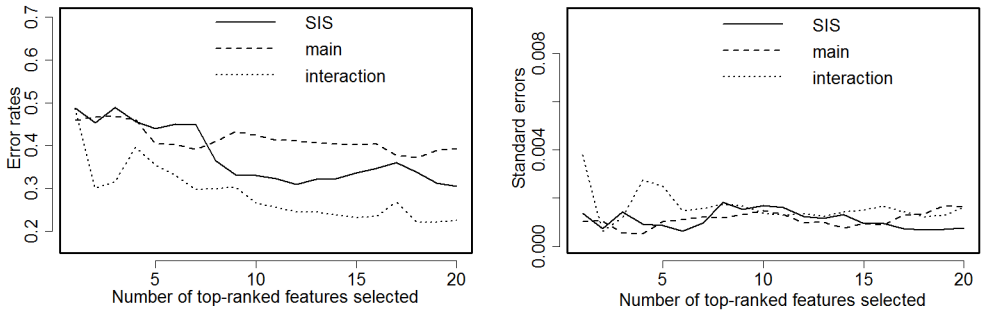
شکل ۶.۴: نمودارهای نرخ خطای رده‌بندی 10-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با سه روش کاهش بعد برای مجموعه داده تصاویر قلبی

دو شکل ۶.۴ و ۷.۴، تحلیل‌های مشابه دو شکل ۴.۴ و ۵.۴ را برای روش 10-NN نشان می‌دهند.

در مطابقت با شکل ۴.۴، برای حالتی که از ضریب همبستگی معمولی استفاده می‌کنیم، می‌توان گفت که روش پیشنهادی دومرحله‌ای از عملکرد خیلی بهتری نسبت به روش SIS برخوردار است. در تمام موارد نیز دقت روش رده‌بندی مبتنی بر تمام روش‌های کاهش بعد، تقریبا مشابه است. نتیجه مشابه نیز، به‌جز برای اثرات اصلی، در روش کاهش بعد مبتنی بر ضریب همبستگی تعمیم‌یافته (شکل ۷.۴) اتفاق افتاده است.



شکل ۷.۴: نمودارهای نرخ خطای رده‌بندی 10-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با پنج روش کاهش بعد برای مجموعه داده تصاویر قلبی

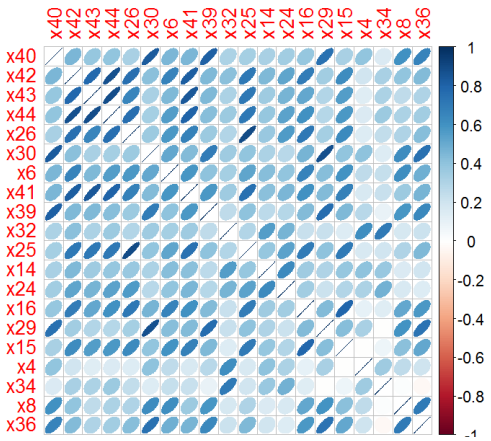


شکل ۸.۴: نمودارهای نرخ خطای رده‌بندی روش مرکز-مبنا (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با سه روش کاهش بعد برای مجموعه داده تصاویر قلبی

برای ارزیابی عملکرد روش‌های مذکور در ترکیب با روش رده‌بندی مرکز-مبنا، داده‌های تصاویر قلبی رو با شیوه مشابه بالا ولی بر اساس روش مرکز-مبنا تحلیل کردیم. در الگوریتم دومرحله‌ای پیشنهادی، به دلیل محدودیت‌های محاسباتی روش مرکز-مبنا، فقط از ضریب همبستگی معمولی استفاده کردیم. نتیجه در شکل ۸.۴ نمایش داده شده است. با توجه به نمودارهای این شکل، واضح است که افزایش تعداد متغیرهای با اهمیت موجب کاهش نرخ خطای رده‌بندی، در هر سه روش کاهش بعد، می‌شود. اما به‌طور کلی، روش انتخاب متغیر با اثرات متقابل دارای کم‌ترین نرخ خطای رده‌بندی است و اثرات

متقابل باعث بهبود قابل توجهی در کاهش نرخ خطای رده‌بندی مرکز-مبنا شده‌اند. دقت خطای رده‌بندی نیز برای هر سه روش کاهش بعد، تقریبا یکسان است.

با تجمیع نتایج حاصل از تحلیل این داده‌ها، می‌توان نتیجه گرفت استفاده از اثرات متقابل در کنار اثرات اصلی مهم و موثر، نقش کاملا تاثیرگذاری بر کاهش نرخ خطای رده‌بندی در روش‌های مختلف رده‌بندی دارد. این تاثیر در استفاده از رهیافت ضریب همبستگی تعمیم‌یافته می‌تواند تقویت شود. شکل ۹.۴ نتیجه همبستگی خطی بین ۲۰ اثر موثر اول توسط روش پیشنهادی دو مرحله‌ای در مجموعه داده تصویر برداری قلبی می‌باشد که نشان دهنده مساله همبستگی خطی بین متغیرهای توضیحی است.

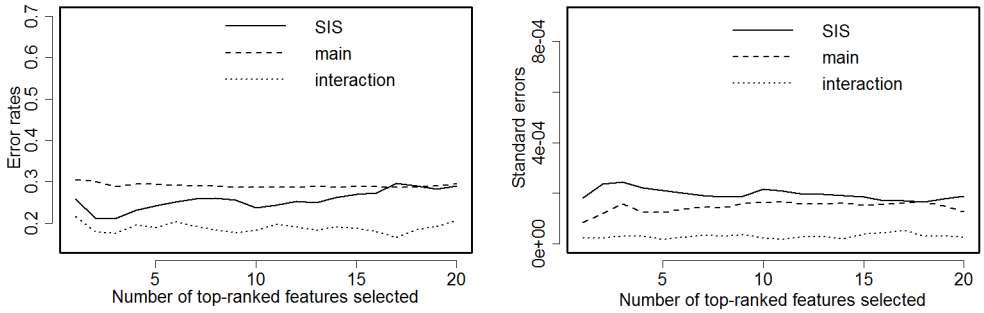


شکل ۹.۴: نمودار همبستگی خطی بین ۲۰ اثر موثر اول در مجموعه داده تصویر برداری قلبی

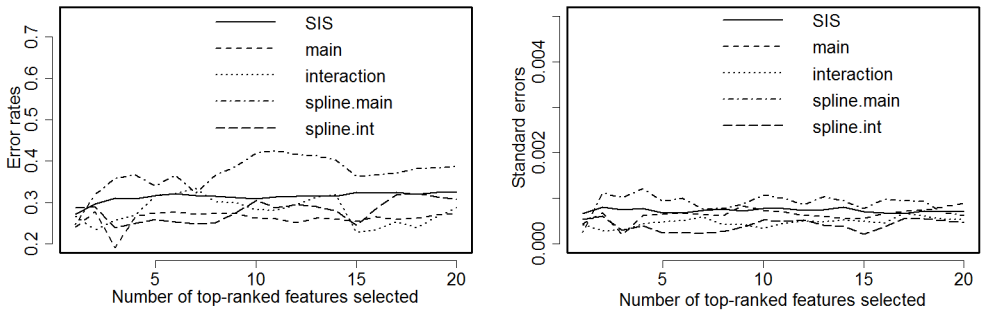
۲.۴ داده‌های ساختار یونی اتمسفر

مجموعه داده Ionosphere شامل ۳۵۱ بازتاب رادار برای شناسایی شواهدی از ساختار یونی اتمسفر جو زمین می‌باشد که در آن ۲۲۵ بازتاب خوب و ۱۲۶ بازتاب بد حضور دارند. بنابراین متغیر پاسخ دوسطحی است. برای هر یک از این بازتاب‌ها ۳۳ مشخصه به عنوان متغیر توضیحی ثبت شده‌اند. برای انجام اعتبارسنجی متقابل، مشابه مثال‌های قبلی، این مجموعه داده را به‌طور تصادفی به دو مجموعه آموزشی و آزمون به‌ترتیب با نسبت‌های دو سوم و یک سوم، تقسیم کردیم.

مشابه داده‌های Heart، برای این داده‌ها از روش‌های رده‌بندی 1-NN، 10-NN و مرکز-مبنا برای رده‌بندی پاسخ در داده‌های آزمون استفاده کرده‌ایم. تعداد نمونه‌های خودگردان نیز برابر ۴۰۰ در نظر گرفته شده است. نتایج در شکل‌های ۱۰.۴ تا ۱۴.۴ گزارش شده‌اند. از شکل ۱۰.۴ مشهود است که افزایش تعداد متغیرهای با اهمیت، نوسان مشخصی را در نرخ خطای رده‌بندی 1-NN، در روش‌های کاهش بعد مشخص‌شده در شکل (SIS و مبتنی بر ضریب همبستگی معمولی) نشان نمی‌دهد. اما واضح است که حضور اثرات متقابل نسبت به دو روش دیگر، همواره کاهش نرخ خطای رده‌بندی بیشتری را (به ازای هر تعداد متغیر) دارد. دقت خطای رده‌بندی 1-NN نیز در هر سه روش کاهش بعد، مشابه است.



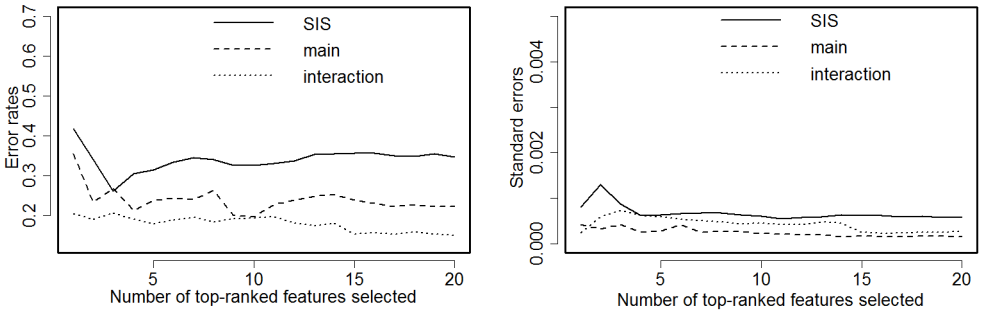
شکل ۱۰.۴: نمودارهای نرخ خطای رده‌بندی 1-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با سه روش کاهش بعد برای مجموعه داده ساختار یونی اتمسفر



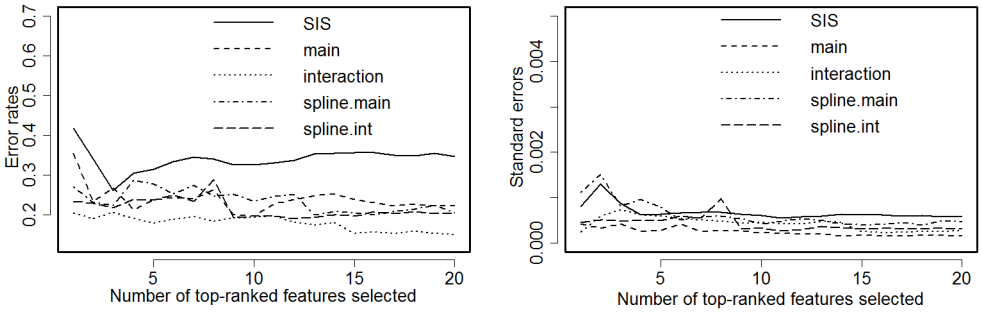
شکل ۱۱.۴: نمودارهای نرخ خطای رده‌بندی 1-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با پنج روش کاهش بعد برای مجموعه داده ساختار یونی اتمسفر

تاثیر رهیافت ضریب همبستگی تعمیم‌یافته در روش پیشنهادی را، برای این داده‌ها، در شکل ۱۱.۴ مشاهده می‌کنید. با توجه به این شکل، هر دو روش انتخاب متغیر مبتنی بر ضریب همبستگی معمولی و تعمیم‌یافته با اثرات متقابل، نسبت به سایر روش‌های کاهش بعد، از خطای رده‌بندی 1-NN کم‌تری برخوردار هستند. میزان کاهش خطا نیز قابل توجه است. این در حالی است که دقت خطای رده‌بندی 1-NN مبتنی بر کاهش بعد توسط همه روش‌ها، مشابه است. بنابراین در این مثال نیز اهمیت حضور اثرات متقابل، در افزایش دقت رده‌بندی که برای این داده‌ها نیز از اهمیت بالایی برخوردار است، کاملا قابل درک است.

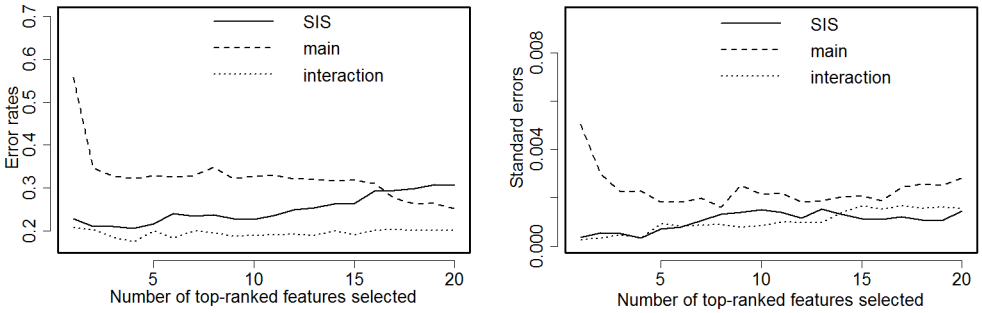
مشابه نتایج بالا، دو شکل ۱۲.۴ و ۱۳.۴، حاصل رده‌بندی بر اساس متغیرهای منتخب روش‌های کاهش بعد را با روش 10-NN نشان می‌دهند. شکل ۱۲.۴ مربوط به روش‌های SIS و ضریب همبستگی معمولی است، در حالی‌که در شکل ۱۳.۴ روش‌های مبتنی بر ضریب همبستگی تعمیم‌یافته نیز اضافه شده‌اند. نتایج این دو شکل نشان می‌دهند که روش پیشنهادی دومرحله‌ای (در هر دو حالت همبستگی معمولی و تعمیم‌یافته) از عملکرد خیلی بهتری نسبت به روش SIS برخوردار است. حضور اثرات متقابل نیز بهبودی بیشتری را ایجاد کرده است.



شکل ۱۲.۴: نمودارهای نرخ خطای رده‌بندی 10-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با سه روش کاهش بعد برای مجموعه داده ساختار یونی اتمسفر



شکل ۱۳.۴: نمودارهای نرخ خطای رده‌بندی 10-NN (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با پنج روش کاهش بعد برای مجموعه داده ساختار یونی اتمسفر

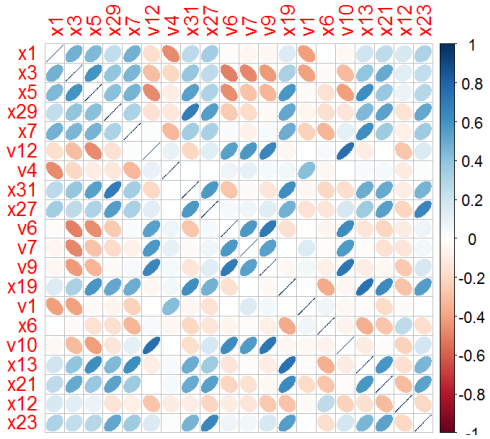


شکل ۱۴.۴: نمودارهای نرخ خطای رده‌بندی مرکز-مبنا (سمت چپ) و انحراف‌های معیار متناظر (سمت راست) با سه روش کاهش بعد برای مجموعه داده ساختار یونی اتمسفر

شکل (۱۴.۴) نیز نتیجه رده‌بندی داده‌های کاهش بعدیافته Ionoshpere را با روش مرکز-مبنا نشان می‌دهد. مشابه داده‌های Heart، در این‌جا نیز حضور اثرات متقابل کم‌ترین خطای رده‌بندی مرکز-مبنا

را به‌طور یکنواخت در پی داشته است.

مشابه مثال‌های قبلی، تحلیل این مجموعه داده نیز بر ورود اثرات متقابل در کاهش نرخ خطای رده‌بندی متغیر پاسخ در مسایل رده‌بندی تاکید دارد. شکل ۱۵.۴ نتیجه همبستگی خطی بین ۲۰ اثر موثر اول توسط روش پیشنهادی دو مرحله‌ای در مجموعه داده ساختار یونی اتمسفر می‌باشد که نشان دهنده مساله همبستگی خطی بین متغیرهای توضیحی است.



شکل ۱۵.۴: نمودار همبستگی خطی بین ۲۰ اثر موثر اول در مجموعه داده ساختار یونی اتمسفر

۳.۴ نتیجه‌گیری و آینده تحقیق

با توجه به نتایج مطالعات موردی و شبیه‌سازی می‌توان نتیجه گرفت که وارد کردن اثرات متقابل مهم در کنار اثرات اصلی تاثیر گذار، تاثیر قابل توجهی بر دقت روش‌های رده‌بندی و کاهش نرخ خطا دارد. این تاثیر در مقایسه با روش SIS و نتایج حاصل از آن، قابل تامل است. شایان ذکر است که مدل رده‌بندی ۱-نزدیک‌ترین همسایگی برای پیشگویی مناسب نیست زیرا ممکن است داده‌ها دارای مشاهدات پرت باشند. برای رفع این مشکل معمولا روش رده‌بندی ۱۰-نزدیک‌ترین همسایگی توصیه می‌شود.

با توجه به مطالب نظری، مطالعات موردی و شبیه‌سازی در این پایان‌نامه، می‌توان به موارد زیر به‌عنوان آینده تحقیق اشاره کرد:

۱. روش دو مرحله‌ای پیشنهاد شده برای انتخاب متغیرهای مهم با سایر روش‌های رده‌بندی نظیر تحلیل ممیزی خطی، درجه دو و شبکه‌های عصبی نیز قابل استفاده است و می‌تواند کاهش نرخ خطای رده‌بندی را در پی داشته باشد.

۲. مشابه پیشنهاد اول می‌توان روش SIS را با سایر روش‌های رده‌بندی استفاده کرده و نتایج را با روش پیشنهادی دو مرحله‌ای مقایسه کرد.

۳. از روش پیشنهادی می‌توان به‌عنوان یک گام پیش‌پردازنده برای تقلیل متغیرهای توضیحی در روش‌های رگرسیونی نظیر لاسو، SCAD و ریج بهره گرفت.

۴. می‌توان دقت کاهش بعد حاصل از روش پیشنهادی را با سایر روش‌های انتخاب متغیر نظیر صافی، پوشه و جداسازی مقایسه کرد.

پیوست آ

آ.۱ فضای برداری

تعریف آ.۱.۱ (میدان[۱]). فرض کنیم F نمایشگر مجموعه‌ی اعداد حقیقی یا مختلط باشد.

۱- جمع خاصیت جابه‌جایی دارد؛ یعنی به ازای هر x و y در F؛
$$x + y = y + x$$

۲- جمع خاصیت شرکت‌پذیری دارد؛ یعنی به‌ازای هر x، y و z در F؛
$$x + (y + z) = (x + y) + z$$

۳- عنصر یکتایی مانند صفر در F وجود دارد، به‌طوری‌که به‌ازای هر x در F؛
$$x + ۰ = x$$

۴- به‌ازای هر x در F عنصر یکتایی $(-x)$ در F متناظر است به‌طوری‌که:
$$x + (-x) = ۰$$

۵- ضرب خاصیت جابه‌جایی دارد؛ یعنی به ازای هر x و y در F؛
$$xy = yx$$

۶- ضرب خاصیت شرکت‌پذیری دارد؛ یعنی به ازای هر x، y و z در F؛
$$x(yz) = (xy)z$$

۷- عنصر غیر صفر یکتایی مانند یک در F وجود دارد، به‌طوری‌که به‌ازای هر x در F؛
$$x.۱ = x$$

[۱] Field

۸- به‌ازای هر x غیر صفر در F عنصر یکتایی $\frac{۱}{x}$ در F متناظر است به‌طوری‌که:
$x.\frac{۱}{x} = ۱$

۹- ضرب بر روی جمع پخش‌پذیر است؛ بدین معنی که، به‌ازای هر x، y و z در F؛
$x(y+z) = xy + xz$

فرض کنیم مجموعه F متشکل از اشیاء $x, y, z, \ldots$ و دو عمل، به‌صورت زیر، روی عناصر آن در دست باشد. عمل اول، که جمع نام دارد، به هر جفت عنصر x و y در F، عنصر $x+y$ در F را مربوط می‌سازد؛ عمل دوم، که ضرب نامیده می‌شود، به هر جفت عنصر x و y در F، عنصر xy در F را وابسته می‌سازد؛ به‌علاوه این دو عمل شرایط ۹-۱ بالا را دارند.

در این‌صورت، مجموعه F همراه با این دو عمل یک میدان نامیده می‌شود. به‌صورت نادقیق، یک میدان عبارت است از مجموعه‌ای همراه با چند عمل روی اشیاء آن که رفتار شبیه به اعمال جمع، ضرب، تفریق و تقسیم معمولی در اعداد دارند؛ بدین معنی که از نه قاعده جبری ذکر شده در بالا تبعیت می‌کنند. به عنوان مثال مجموعه اعداد مختلط و هم‌چنین اعداد حقیقی، همراه با اعمال جمع و ضرب معمولی یک میدان را تشکیل می‌دهند زعفرانی ج. (۱۳۵۴).

تعریف آ.۲.۱. یک فضای برداری متشکل از:

۱- یک میدان F از اسکالرها؛

۲- یک مجموعه V از اشیایی به نام بردارها؛

۳- یک قاعده (یا عمل) به‌نام جمع برداری که به هر جفت از بردارهای α و β از V، بردار $\alpha+\beta$ از V را که مجموع α و β نامیده می‌شود وابسته می‌سازد با این شرایط که:

الف- جمع خاصیت جابه‌جایی دارد؛ یعنی $\alpha + \beta = \beta + \alpha$

ب- جمع خاصیت شرکت پذیری دارد؛ یعنی $(\alpha+\beta)+\gamma = \alpha+(\beta+\gamma)$

ج- بردار یکتای $\underline{۰}$ به‌نام بردار صفر در V موجود است به‌طوری که به‌ازای هر α در V، $\alpha + \underline{۰} = \alpha$

د- به‌ازای هر بردار α در V، بردار یکتای $-\alpha$ در V موجود است که به‌طوری که $\alpha + (-\alpha) = \underline{۰}$

۴- یک قاعده (یا عمل) به‌نام ضرب اسکالری که به هر اسکالر c از F و هر بردار α از V بردار $c\alpha$ در V را که حاصل‌ضرب c و α نامیده می‌شود وابسته سازد با این شرایط که:

الف- به‌ازای هر بردار α در V، $۱.\alpha = \alpha$

ب- $(c_۱c_۲)\alpha = c_۱(c_۲\alpha)$

ج- $c(\alpha+\beta) = c\alpha + c\beta$

د- $(c_1 + c_2)\alpha = c_1\alpha + c_2\alpha$

توجه به این نکته مهم است که طبق تعریف، یک فضای برداری شیئی است مرکب، متشکل از یک میدان، یک مجموعه از بردارها، و دو عمل با خواص ویژه معین، مجموعه‌ی مفروضی از بردارها می‌تواند بخشی از چند فضای برداری متمایز باشد.

تعریف آ.۳.۱. تابع f به دامنه‌ی R^p و برد R^q را در صورتی خطی گوییم که به‌ازای هر a و b در R و هر x و y در R^p داشته باشیم:

$$f(ax + by) = af(x) + bf(y)$$

قضیه آ.۴.۱. اگر f یک تابع خطی به دامنه‌ی R^p و برد R^q باشد، آنگاه تعداد pq عدد حقیقی مانند c_{ij}، $1 \le i \le q$ و $1 \le j \le p$، وجود دارد به‌طوری که اگر $x = (x_1, \ldots, x_p)$ نقطه‌ای در R^p و $y = f(\underline{x}) = (y_1, \ldots, y_p)$ تصویر آن تحت f باشد، آنگاه:

$$\underline{c} = \begin{pmatrix} c_{11} & \ldots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{q1} & \ldots & c_{2p} \end{pmatrix}$$

$\underline{c}$ ماتریس تابع خطی f بوده و هم‌چنین داریم:

$$\begin{aligned} y_1 &= c_{11}x_1 + c_{12}x_2 + \cdots + c_{1p}x_p \\ y_2 &= c_{21}x_1 + c_{22}x_2 + \cdots + c_{2p}x_p \\ &\vdots \\ y_q &= c_{q1}x_1 + c_{q2}x_2 + \cdots + c_{qp}x_p \end{aligned} \qquad (آ.۱)$$

برعکس، هرگاه c_{ij} دسته‌ای از pq عدد حقیقی باشد، تابعی که به هر x در R^p عنصر y در R^q را طبق معادلات (آ.۱) مربوط کند یک تابع خطی به دامنه‌ی R^p و برد در R^q است.

تعریف آ.۵.۱. فرض کنید H دسته‌ی تمام توابع خطی $h : R^p \to R$ باشد آنگاه (R^p, R) یک فضای برداری از توابع خطی می‌باشد که در آن به‌ازای $a, b \in R$ و $\underline{x}, \underline{y} \in R^p$ داریم:

$$h(a\underline{x} + b\underline{y}) = ah(\underline{x}) + bh(\underline{y})$$

آ.۲ توابع به‌طور یکنواخت کراندار

تابع به‌طور یکنواخت کراندار برای اعداد حقیقی و مختلط:
فرض کنیم $\mathrm{F} = \{f_i; x \to k, i \in I\}$ که یک خانواده از توابع با مجموعه اندیس‌های مشخص I، که در آن x یک مجموعه دلخواه و k یک مجموعه از اعداد حقیقی یا مختلط می‌باشد. خانواده F را به‌طور یکنواخت کراندار گوییم هرگاه عدد حقیقی M وجود داشته باشد به‌طوری که:

$$|f_i(x)| \leq M \qquad \forall i \in I; \qquad \forall x \in X$$

برای مطالعه بیشتر، می‌توان به زعفرانی ج. (۱۳۵۴) مراجعه کرد.

پیوست ب
دستورات نرم افزار R

در این پیوست دستورات مربوط به ارزیابی اعتبار روش انتخاب متغیر دومرحله‌ای پیشنهادی و غربال‌گری مستقل مطمئن برای داده‌های تصویر برداری قلبی، ارایه شده است.
ابتدا بسته‌های زیر فراخوانی می‌شوند.

```
library(SIS)
library(class)
library(mgcv)
library(MASS)
```

دستورات مربوط به استاندارد سازی داده‌ها به‌صورت زیر می‌باشند:

```
set.seed(12110)
data1=read.table("D:\\data.txt",sep="",header=T)
n = length(data1[,1]) ; p = length(data1[1,])-1 ; n2 = 400 ; p2 = 20
standardf=function(data1,n,p){
  for(k in 1:p){
    data1[,k]=(data1[,k]-mean(data1[,k]))/sqrt(var(data1[,k]))
  }
  return(data1)
}
data1=as.data.frame(standardf(data1,n,p))
attach(data1)
```

دستورات مربوط به تولید داده‌های آموزشی و آزمون به‌صورت زیر می‌باشند:

```
splitdf <- function(data1) {
```

```
  index <- 1:nrow(data1)
  trainindex <- sample(index,trunc(length(index)*(2/3)))
  trainset <- data1[trainindex,]
  testset <- data1[-trainindex,]
 return(list(trainset=trainset,testset=testset,trainindex=trainindex))
}
trainset=as.data.frame(splitdf(data1)$trainset)
testset=as.data.frame(splitdf(data1)$testset)
attach(trainset)
attach(testset)
```

دستورات مربوط به محاسبه رده‌بندی مرکز- مبنا به‌صورت زیر می‌باشند:

```
fanavrege=function(trainset,p){
  avrege0=avrege1=rep(0,p)
  data0=trainset[trainset[,p+1]==0,]
  data1=trainset[trainset[,p+1]==1,]
  for(i in 1:p){
    avrege0[i]=mean(data0[,i])
    avrege1[i]=mean(data1[,i])
  }
  return(list(avrege0=avrege0,avrege1=avrege1))
}
b=dim(testset)[1]
fancentroid=function(testset,b){
  testset1=testset[,-(p+1)]
  labal=rep(0,b)
  for(i in 1:b){
    zz=(fanavrege(trainset,p)$avrege0
    -fanavrege(trainset,p)$avrege1)*(fanavrege(trainset,p)
    $avrege0+fanavrege(trainset,p)$avrege1-2*((testset1[i,])))
    zz1=(zz)
    z=sum(zz1)
    if(z>0){labal[i]=1}else{labal[i]=0}
    labal
  }
```

```
  ER=1-mean(labal==testset[,(p+1)])
  return(ER)
}
```

دستورات مربوط به رتبه‌دهی اثرات اصلی به‌همراه اثرات متقابل به‌صورت زیر می‌باشند:

```
fangen=function(data1,n,p){
  lhat = lhat.gam =rep(0,p)
  data1=as.data.frame(standardf(data1,n,p))
  for(j in 1:p){
   logistic.fit=glm(data1[,(p+1)]~data1[,j],data=data1,family=binomial)
    lhat[j] = as.numeric(-logLik(logistic.fit))
  }
  return(list(lhat))
}
fansub=function(data1,n,p){
  lhat =  rank = rep(0,p)
  r = fangen(data1, n, p)
  lhat = r[[1]]
  stlhat = sort(lhat)
  for(i in 1:p){
    for(j in 1:p){
      if(stlhat[i]==lhat[j]){rank[i]=j}
    }
  }
  return((rank))
}
index.main = fansub(data1,n,p)
data.main = cbind(data1[,index.main], data1[,(p+1)])
attach(data.main)
fanint=function(data1,n,p){
  dat = index.main
  m = floor(sqrt(p))+1
  k = (m/2)*(m-1)+p
  data = data.gam = matrix(0,n,m)
  data2 = matrix(0,n,k+1)
```

```
  nam = rep(0,k-p)
  data0 = data0.gam = matrix(0,n,k-p)
  data = data1[,c(dat[1:m])]
  t=1
  g=2
  for(i in 1:(m-1)){
    for(j in g:m){
      data0[,t] = data[,i]*data[,j]
      t=t+1
    }
    g=g+1
  }
  colnames(data0)=colnames(data0,do.NULL=FALSE,prefix="z")
  data2=cbind(data1[,c(dat[1:p])],data0,y)
  data2=as.data.frame(data2)
  return((data2))
}
m=floor(sqrt(p))+1
k=(m/2)*(m-1)+p
data2 = data2.gam = matrix(0,n,k+1)
data2 = ddd
data2 = as.data.frame(data2)
attach(data2)
xx = fansub(data2,n,k)
data3 = cbind(data2[,xx[1:length(xx)]], data2$y)
attach(data3)
data.interaction = as.data.frame(data3)
attach(data.interaction)
```

دستورات مربوط به انتخاب متغیر با غربال‌گری مستقل مطمئن به‌صورت زیر می‌باشند:

```
data1=as.matrix(data1)
x.train=data1[,-(p+1)]
y.train=data1[,p+1]
model2 = SIS(x.train, y.train, family="binomial", penalty="SCAD",
         concavity.parameter = 90, tune = "aic", type.measure = "deviance",
```

```
            nsis = p2, iter = TRUE, iter.max = ifelse(30,
            floor(2*nrow(x.train)/log(nrow(x.train)))),
            varISIS = "vanilla", perm = FALSE, seed = 10)
```

دستورات مربوط به محاسبه رده‌بندی مرکز-مبنا با روش انتخاب متغیر غربال‌گری مستقل مطمئن به‌صورت زیر می‌باشند:

```
classfier.sis=function(data1,n2,p2){
  test=as.data.frame(splitdf(data1)$testset)
  testy=test[,(p+1)]
  train=as.data.frame(splitdf(data1)$trainset)
  test=test[,model2$ix]
  rate=matrix(0,n2,p2)
  for(j in 1:n2){
    trainindex=sample(1:nrow(train),nrow(train),replace=T)
    train=train[trainindex,]
    ytrain=train[,(p+1)]
    train2=train[,model2$ix]
    for(i in 1:p2){
      train1=cbind(train2[,c(1:i)],ytrain)
      train1=as.data.frame(train1)
      test1=as.data.frame(test[,c(1:i)])
      fanavrege(train1,i)
      labal=rep(0,b)
      for(l in 1:b){
        zz=(fanavrege(train1,i)$avrege0
        -fanavrege(train1,i)$avrege1)*(fanavrege(train1,i)
        $avrege0+fanavrege(train1,i)$avrege1-2*((test1[l,])))
      zz1=(zz)
      z=sum(zz1)
      if(z>0){labal[l]=1}else{labal[l]=0}
      }
    rate[j,i]=1-mean(labal==testy)
      }
  }
  return(rate)
```

```
}
```

دستورات مربوط به محاسبه رده‌بندی مرکز-مبنا با روش انتخاب متغیر دو مرحله‌ای به‌صورت زیر می‌باشند:

```
data.main = cbind(data1[,index.main], data1[,(p+1)])
classfier.main=function(data.main,n2,p2){
  test=as.data.frame(splitdf(data.main)$testset)
  testy=test[,(p+1)]
  train=as.data.frame(splitdf(data.main)$trainset)
  test=test[,(1:p2)]
  rate=matrix(0,n2,p2)
  for(j in 1:n2){
    trainindex=sample(1:nrow(train),nrow(train),replace=T)
    train=train[trainindex,]
    ytrain=train[,(p+1)]
    train2=train[,(1:p2)]
    for(i in 1:p2){
      train1=cbind(train2[,c(1:i)],ytrain)
      train1=as.data.frame(train1)
      test1=as.data.frame(test[,c(1:i)])
      fanavrege(train1,i)
      labal=rep(0,b)
      for(l in 1:b){
        zz=(fanavrege(train1,i)$avrege0
        -fanavrege(train1,i)$avrege1)*(fanavrege(train1,i)
        $avrege0+fanavrege(train1,i)$avrege1-2*((test1[l,])))
        zz1=(zz)
        z=sum(zz1)
        if(z>0){labal[l]=1}else{labal[l]=0}
      }
      rate[j,i]=1-mean(labal==testy)
    }
  }
  return(rate)
```

```
}
classfier.interaction=function(data3,n2,p2){
  test=as.data.frame(splitdf(data3)$testset)
  testy=test[,k+1]
  train=as.data.frame(splitdf(data3)$trainset)
  test=test[,1:p2]
  rate=matrix(0,n2,p2)
  for(j in 1:n2){
    trainindex=sample(1:nrow(train),nrow(train),replace=T)
    train=train[trainindex,]
    ytrain=train[,k+1]
    train2=train[,1:p2]
    for(i in 1:p2){
      train1=cbind(train2[,c(1:i)],ytrain)
      train1=as.data.frame(train1)
      test1=as.data.frame(test[,c(1:i)])
      fanavrege(train1,i)
      labal=rep(0,b)
      for(l in 1:b){
        zz=(fanavrege(train1,i)$avrege0
        -fanavrege(train1,i)$avrege1)*(fanavrege(train1,i)
        $avrege0+fanavrege(train1,i)$avrege1-2*((test1[l,])))
        zz1=(zz)
        z=sum(zz1)
        if(z>0){labal[l]=1}else{labal[l]=0}
      }
      rate[j,i]=1-mean(labal==testy)
    }
  }
  return(rate)
}
```

دستورات مربوط به محاسبه خطا و انحراف معیار رده‌بندی مرکز- مبنا به‌صورت زیر می‌باشند:

```
Error.rate = function(rate, p2){
```

```
  RE=rep(0,p2)
  for(i in 1:p2){
    RE[i]=mean(rate[,i])
  }
  return(RE)
}
SE.Error.rate = function(rate,p2){
  SE=rep(0,p2)
  for(i in 1:p2){
    SE[i]=sqrt(var(rate[,i]))/n2
  }
  return(SE)
}
ratesis = classfier.sis(data1,n2,p2)
rate.sis = Error.rate(ratesis, p2)
SErate.sis = SE.Error.rate(ratesis, p2)
rate.glm.main = classfier.main(data.main,n2,p2)
rate.glm.interaction = classfier.interaction(data.interaction,n2,p2)
rate.glm.m = Error.rate(rate.glm.main, p2)
SErate.glm.m = SE.Error.rate(rate.glm.main, p2)
rate.glm.int = Error.rate(rate.glm.interaction, p2)
SErate.glm.int = SE.Error.rate(rate.glm.interaction, p2)
```

دستورات مربوط به رسم نمودارها به‌صورت زیر می‌باشند:

```
par(mfrow=c(1,2))
plot(rate.sis,lty=1,type="l",xlab=
"Number of top-ranked features selected",ylab=
" Error rates", ylim=c(.17,.7),axes=T, ann=T,
     cex.lab=1,cex.axis=1, lwd=2, mgp=c(2.4,.8,.5))
lines(rate.glm.m,type="l", lty=2, lwd=2)
lines(rate.glm.int,type="l", lty=3, lwd=2)
box(lwd=2.5)
legend(6.5,.75,c("SIS","main","interaction"),cex=1,lty=1:3,lwd=2,bty="n")
plot(SErate.sis,lty=1,type="l",xlab=
"Number of top-ranked features selected",ylab=
```

```
" Standard errors", ylim=c(.0001,.0095),axes=T, ann=T,
     cex.lab=1,cex.axis=1, lwd=2, mgp=c(2.6,.8,.5))
lines(SErate.glm.m,type="l", lty=2, lwd=2 )
lines(SErate.glm.int,type="l", lty=3, lwd=2 )
box(lwd=2.5)
legend(6.5,.0103,c("SIS","main","interaction"),cex=1,lty=1:3,lwd=2,bty="n")
```

مراجع

[۱] بارتل جی. آر، (۱۳۵۴)، "اصول آنالیز حقیقی"، زعفرانی ج.، چاپ دوم، انتشارات دانشگاه تهران، تهران.

[۲] محمدی، س.، شاهسونی د.، باغیشنی ح. (۱۳۹۴)، "انتخاب متغیر در داده‌های بزرگ"، نخستین کنفرانس ملی محاسبات نرم، دانشگاه گیلان، گیلان، ایران.

[۳] ژیاوی هان، میشلین کمبر و ژان پی، (۱۳۹۳)، "داده کاوی (مفاهیم و تکنیک‌ها)"، اسماعیلی، م.، چاپ اول، انتشارات نیاز دانش، تهران

[4] Abe, S. (2005), *Support Vector Machines for Pattern Classification*, Vol. 2, Springer, London.

[5] Altman, N. S. (1992), An introduction to kernel and nearest-neighbor nonparametric regression, *Journal of the American Statistician*, **46**(3), 175 – 185.

[6] Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, Vol. 1, Springer, New York.

[7] Bain, BJ. (2003), Bone marrow biopsy morbidity and mortality, *British Journal of Haematology*, **121**(6), 949 – 51.

[8] Bickel, PJ. (2008), Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society*, B, **70**, 883 – 884.

[9] Bickel, PJ. and Levina, E. (2004), Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations, *Journal of the Bernoulli*, **10**, 989 – 1010.

[10] Bickel, PJ. and Li, B. (2006), Regularization in statistics (with discussion), *Test*, **15**, 271 – 344.

[11] Bickel, PJ., Ritov, Y. and Tsybakov, A. (2008), Simultaneous analysis of LASSO and Dantzig selector, *The Annals of Statistics, To appear.*

[12] Boulesteix, A. (2004), PLS Dimension reduction for classification with microarray data, *Journal of the Statistical Applications in Genetics and Molecular Biology*, **3**, 1 – 33.

[13] Bradic, J., Fan, J. and Wang, W. (2009), Penalized composite quasi-likelihood for high-dimensional variable selection, *Manuscript*.

[14] Buhlmann, P. and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics, New York.

[15] Bunea, F., Tsybakov, A. and Wegkamp, MH. (2007), Sparsity oracle inequalities for the LASSO, *Electronic Journal of Statistics*, **1**, 169 – 194.

[16] Cai, T. and Lv, J. (2007), The Dantzig selector: statistical estimation when p is much larger than n, *The Annals of Statistics*, **35**, 2365 – 2369.

[17] Candes, E. and Tao, T. (2007), The Dantzig selector: Statistical estimation when p is much larger than n (with discussion), *The Annals of Statistics*, **35**, 2313 – 2404.

[18] Cao, HY. (2007), Moderate deviations for two sample t-statistics, *Probability and Statistics, ESAIM*, **11**, 264 – 627.

[19] Chiaromonte, F. and Martinelli J. (2002), Dimension reduction strategies for analyzing global gene expression data with a response, *Mathematical Biosciences*, **176**, 123 – 144.

[20] Craven M.W., and Shavlik J.W. (1997), Using neural networks for data mining, *Future Generation Computer Systems*, **13**, 221 – 229.

[21] Dabney, A. R. (2005), Classification of microarrays to nearest centroids, *Bioinformatics*, **21**, 4148 – 4154.

[22] Dahan, H., Cohen, S and Rokachan, L. (2014), *Proactive Data Mining with Decision Trees*, Springer, New York.

[23] Donoho, D. and Jin, J. (2008), Higher criticism thresholding: optimal feature selection when useful features are rare and weak, *Proceedings of the National Academy of Sciences*, **105**, 14790 – 14795.

[24] Dudoit, S., Fridlyand, J. and Speed,T.P. (2002), Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **97**(457), 77 – 87.

[25] Efron, B., Hastie, T. and Tibshirani, R. (2007), Discussion: The Dantzig selector: statistical estimation when p is much larger than n, *The Annals of Statistics*, **35**, 2358 – 2364.

[26] Fan and liv. (2008), Sure independence screening for ultrahigh dimensional feature space, *Royal Statistical Society*, B, **70**(5), 849 – 911.

[27] Fan, J. (1997), Comments on "Wavelets in statistics: a review," by A. Antoniadis, *Italian Statistical Society*, **6**, 131 – 138.

[28] Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *American Statistical Association*, **96**, 1348 – 1360.

[29] Fan, J. and Lv, J. (2010), A selective overview of variable selection in high dimensional feature space, *Statistica Sinica*, **20**(1), 101 – 148.

[30] Fan, J. and Peng, H. (2004), Nonconcave penalized likelihood with diverging number of parameters, *The Annals of Statistics*, **32**, 928 – 961.

[31] Fan, J. and Ren, Y. (2006), TStatistical analysis of DNA microarray data, *Clinical Cancer Research*, **12**, 4469 – 4473.

[32] Fan, J. and Song, R. (2009), Sure independence screening in generalized linear models with NP-dimensionality, *Revised for Journal of the Annals of Statistics*.

[33] Fan J, Fan. and Y, Lv J. (2008), High dimensional covariance matrix estimation using a factor model, *Econometrics*, **147**, 186 – 197.

[34] Fan, J., Fan, Y. and Lv, J. (2008), High-dimensional classification using features annealed independence rules, *The Annals of Statistics*, **36**(6), 2605 – 2637.

[35] Fan, J., Samworth, R. and Wu, Y. (2009), Ultrahigh dimensional variable selection: beyond the linear model, *Machine Learning Research*, **10**, 1829 – 1853.

[36] Friedman, J.H. (1991), Multivariate adaptive regression splines, *The Annals of Statistics*, **19**(1), 1 - 67.

[37] Golub, T. R., Slomin, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M. L., Down-ing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999), Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531 – 537.

[38] Greenshtein, E. (2006), Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint, *The Annals of Statistics*, **34**, 2367 – 2386.

[39] Gut A. (2005), *Probability: A Graduate Course*, Springer, New York.

[40] Guyon, I. and Elisseeff, A. (2003), An introduction to variable and feature selection, *Machine Learning Research*, **3**, 1157 – 1182.

[41] Hall, M.A. (2000), Correlation-based feature selection for discrete and numeric class machine learning, *InInternational Conference on Machine Learning*, Stanford, CA, 359 – 366.

[42] Hall,P. and Hao Xue,J. (2014), On selecting interacting features from high-dimensional data, *Computational Statistics and Data Analysis*, **71**, 694 – 708.

[43] Hall, P. and Miller, H. (2009), Using generalized correlation to effect variable selection in very high dimensional problems, *Computational and Graphical Statistics*, **18**(3), 533 – 550.

[44] Hall, P. and Miller, H. (2009b), Recursive methods for variable selection in very high dimensional classification, *Manuscript*.

[45] Hall, P., Titterington, DM. and Xue, J-H. (2008), Discussion of Sure independence screening for ultrahigh dimensional feature space, *Royal Statistical Society*, B, **70**, 889 – 890.

[46] Hall, P., Titterington, D. M. and Xue, J.-H. (2009), Median-based classifiers for high-dimensional data, *Manuscript*.

[47] Hall, P., Titterington, DM. and Xue J-H. (2009), Tilting methods for assessing the influence of components in a classifier, *Royal Statistical Society*, B, **71**(4), 783 – 803.

[48] Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edition, Springer - Verlag, New York.

[49] Hirotugu, A. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**(6), 716 – 723.

[50] Hosmer, Jr., David, W., and Lemeshow, S. (2004), *Applied logistic regression* John Wiley Sons.

[51] Hua, J., Tembe, W.D. and Dougherty, E.R. (2009), Performance of feature-selection methods in the classification of high-dimension data, *Pattern Recognition*, **42**(3), 409 – 424.

[52] Huang, J., Horowitz, J. and Ma, S. (2008), Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *The Annals of Statistics*, **36**, 587 – 613.

[53] Huang, X. and Pan, W. (2003), Linear regression and two-class classification with gene expression data, *Bioinformatics*, **19**, 2072 – 2078.

[54] Izenman, A.J. (2008), *Modern multivariate statistical techniques*, Wiley.

[55] Kim, Y. and Kwon, S. (2009), On the global optimum of the SCAD penalized estimator, *Manuscript*.

[56] Kim, Y., Choi, H. and Oh, HS. (2008), Smoothly clipped absolute deviation on high dimensions, *American Statistical Association*, **103**, 1665 – 1673.

[57] Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M.R. and Goodenday, L.S. (2001), Knowledge discovery approach to automated cardiac SPECT diagnosis, *Artificial Intelligence in Medicine*, **23**(2), 149 – 169.

[58] Lee, Y., Lin, Y. and Wahba, G. (2004), Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data, *American Statistical Association*, **99**, 67 – 81.

[59] Lei, Y. and Huan, L. (2003), Feature selection for high-dimensional data: a fastcorrelation-based filter solution, *In International Conference on Machine Learning*, 856 – 863, Washington DC, USA.

[60] Liu, H. and Motoda, H. (2008), *Computational methods of feature selection*, Chapman Hall/CRC: Data Mining and Knowledge Discovery Series, New York.

[61] Li, X. and Xu, R. (2009), *High-dimensional data analysis in cancer research*, Springer , New York.

[62] Lv, J. and Fan, Y. A. (2009), unified approach to model selection and sparse recovery using regularized least squares, *The Annals of Statistics*, **37**, 3498 – 3528.

[63] Mallows, CL. (1973), Some comments on C_p, *Technometrics*, **15**, 661 – 675.

[64] Meier, L., van de Geer, S. and Bühlmann, P. (2008), The group LASSO for logistic regression, *Royal Statistical Society*, B, **70**, 53 – 71.

[65] Melanie, M. (1996), *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, London, England.

[66] Mendel, J.M. (2001), *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*, IEEE, Computational Intelligence Magazine, Volume:2 , Issue: 1.

[67] Nelder, J. A. and R.W.M. Wedderburn (1972), Generalized Linear Models, *Royal Statistical Society*, Series A, **135**, 370 - 384.

[68] Nilsson, R. (2007), *Statistical feature selection with applications in life science*, Department of Physcis, Chemistry and Biology, Linköping University.

[69] Press, W.H., Teukolosky, S.A., Vetterling, W.T. and Flannery, B.P. (2007), *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge.

[70] Roger, P. (1955), A generalized inverse for matrices, *Proceedings of the Cambridge Philosophical Society*, **51**, 406 – 413.

[71] Robert, P.C. and Casella, G. (2010), *Introducing monte carlo methods with R*, Springer, New York.

[72] Saeys, Y., Iñza, I. and Larrañaga, P. (2007), A review of feature selection techniq ues in bioinformatics, *Bioinformatics*, **23**(19), 2507 – 2517.

[73] Schwarz, Gideon E. (1978), Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461 – 464.

[74] Segal, M. R., Dahlquist, K. D. and Conklin, B. R. (2003), Regression Approach for Microarray Data Analysis, *Computational Biology*, **10**, 961 – 980.

[75] Shao, J. and Deng, X. (2012), Estimation in high-dimensional linear models with deterministic design matrices, The Annals of Statistics, **40**(2) , 812 - 831.

[76] Sharon L. Lohr. (2009), *Sampling: Design and Analysis*, BROOKS/COLE, Boston, USA.

[77] Sigillito, V.G., Wing, S.P., Hutton, L.V. and Baker, K.B. (1989), Classification of radar returns from the ionosphere using neural networks, *Johns Hopkins APL Technical Digest*, **10**, 262 – 266.

[78] Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub,T.R. and Sellers, W.R. (2002), Geneexpression correlates of clinical prostate cancer behavior, *Cancer Cell1*, **2**, 203 – 209.

[79] Storey, JD. and Tibshirani, R. (2003), Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences*, **100**, 9440 – 9445.

[80] Tibshirani, R. (1966), Regression shrinkage and selection via the LASSO, *Royal Statistical Society*, B, **58**, 267 – 288.

[81] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003), Class prediction by nearest shrunken centroids, with applications to DNA microarrays, *Statistical Science*, **18**(1), 104 – 117.

[82] Tony Cai, T. and Shen, X. (2011), *High-dimensional data analysis*, World Scientific, China.

[83] Van de Geer, S. (2008), High-dimensional generalized linear models and the LASSO, *Annals of Statistics*, **36**, 614 – 645.

[84] Van der vaart A. W. (1998), *Asymptotic Statistics*, 1st Ed., Cambridge University Press, New York.

[85] Vapnik, V. (1995), *The Nature of Statistical Learning* Springer-Verlag, New York.

[86] Wang, H., Li, R, and Tsai, C-L. (2007), Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, **94**, 553 – 568.

[87] Wolberg, J. (2006), *Data analysis using the method of Least squares*, Springer, New York.

[88] Yuan, M. and Lin, Y. (2006), Model selection and estimation in regression with grouped variables, *Royal Statistical Society*, B, **68**, 49 – 67.

[89] Zhao, P. and Yu, B. (2006), On model selection consistency of LASSO, *Machine Learning Research*, **7**, 2541 – 2563.

[90] Zou, H. (2006), The adaptive LASSO and its oracle properties, *American Statistical Association*, **101**, 1418 – 1429.

واژه‌نامه فارسی به انگلیسی

هوش مصنوعی ..... Intelligence inteligence
داده کاوی ..... Data mining
اندازه ..... Measure
پایداری ..... Stability
یادگیری ماشین ..... Machine learning
پزشکی ..... Medical
فضای برداری از توابع ..... Vector space of functions
شیمی ..... Chemistry
اقتصاد ..... Economics
مهندسی ..... Engineering
پردازش متن ..... Text processing
پردازش تصویر ..... Image processing
آرایه های بیان ژن ..... Gene expression arrays
کاهش بعد ..... Dimensionality reduction
انتخاب ویژگی ..... Feature selection
انتخاب متغیر ..... Variable selection
انتخاب صفت ..... Attribute selection
بعد بالا ..... High dimensional
انتخاب ویژگی مبتنی بر همبستگی ..... Correlation-beasd feature selection
انبار داده‌ها ..... Data warehouse
بیش برازش ..... Over fitting
استخراج ویژگی ..... Feature extraction
داده ریز آرایه دی ان آ ..... DNA microarrays data
روش پوشه ..... Wrappers method
روش فیلتر ..... Filters method

روش جدا سازی ..... Embedded method
همبستگی مبتنی بر ویژگی ..... Correlation-based feature
استوار ..... Robust
لاسو ..... Lasso
ماشین بردار پشتیبان ..... Support vector machines
رگرسیون گام به گام ..... Stepwise regression
الگوریتم حریصانه ..... Greedy algorithm
تحلیل مولفه‌های اصلی ..... Principal component analysis
مقادیر ویژه ..... Eigen values
بردارهای ویژه ..... Eigen vectors
ماتریس کواریانس ..... Covariance matrix
متعامد ..... Orthogonal
نمونه‌گیری ..... Sampling
بدون جایگذاری ..... Without replacment
با جایگذاری ..... With replacment
کمی ..... Quantitative
کیفی ..... Qualitative
پیوسته ..... Continuous
گسسته ..... Discrete
ترتیبی ..... Ordinal
اسمی ..... Nominal
با نظارت ..... Superviesd
درخت تصمیم ..... Decision tree
شبکه‌های عصبی ..... Artificial neural networks
الگوریتم ژنتیک ..... Genetic algorithm
روش مارس ..... Mars method
فازی ..... Fuzzy
رده‌بندی مرکز مبنا ..... Centroid beasd classifacation
نزدیک‌ترین همسایگی ..... Nearest neighbor
یادگیرنده تنبل ..... Lazy learner
داده آزمون ..... Testing data
داده آموزشی ..... Training data

اعتبار سنجی متقابل ..... Cross validation
نیکویی برازش ..... Goodness of fit
پارامترهای میزان‌ساز ..... Tuning parameters
رگرسیون مولفه به مولفه ..... Componentwise regression
غربال‌گری مستقل مطمئن تکراری ..... Iterative sure independence screening

واژه‌نامه انگلیسی به فارسی

هوش مصنوعی ..... Intelligence inteligence
داده کاوی ..... Data mining
اندازه ..... Measure
پایداری ..... Stability
یادگیری ماشین ..... Machine learning
پزشکی ..... Medical
فضای برداری از توابع ..... Vector space of functions
شیمی ..... Chemistry
اقتصاد ..... Economics
مهندسی ..... Engineering
پردازش متن ..... Text processing
پردازش تصویر ..... Image processing
آرایه های بیان ژن ..... Gene expression arrays
کاهش بعد ..... Dimensionality reduction
انتخاب ویژگی ..... Feature selection
انتخاب متغیر ..... Variable selection
انتخاب صفت ..... Attribute selection
بعد بالا ..... High dimensional
انتخاب ویژگی مبتنی بر همبستگی ..... Correlation-beasd feature selection
انبار داده‌ها ..... Data warehouse
بیش برازش ..... Over fitting
استخراج ویژگی ..... Feature extraction
داده ریز آرایه دی ان آ ..... DNA microarrays data
روش پوشه ..... Wrappers method
روش فیلتر ..... Filters method

روش جدا سازی ..... Embedded method
همبستگی مبتنی بر ویژگی ..... Correlation-based feature
استوار ..... Robust
لاسو ..... Lasso
ماشین بردار پشتیبان ..... Support vector machines
رگرسیون گام به گام ..... Stepwise regression
الگوریتم حریصانه ..... Greedy algorithm
تحلیل مولفه‌های اصلی ..... Principal component analysis
مقادیر ویژه ..... Eigen values
بردارهای ویژه ..... Eigen vectors
ماتریس کواریانس ..... Covariance matrix
متعامد ..... Orthogonal
نمونه‌گیری ..... Sampling
بدون جایگذاری ..... Without replacment
با جایگذاری ..... With replacment
کمی ..... Quantitative
کیفی ..... Qualitative
پیوسته ..... Continuous
گسسته ..... Discrete
ترتیبی ..... Ordinal
اسمی ..... Nominal
با نظارت ..... Superviesd
درخت تصمیم ..... Decision tree
شبکه‌های عصبی ..... Artificial neural networks
الگوریتم ژنتیک ..... Genetic algorithm
روش مارس ..... Mars method
فازی ..... Fuzzy
رده‌بندی مرکز مبنا ..... Centroid beasd classifacation
نزدیک‌ترین همسایگی ..... Nearest neighbor
یادگیرنده تنبل ..... Lazy learner
داده آزمون ..... Testing data
داده آموزشی ..... Training data

اعتبار سنجی متقابل ..... Cross validation
نیکویی برازش ..... Goodness of fit
پارامترهای میزان‌ساز ..... Tuning parameters
رگرسیون مولفه به مولفه ..... Componentwise regression
غربال‌گری مستقل مطمئن تکراری ..... Iterative sure independence screening

Abstract

New technologies make big data that their analysis cause to invent new statistical methods. Before using these methods, some techniques are normally used to reduce the data dimension and provide a basis to quantify the importance of features. The methods such as sure independence screening (SIS) detect the effective variables by ranking. These methods don't consider the interaction effects, whereas the experience of researcher may indicate the necessity of interactions. In this thesis, by using generalized correlation coefficients, a two-stage algorithm are presented to determine the effective main and interaction effects. The advantage of the proposed method in comparison with SIS is that it can improve the performance of classification techniques, like K-nearest neighbors and centroid based.

Keywords: Classification, Sure independence screening, Feature ranking, Logistic regression, Generalised correlation.

Division of Statistics

Shahrood University of Technology

Faculty Of Mathematical Sciences

Recognition of effective interactions for high dimensional data

Saeed Mohammadi

Supervisors

Davood Shahsavani and Hossein Baghishani

January 2016