

الحمد لله
الذي هدانا لهذا
الذي كنا لنهتدي لولا
هدايتنا



دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد آمار

جنگل‌های تصادفی بهبود یافته برای متغیر پاسخ رتبه‌ای

نگارنده: مریم علی بیگی

استاد راهنما

دکتر داود شاهسونی

بهمن ۱۳۹۶

شماره:

تاریخ:

باسمه تعالی



مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم **مریم علی بیگی** با شماره دانشجویی **۹۴۱۲۵۴۴** رشته ریاضی کاربردی گرایش آنالیز عددی تحت عنوان **جنگل های تصادفی بهبود یافته برای متغیر پاسخ رتبه ای** که در تاریخ **۹۶/۱۱/۱۰** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

☐ مردود ☒ قبول (با درجه: خیلی خوب)			
نوع تحقیق: ☒ نظری ☐ عملی			
عضو هیأت داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱- استاد راهنمای اول	دکتر داود شاهسونی	دانشیار	
۳- استاد مشاور	_____	_____	_____
۴- نماینده تحصیلات تکمیلی	دکتر محمدرضا ربیعی	استادیار	
۵- استاد ممتحن اول	دکتر محمد آرشی	دانشیار	
۶- استاد ممتحن دوم	دکتر حسین باغیشنی	استادیار	

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده: ۹۶/۱۱/۱۰



تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند از پایان نامه خود دفاع نماید (دفاع

مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم به آستان حقیقت و آنان که وصالش را می جویند و آنان که
خود عین حقیقت اند.

تقدیم به پدر، مادر، مهربانم که هر لحظه وجودم را از چشمه سار پر از عشق
چشمانشان سیراب می کنند.

تقدیم به خواهران و برادران عزیزم که وجودشان شادی، بخش
و صفایشان مایه آرامش من است.

به نام پروردگار بخشنده و مهربان

شکر خدا که هرچه طلب کردم از خدا..... بر منتهای همت خود کامران شدم

سپاسگزار کسانی هستم که سرآغاز تولد من هستند. از یکی زاده می‌شوم و از دیگری جاودانه. استادانی که سپیدی را بر تخته سیاه زندگیم نگاشتند و مادری که تار مویی از او بیای من سیاه نماند. از قبیله‌ی سبز نجابت بوده است و با زبان مردم بهشت سخن می‌گوید، جاده‌ای از ابریشم ایمان بر سر دارد و قبلش به عرش خدا می‌ماند که به اندازه حقیقت خدا بزرگ است. بی آنکه کسی بداند آفتاب از مشرق صدای او در من طلوع می‌کرد. من صدای خدا را از صدای قلب او می‌شنوم، مادر و خواهران عزیزم سپاسگزارم... پدر و برادران گرانقدرم به خاطر وجودتان خدا را شاکرم.

جناب آقای دکتر شاهسونی استاد راهنمای گرامی: شما روشنایی بخش تاریکی جان هستی و ظلمت اندیشه را نور می‌بخشی. چگونه سپاس گویم مهربانی و لطف شما را که سرشار از معرفت و اخلاق است. چگونه سپاس گویم تاثیر علم آموزی شما را که چراغ روشن هدایت را بر کلبه‌ی محقر وجودم فروزان ساخته است. آری در مقابل این همه عظمت و شکوه شما مرا نه توان سپاس است و نه کلام وصف.

از اساتید گرانقدر آقای دکتر آرشی و آقای دکتر باغیشنی که زحمت داوری این پایان‌نامه را متقبل شدند، همچنین از خانم دکتر اقبال، آقای دکتر خدای و آقای دکتر نزاکتی کمال تشکر را دارم.

در پایان از دوستان عزیزم خانم‌ها زهره فرهادی، فاطمه سهرابی و اکرم قائمی‌زاده به پاس عاطفه سرشار و گرمای امید بخش وجودشان که صمیمانه در پیشبرد این پروژه از هیچ کمکی دریغ نکردند قدردانی می‌کنم.

مریم علی بیگی
بهمن ۱۳۹۶

تعهد نامه

اینجانب **مریم علی بیگی** دانشجوی کارشناسی ارشد رشته آمار دانشکده علوم ریاضی دانشگاه صنعتی شاهرود، نویسنده پایان نامه با عنوان **جنگل های تصادفی بهبود یافته برای متغیر پاسخ رتبه ای**، تحت راهنمایی **داود شاهسونی** متعهد می شوم:

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش های دیگر پژوهش گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا 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 به دست آوردن نتایج اصلی پایان نامه تاثیرگذار بوده اند، در مقالات مستخرج از پایان نامه رعایت می گردد.
- در تمام مراحل انجام این پایان نامه، در مواردی که از موجود زنده (یا بافت های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

مریم علی بیگی

بهمن ۱۳۹۶

مالکیت نتایج و حق نشر

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

چکیده

جنگل تصادفی یکی از روش‌های فراگیر و مدرن داده‌کاوی است که در زمره روش‌های اجماع خودگردان یا Bagging قرار گرفته و ترکیبی از درخت‌های تصمیم می‌باشد. این روش، یک ابزار معمول برای رده‌بندی داده‌ها در ابعاد بالا است به طوری که قادر به پیش‌بینی رده مشاهدات و تعیین رتبه متغیرهای پیش‌گو براساس میزان اهمیت آن متغیرها بوده و قابل استفاده برای متغیر پاسخ از نوع کمی، اسمی یا بقا است. اما وقتی متغیر پاسخ از نوع رتبه‌ای است، بطور معمول آن را به صورت اسمی در نظر می‌گیرند که باعث نادیده گرفتن اطلاعات رتبه‌ای و کاهش دقت پیش‌بینی می‌شود. در این پایان‌نامه با اعمال تغییراتی در الگوریتم درخت تصمیم، گونه‌ای متفاوت از روش جنگل تصادفی که مختص متغیر پاسخ رتبه‌ای است مورد مطالعه قرار می‌گیرد. نتایج داده‌های شبیه‌سازی و واقعی حاکی از آن است که با اعمال این تغییرات، عملکرد پیش‌بینی نسبت به روش کلاسیک جنگل تصادفی بهبود یافته است.

کلمات کلیدی: درخت تصمیم، درخت استنباط شرطی، جنگل تصادفی، جنگل تصادفی رتبه‌ای.

فهرست مطالب

ک فهرست تصاویر

م فهرست جداول

۱ مقدمه و مفاهیم درخت تصمیم

۱.۱ مقدمه ۱

۲.۱ آزمون استنباط شرطی ۳

۱.۲.۱ آزمون جایگشتی ۵

۳.۱ درخت تصمیم ۵

۱.۳.۱ الگوریتم‌های مطرح درخت تصمیم ۶

۲.۳.۱ CART ۷

۳.۳.۱ پیش‌بینی درخت تصمیم ۱۲

۴.۳.۱ دقت درخت تصمیم ۱۲

۵.۳.۱ توقف ساخت درخت ۱۲

۶.۳.۱ روش هرس درخت تصمیم ۱۳

۷.۳.۱ مزایا و معایب درخت تصمیم ۱۴

۴.۱ درخت استنباط شرطی ۱۴

۱.۴.۱ انتخاب متغیر برای افراز فضا ۱۵

۲.۴.۱ انتخاب نقطه افراز ۱۶

۳.۴.۱ ضابطه توقف ۱۶

۴.۴.۱ پیش‌بینی درخت استنباط شرطی ۱۶

۵.۴.۱ الگوریتم درخت استنباط شرطی ۱۷

۲ جنگل تصادفی ۱۹

۱.۲ روش رده‌بندی اجماعی ۱۹

۱.۱.۲ اجماع‌سازی خودگردان ۲۰

۲۱	جنگل تصادفی	۲.۲
۲۱	الگوریتم جنگل تصادفی کلاسیک	۱.۲.۲
۲۲	رده‌بندی با جنگل تصادفی	۲.۲.۲
۲۳	دقت رده‌بندی و پیش‌بینی جنگل تصادفی کلاسیک	۳.۲.۲
۲۳	اهمیت متغیرها در جنگل تصادفی	۴.۲.۲
۲۴	جنگل تصادفی رتبه‌ای	۳.۲
۲۵	الگوریتم جنگل تصادفی رتبه‌ای	۱.۳.۲
۲۵	معیارهای عملکرد	۲.۳.۲
۲۹	اهمیت متغیر جنگل تصادفی رتبه‌ای	۳.۳.۲
۲۹	ارزیابی کیفیت رتبه‌بندی متغیرها	۴.۳.۲
۳۱	مطالعات شبیه‌سازی و داده‌های واقعی	۳
۳۱	داده‌های شبیه‌سازی شده	۱.۳
۳۳	بررسی عملکرد جنگل‌های تصادفی	۱.۱.۳
۳۶	بررسی اهمیت متغیرها	۲.۱.۳
۴۳	معرفی مجموعه داده‌های واقعی	۲.۳
۴۴	اجرای الگوریتم بر روی مجموعه داده‌ها	۱.۲.۳
۴۵	بررسی اهمیت متغیرها	۲.۲.۳
۴۹		آ
۴۹	مجموعه‌های آموزشی و آزمون	۱.آ
۴۹	اعتبارسنجی متقابل	۲.آ
۵۰	اعتبارسنجی K-Fold	۱.۲.آ
۵۱		ب
۶۷	مراجع	
۷۱	واژه‌نامه فارسی به انگلیسی	
۷۳	واژه‌نامه انگلیسی به فارسی	

فهرست تصاویر

۸	۱.۱	نمودار افراز فضای مشاهدات در یک درخت تصمیم
۱۰	۲.۱	نمودار پراکنش مشاهدات
۱۰	۳.۱	نمودار حاصل از انتخاب اولین افراز
۱۱	۴.۱	مراحل افراز فضای مشاهدات
۱۱	۵.۱	ساختار درختی مشاهدات
۲۰	۱.۲	ساختار رده‌بندی اجماعی
۲۲	۲.۲	ساختار جنگل تصادفی
	۱.۳	نمودار جعبه‌ای نسبت خطای جنگل تصادفی رتبه‌ای به جنگل تصادفی کلاسیک
۳۵		وقتی متغیرهای توضیحی فاقد همبستگی هستند
	۲.۳	نمودار جعبه‌ای نسبت خطای جنگل تصادفی رتبه‌ای به جنگل تصادفی کلاسیک
۳۵		با فرض همبستگی درونی متغیرها
	۳.۳	نمودار جعبه‌ای AUC برای $K = 9$ با استفاده از معیارهای عملکرد برای مقایسه
		جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک وقتی متغیرهای توضیحی
۳۷		فاقد همبستگی هستند
	۴.۳	نمودار جعبه‌ای AUC برای $K = 9$ با استفاده از معیارهای عملکرد برای مقایسه
		جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک با فرض همبستگی درونی
۳۷		متغیرها
	۵.۳	نمودار جعبه‌ای AUC برای $K = 6$ با استفاده از معیارهای عملکرد برای مقایسه
		جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک وقتی متغیرهای توضیحی
۳۹		فاقد همبستگی باشند
	۶.۳	نمودار جعبه‌ای AUC برای $K = 6$ با استفاده از معیارهای عملکرد برای مقایسه
		جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک با فرض همبستگی درونی
۳۹		متغیرها

۷.۳	نمودار جعبه‌ای AUC برای $K = 3$ با استفاده از معیارهای عملکرد برای مقایسه جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک وقتی متغیرهای توضیحی فاقد همبستگی هستند	۴۰
۸.۳	نمودار جعبه‌ای AUC برای $K = 3$ با استفاده از معیارهای عملکرد برای مقایسه جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک با فرض همبستگی درونی متغیرها	۴۰
۹.۳	نسبت تعداد درختان برای انتخاب متغیر با اهمیت در اولین افراز، وقتی متغیرهای توضیحی فاقد همبستگی هستند	۴۲
۱۰.۳	نسبت تعداد درختان برای انتخاب متغیر با اهمیت در اولین افراز، با فرض همبستگی درونی متغیرها	۴۲
۱۱.۳	نمودار جعبه‌ای نسبت خطای جنگل تصادفی رتبه‌ای به جنگل تصادفی کلاسیک برحسب معیار ER (نمودار سمت چپ) و معیار RPS (نمودار سمت راست)	۴۵
۱۲.۳	مقایسه معیارهای خطا برای ارزیابی اهمیت متغیرها در جنگل تصادفی رتبه‌ای	۴۶
۱۳.۳	مقایسه معیارهای خطا برای ارزیابی اهمیت متغیرها در جنگل تصادفی کلاسیک	۴۶

فهرست جداول

۲۷	مشاهدات با متغیر پاسخ اسمی	۱.۲
۲۸	مشاهدات دارای متغیر پاسخ رتبه‌ای	۲.۲
۳۲	ضرایب ثابت γ	۱.۳
۳۲	ضرایب متغیرهای پیش‌گو موثر و بدون اثر	۲.۳

فصل ۱

مقدمه و مفاهیم درخت تصمیم

۱.۱ مقدمه

رده‌بندی یکی از موضوعات آماری و مفاهیم داده‌کاوی^۱ است که در آن، رابطه‌ی متغیر پاسخ از نوع کیفی، با چندین متغیر توضیحی مورد بررسی قرار می‌گیرد و دستورالعملی موسوم به قاعده تصمیم استخراج می‌شود. در این صورت می‌توان با داشتن مشاهدات متغیرهای توضیحی، رده‌ای برای متغیر پاسخ براساس آن قاعده تصمیم پیش‌گویی کرد. از آنجا که داده‌های کیفی به دو نوع اسمی^۲ و ترتیبی^۳ تقسیم می‌شوند لذا روش‌های رده‌بندی متغیر پاسخ اسمی و متغیر پاسخ ترتیبی متفاوت هستند. رگرسیون لجستیک^۴، رگرسیون ماتریس نشانگر، تحلیل ممیزی خطی، درخت تصمیم^۵ و جنگل‌های تصادفی^۶ مثال‌هایی از روش‌های آماری متغیر پاسخ اسمی هستند. برای رده‌بندی متغیر پاسخ رتبه‌ای نیز روش‌هایی چون بخت‌های متناسب^۷ و رده‌ی همسایه^۸ پیشنهاد شده‌اند [۲]، که در این دو

^۱ Data mining:

^۲ Nominal

^۳ Ordinal

^۴ Logistic regression

^۵ Decision tree

^۶ Random forests

^۷ Proportional odds

^۸ Adjacent category

روش، مدل‌های در نظر گرفته شده به ترتیب عبارتند از:

$$\log \left(\frac{p(Y \leq j)}{p(Y > j)} \right) = \alpha_j + \beta_1 X_1 + \dots + \beta_p X_p$$

$$\log \left(\frac{p(Y = j)}{p(Y = j + 1)} \right) = \alpha_j + \beta_1 X_1 + \dots + \beta_p X_p \quad j = 1, \dots, K$$

که K تعداد سطوح پاسخ است. این روش‌ها هنگامی که رابطه‌ی بین متغیر پاسخ و متغیرهای پیش‌گو از پیچیدگی ذاتی خاصی، شامل همبستگی و اثر متقابل بین متغیرها، بخصوص در بعد بالا برخوردار است، کارایی لازم را ندارند [۲۵]. علاوه بر این اگر تعداد متغیرهای پیش‌گو، بسیار بیشتر از تعداد مشاهدات باشد برآورد پارامترهای مدل با مشکل عدم پایداری محاسبات عددی مواجه خواهد شد. همچنین این روش‌ها مبتنی بر فرضیه‌هایی هستند که در کاربردهای واقعی، عملیاتی نیستند.

یکی از راه‌حل‌ها برای متغیر پاسخ رتبه‌ای، استفاده از توانایی‌های ذاتی روش‌های رده‌بندی متغیر اسمی است که با تغییراتی جزئی در ساختار آن بتوان، موضوع رتبه‌ای بودن متغیر پاسخ را لحاظ کرد. درخت تصمیم و جنگل تصادفی که در زمره‌ی روش‌های ناپارامتری قرار می‌گیرند، دارای انعطاف‌پذیری و ویژگی‌های خاصی هستند که می‌توان با تغییراتی در الگوریتم، از آن‌ها برای پاسخ‌های رتبه‌ای نیز استفاده نمود.

جنگل تصادفی یکی از روش‌های یادگیری با نظارت است [۵] و بر پایه‌ی درخت تصمیم بنا شده که در زمینه‌های مختلف علوم از جمله بیوانفورماتیک، ژنتیک و همه‌گیرشناسی به کار رفته و قابل کاربرد در مسائلی است که غیرخطی بودن، بعد بالا، همبستگی و اثر متقابل متغیرهای توضیحی جزء لاینفک ذاتی این متغیرها است [۲۴، ۲۵، ۳۴]. این روش، ترکیبی از درخت‌های رگرسیون یا رده‌بندی است که در آن، مشاهدات (داده‌ها) در نظر گرفته‌شده برای هر درخت، نمونه تصادفی خودگردان^۹ از داده‌های اصلی است. همچنین از آن می‌توان برای رده‌بندی داده‌ها با بعد بالا استفاده کرد و قادر به تعیین رتبه متغیرهای پیش‌گو از طریق میزان اهمیت آن متغیرها است. هنگامی که متغیر پاسخ از نوع کمی و اسمی است، رهیافت کلاسیک این روش به خوبی از عهده پیش‌گویی این نوع متغیر پاسخ برمی‌آید، اما وقتی متغیر پاسخ رتبه‌ای است جنگل‌های تصادفی معمولاً متغیر رتبه‌ای را به صورت اسمی در نظر می‌گیرند که بدین ترتیب اطلاعات رتبه‌ای داده‌ها نادیده گرفته شده و دقت پیش‌گویی کاهش می‌یابد.

موضوع اصلی این پایان‌نامه، ارائه روشی برای رده‌بندی متغیر پاسخ رتبه‌ای براساس روش معروف جنگل‌های تصادفی است و لذا ساختار این پایان‌نامه به شرح زیر است:

۱. در فصل اول، موضوع همبستگی متغیر رتبه‌ای با انواع متغیرهای رتبه‌ای، اسمی و کمی در قالب آزمون استقلال شرطی و بوسیله‌ی آماره این آزمون، بررسی می‌شود.

^۹ Bootstrap

سپس درخت کلاسیک تصمیم و درخت استنباط شرطی معرفی می‌شوند. درخت تصمیم، خاص متغیر پاسخ اسمی و درخت استنباط شرطی، فرم تغییر یافته درخت کلاسیک برای متغیر پاسخ رتبه‌ای است.

۲. جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای، که نقطه هدف این پایان نامه است در فصل دوم مطرح می‌شود.

۳. در فصل سوم، با تولید داده‌های شبیه‌سازی شده و استفاده از داده‌های واقعی، عملکرد جنگل تصادفی رتبه‌ای، مورد بحث واقع شده است.

۲.۱ آزمون استنباط شرطی

فرض کنید X ، متغیر توضیحی و Y متغیر پاسخ باشد و $\{(x_i, y_i); i = 1, \dots, n\}$ مجموعه مشاهدات را تشکیل دهند. برای اندازه‌گیری میزان همبستگی این دو متغیر، آزمون فرضیه استقلال X و Y یعنی $H_0: D(Y|X) = D(Y)$ مورد بحث قرار می‌گیرد [۱۷] که آماره آزمون یا ابزار سنجش همبستگی متغیر پاسخ رتبه‌ای با متغیر توضیحی به صورت (۱.۱) تعریف می‌شود:

$$T = \sum_{i=1}^n g(X_i) s(Y_i) \quad (1.1)$$

که در آن $g: X \rightarrow \mathbb{R}^p$ یک تبدیل غیرتصادفی از X ، فضای متغیر توضیحی است، X_i معرف مشاهده‌ی i ام متغیر X و $s(Y_i)$ نمره یا امتیاز Y_i نامیده می‌شود. تابع s تابعی است که ماهیت کیفی متغیر پاسخ را با رتبه‌های $1, \dots, K$ به ماهیت کمی تبدیل می‌کند به طوری که $s(1) < s(2) < \dots < s(K)$. در اینجا ساده‌ترین فرم آن یعنی $s(r) = r$ مورد استفاده قرار گرفته است.

تابع g وابسته به نوع متغیر توضیحی مربوطه است:

- اگر X متغیری از نوع کمی باشد، g یک تابع همانی است. بنابراین

$$T = \sum_{i=1}^n X_i s(Y_i) \in \mathbb{R}$$

- اگر X متغیری از نوع اسمی با J سطح باشد، آنگاه $p = J$ و بنابراین $g(X_i)$ یک بردار J تایی است که مولفه‌ی j ام مقدار یک و بقیه اعضای آن، مقدار صفر را می‌گیرند. به عنوان مثال اگر سطوح X عبارت از $(5, 7, 8, 11)$ باشند و مشاهده اول این متغیر $X_1 = 8$ باشد آنگاه $g(X_1) = (0, 0, 1, 0)$ است. بنابراین

$$t = \sum_{i=1}^n g(X_i) s(Y_i) \in \mathbb{R}^J$$

بنابراین با استانداردسازی آماره آزمون با استفاده از امید شرطی μ و ماتریس کواریانس Σ داریم:

$$T = \max \left| \frac{t - \mu}{\text{diag} \Sigma} \right|$$

که در آن

$$\mu = \mathbb{E}(T|B) = \text{vec} \left(\left(\sum_{i=1}^n g(X_i) \right) \mathbb{E}(s|B) \right)$$

$$\Sigma = \mathbb{V}(T|B) = \frac{n}{n-1} \mathbb{V}(s|B) \otimes \left(\sum_i g(X_i) \otimes g(X_i)^T \right)$$

$$- \frac{1}{n-1} \mathbb{V}(s|B) \otimes \left(\sum_i g(X_i) \right) \otimes \left(\sum_i g(X_i) \right)^T$$

که عملگر $\otimes$ نشان دهنده ضرب کرونکر است. حاصل عبارت $(\sum_{i=1}^n g(X_i) \mathbb{E}(s|B))$ یک ماتریس است که با استفاده از عملگر vec به بردار تبدیل می شود. همچنین

$$\mathbb{E}(s|B) = \frac{1}{n} \left(\sum_i s(Y_i) \right)$$

$$\mathbb{V}(s|B) = \frac{1}{n} \left(\sum_i (s(Y_i) - \mathbb{E}(s|B))(s(Y_i) - \mathbb{E}(s|B))^T \right).$$

• اگر X متغیری رتبه ای با J سطح باشد آماره ی آزمون به صورت (۲.۱) تعریف می شود:

$$T = \sum_{i=1}^n \gamma^T g(X_i) s(Y_i) \in \mathbb{R} \quad (2.1)$$

که $\gamma \in \mathbb{R}^J$ بردار امتیاز متغیر توضیحی است.

توزیع T به توزیع توأم Y و X بستگی دارد که در عمل نامعلوم است و به همین دلیل، از آزمون جایگشتی استفاده می شود که در آن متغیر توضیحی ثابت، و جایگشت های متفاوت متغیر پاسخ، در نظر گرفته می شود. بدین صورت مقدار آماره ی آزمون به ازای همه متغیرهای توضیحی حاصل می شود. با به دست آوردن توزیع شرطی $P(T \leq z|B)$ که در آن B تعداد کل جایگشت ها و z نسبت تعداد جایگشت هایی است که مقدار آماره آزمون به ازای آن ها بیشتر از مقدار آماره آزمون داده های اصلی است. با انجام مکرر این کار، آماره های متناظر برای همان متغیر نیز به دست آمده و نسبت تعداد دفعاتی که آماره آزمون اصلی از آماره های متناظر با داده های جایگشتی کمتر باشد، p - مقدار به دست می آید. بدین صورت p - مقدار به ازای متغیرهای توضیحی حاصل می شود. در نهایت اندازه همبستگی متغیرهای توضیحی با متغیر پاسخ رتبه ای حاصل می شود، متغیرهایی که p - مقدار کمتری دارند، در واقع بیشترین ارتباط را با Y خواهند داشت.

مقادیر بزرگ آماره یعنی تمایل بیشتری برای رد فرض استقلال وجود دارد، بنابراین تمایل بیشتری وجود دارد که پذیرفته شود X و Y به هم وابسته هستند یعنی هر چقدر p - مقدار کوچکتر شود معادل آن است که X و Y به هم وابستگی بیشتری دارند [۳۵، ۱۷].

۱.۲.۱ آزمون جایگشتی

آزمون جایگشتی^{۱۰} شاخه‌ای از آزمون‌های ناپارامتری و جزء آزمون‌های آزاد توزیع است. در تحلیل داده‌ها، نامطلوب‌ترین حالت زمانی است که هم توزیع جامعه نامعلوم باشد و هم حجم نمونه‌ها کم که می‌توان از بازنمونه‌گیری استفاده کرد. از روش‌های معمول بازنمونه‌گیری می‌توان به خودگردان‌سازی و جک‌نایف^{۱۱} و آزمون جایگشتی اشاره کرد. این روش نوعی شبیه‌سازی محسوب می‌شود و روند کار به گونه‌ای است که به اندازه دلخواه نمونه با جایگذاری و بدون جایگذاری از مشاهدات برداشته شود و از میانگین این نمونه‌ها به عنوان برآوردی بهتر استفاده می‌شود. بنابراین آزمون‌های جایگشتی از روش‌های بازنمونه‌گیری محسوب می‌شود [۳۵].

مراحل آزمون جایگشتی

برای آزمون جایگشتی مراحل زیر انجام می‌شود:

۱. فرض صفر آزمون تعریف می‌شود.
۲. یک آماره آزمون با توجه به آزمون موردنظر انتخاب می‌شود.
۳. آماره آزمون برای داده‌های اصلی در دسترس محاسبه می‌شود.
۴. یک جایگشت از داده‌های اصلی حاصل می‌شود.
۵. آماره آزمون برای این دسته از داده‌های جدید محاسبه می‌شود.
۶. مراحل ۴ و ۵ بارها و بارها تکرار می‌شوند.
۷. اگر آماره آزمون برای داده‌های اصلی ۹۵ درصد از مقادیر آماره برای داده‌های جایگشتی بزرگ‌تر باشد فرض صفر را می‌توان در سطح ۵٪ رد کرد.

۳.۱ درخت تصمیم

درخت تصمیم یک روش ساده و توانمند برای رده‌بندی یک مجموعه داده به رده‌های متمایز و همگن است که یک گراف غیرچرخشی شبیه درخت دارد. گراف درخت تصمیم

^{۱۰} Permutation test

^{۱۱} Jackknife

از سه جزء اصلی ریشه، گره داخلی^{۱۲} و گره خارجی^{۱۳} (برگ) تشکیل شده است و روش کار بدین گونه است که ابتدا یک متغیر توضیحی به عنوان ریشه انتخاب می‌گردد و با توجه به تعدادی از سؤالات و ویژگی‌ها به چندین گره داخلی تقسیم می‌شود. به تقسیمات ایجادشده یک زیردرخت می‌گویند. به عنوان مثال مشاهدات بر اساس گروه‌های سنی مختلف به دو یا چند گروه تقسیم می‌شوند یا اینکه بر اساس جنسیت به دو گروه تقسیم می‌شوند. هر گره داخلی نیز مانند ریشه به گره‌های دیگری تقسیم می‌شود تا در نهایت به هر گره یک رده از متغیر پاسخ منتسب می‌شود که این گره‌ها، برگ نامیده می‌شوند [۷، ۱۱، ۳۸].

درخت تصمیم در موارد فراوانی، از جمله تشخیص و پیش‌گویی در علوم پزشکی، بازرگانی، تصمیم‌گیری در زمینه‌ی آب و هوا و غیره به کار برده می‌شود، به جرأت می‌توان گفت که مهم‌ترین جایگاه و کاربرد درخت تصمیم، در داده‌کاوی است. معمولاً برای معرفی یک درخت تصمیم، مشاهدات موجود به صورت تصادفی به دو دسته نمونه یادگیری (آموزشی) و نمونه آزمون تقسیم‌بندی می‌شوند. از نمونه‌ی اول برای ساخت درخت تصمیم و از نمونه‌ی دوم برای بررسی دقت درخت معرفی‌شده، استفاده می‌گردد.

۱.۳.۱ الگوریتم‌های مطرح درخت تصمیم

با توجه به اینکه شاخص‌های متنوع و روش‌های گوناگونی برای معرفی یک درخت تصمیم به کار برده می‌شوند، الگوریتم‌های متنوعی بدین منظور ارائه شده‌اند. مهم‌ترین و شناخته شده‌ترین این الگوریتم‌ها عبارتند از:

۱. AID^{۱۴}

۲. CHAID^{۱۵}

۳. ID3^{۱۶}

۴. C4.5

۵. درخت رگرسیون و رده‌بندی^{۱۷} (CART)

می‌توان از CART به عنوان متداول‌ترین آن‌ها نام برد [۱۱، ۲۶]. در الگوریتم CART متغیر پاسخ می‌تواند از نوع کمی یا کیفی باشد که به ترتیب قابلیت رگرسیونی و رده‌بندی آن را به نمایش می‌گذارد. در این پایان‌نامه بحث ما صرفاً متمرکز بر قابلیت رده‌بندی این

^{۱۲} Node

^{۱۳} Terminal Node

^{۱۴} Automatic Interaction Detection

^{۱۵} Chi-squared Automatic Interaction Detector

^{۱۶} Induction Decisin Tree

^{۱۷} Classification and regression tree

الگوریتم است. در بخش ۲.۳.۱ الگوریتم کلاسیک CART را برای متغیر پاسخ کیفی از جنس “اسمی” تحت عنوان درخت تصمیم مورد بررسی قرار می‌گیرد. در بخش ۴.۱ گونه دیگری از درخت تصمیم را برای متغیر پاسخ رتبه‌ای که موضوع این پایان‌نامه است را تحت عنوان “درخت استنباط شرطی” مطالعه خواهیم کرد.

۲.۳.۱ CART

در درخت تصمیم (CART) فضای متغیرهای توضیحی به‌طور سلسله‌مراتبی به فضاهای کوچک تقسیم می‌شود به‌طوری که مشاهدات واقع در هر ناحیه دارای تشابه حداکثری باشند. برای تشریح درخت تصمیم، شکل ۱.۱ را که در آن فضای p بعدی متغیرهای توضیحی در نقش t_P یا t_{Parent} ظاهر شده است، در نظر بگیرید. در اینجا فرض می‌شود $Z = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ مجموعه‌ی داده‌ها باشد که در آن $\{x_i = (x_{i1}, \dots, x_{ip}); i = 1, \dots, n\}$ است. این فضا بایستی در راستای محور یکی از متغیرها و از نقطه‌ای به دو زیرفضا تقسیم شود. پیدا کردن بهترین راستا و مناسب‌ترین نقطه‌ی تقسیم، اصلی‌ترین موضوع در رشد درخت است. واضح است که همه متغیرهای توضیحی و همه‌ی مقادیر آن می‌توانند به ترتیب نامزدی برای تعیین بهینه راستا و محل تقسیم فضا باشند. به طور قطع بایستی برای هر فرم تقسیم فضا، اندازه‌ای واحد تعریف کرد تا بهینگی مذکور بر اساس آن اندازه تعیین شود. فرض کنید راستای متغیر X_j و مشاهده x_{cj} یکی از گزینه‌های مورد نظر باشد به‌طوری که فضای متغیرها در مسیر X_j و از نقطه x_{cj} به دو فضای p بعدی راست t_R و چپ t_L تقسیم شده است. در اینجا تابعی موسوم به تابع ناخالصی^{۱۸} $i(t)$ تعریف شده و تغییرات این تابع به صورت (۳.۱) در نظر گرفته می‌شود:

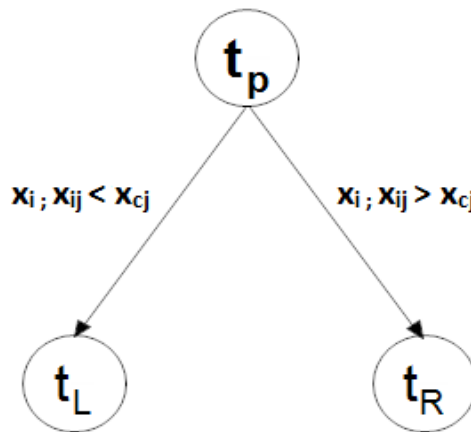
$$\Delta(X_j, x_{cj}) = i(t_p) - E(i(t_p)) = i(t_p) - p_L i(t_L) - p_R i(t_R) \quad (3.1)$$

که در آن t_L ، t_R و t_p به ترتیب فضاهای چپ، راست و فضای اصلی و p_L و p_R احتمال تعلق مشاهدات به فضاهای چپ و راست هستند به‌طوری که $p_L + p_R = 1$. در عمل این مقادیر عبارتند از $p_L = \frac{n(t_L)}{n(t_P)}$ و $p_R = \frac{n(t_R)}{n(t_P)}$. این فرآیند را می‌توان برای همه متغیرها و مقادیر آن‌ها انجام داد و $\Delta(X_j, x_{cj})$ مربوطه را به دست آورد. سرانجام افرازی که بیشترین تغییرات یعنی $\Delta(X_j, x_{cj})$ را نتیجه دهد، به عنوان بهترین راستای افراز انتخاب می‌شود [۲۶، ۳۲، ۳۸، ۳۹]. به عبارت دیگر فضا در راستای متغیر X_l و از نقطه‌ی a افراز می‌شود اگر

$$x_{(la)} = \arg \max_{c,j} \Delta(X_j, x_{cj}).$$

بدین صورت، اولین شاخه‌بندی درخت شکل می‌گیرد. برای ایجاد شاخه‌های جدید هریک

^{۱۸} Impurity Function



شکل ۱.۱: نمودار افراز فضای مشاهدات در یک درخت تصمیم

از گره‌های t_L و t_R را به عنوان t_p جدید تصور کرده و روند مشابه برای افرازهای جدید t_R و t_L به طور مجزا تکرار می‌شود. رشد سلسله‌مراتبی درخت تا مرحله‌ای ادامه می‌یابد که مشاهدات هر ناحیه دارای تشابه حداکثری باشند. شاخص‌های متعددی برای تابع ناخالصی $i(t)$ معرفی شده است که به معرفی آن‌ها خواهیم پرداخت.

شاخص جینی

این شاخص توسط بریمن و همکارانش در سال ۱۹۸۴ در مدل‌های درختی با تقسیمات دوتایی در هر گره معرفی شد. شاخص جینی به صورت (۴.۱) تعریف می‌شود:

$$i(t) = 1 - \sum_{j=1}^K p_{j|t}^2 \quad (4.1)$$

که K تعداد رده‌ها و $p_{j|t}$ احتمال آن است که مشاهده‌ی به تصادف انتخاب شده در فضای t ، متعلق به رده‌ی j باشد [۶، ۳۸].

شاخص آنتروپی

شاخص آنتروپی کاربردهای زیادی در داده‌کاوی دارد و برای ساخت درخت رده‌بندی مورد استفاده قرار می‌گیرد که به صورت (۵.۱) تعریف می‌شود:

$$i(t) = - \sum_{j=1}^K p_{j|t} \log_2 p_{j|t}. \quad (5.1)$$

مثال ۱.۳.۱. نمودار پراکنش ۴۰ مشاهده در شکل ۲.۱ نمایش داده شده است. در این نمودار با فرض دو متغیر توضیحی و سه رده پاسخ قرمز، سبز و آبی، تعداد مشاهدات مربوط به این رده‌ها به ترتیب ۱۷، ۱۱ و ۱۲ هستند. فرض کنید فضای متغیرهای X_1 و X_2 در راستای محور X_1 و از نقطه‌ای a به دو قسمت t_L و t_R تقسیم شده باشد (شکل ۳.۱). مقادیر تابع ناخالصی i برای هر یک از گره‌های فوق به صورت زیر محاسبه می‌شوند:

$$i(t_P) = 1 - \sum_{j=1}^k p_{j|t}^2 = 1 - ((\frac{17}{40})^2 + (\frac{11}{40})^2 + (\frac{12}{40})^2) = 0.654$$

$$i(t_R) = 1 - \sum_{j=1}^k p_{j|t}^2 = 1 - ((\frac{7}{21})^2 + (\frac{10}{21})^2 + (\frac{4}{21})^2) = 0.626$$

$$i(t_L) = 1 - \sum_{j=1}^k p_{j|t}^2 = 1 - ((\frac{10}{19})^2 + (\frac{2}{19})^2 + (\frac{7}{19})^2) = 0.576$$

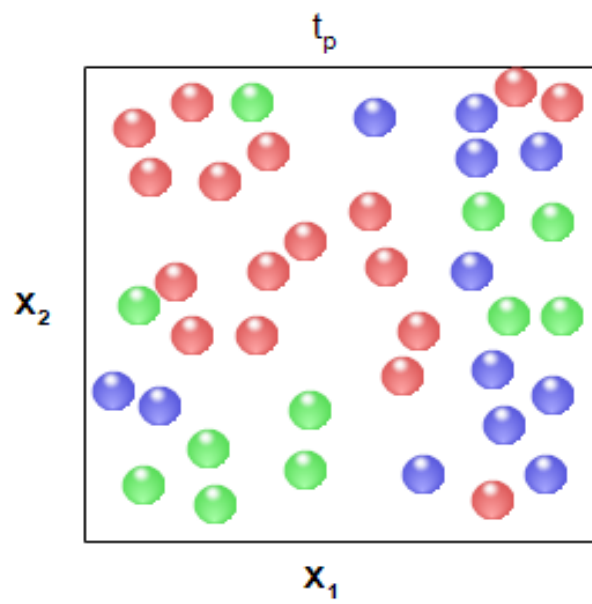
$$\Delta(X_1, a_1) = i(t_P) - p_R i(t_R) - p_L i(t_L)$$

$$= 0.624 - (\frac{21}{40})(0.626) - (\frac{19}{40})(0.576) = 0.51$$

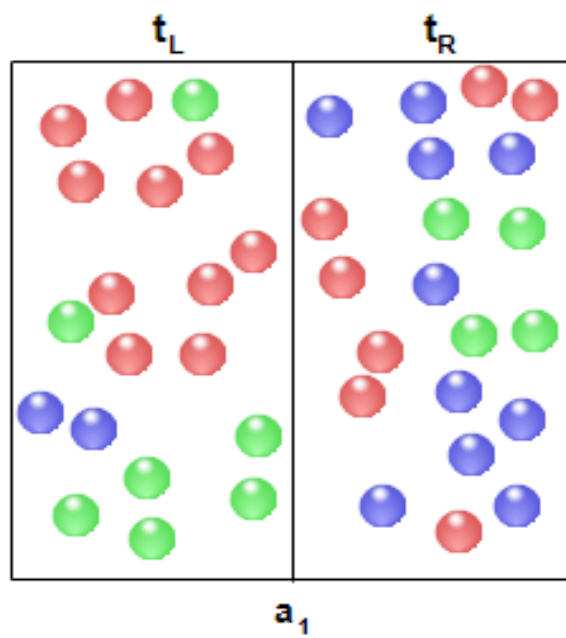
برای تعیین راستا و نقطه‌ای افراز، مقادیر $\Delta(X_1, a_1)$ و $\Delta(X_2, a_2)$ به ازای همه‌ی مقادیر ممکن a_1 و a_2 محاسبه می‌شوند. سرانجام بیشترین مقدار Δ ، تعیین‌کننده‌ی اولین افراز است و در شکل ۴.۱ اولین افراز در مسیر X_1 و از نقطه‌ی a انجام شد. لذا

$$(1, a) = \arg \max_{i,j} \Delta(X_j, x_{ij}) \quad j = 1, 2 \quad i = 1, \dots, 40.$$

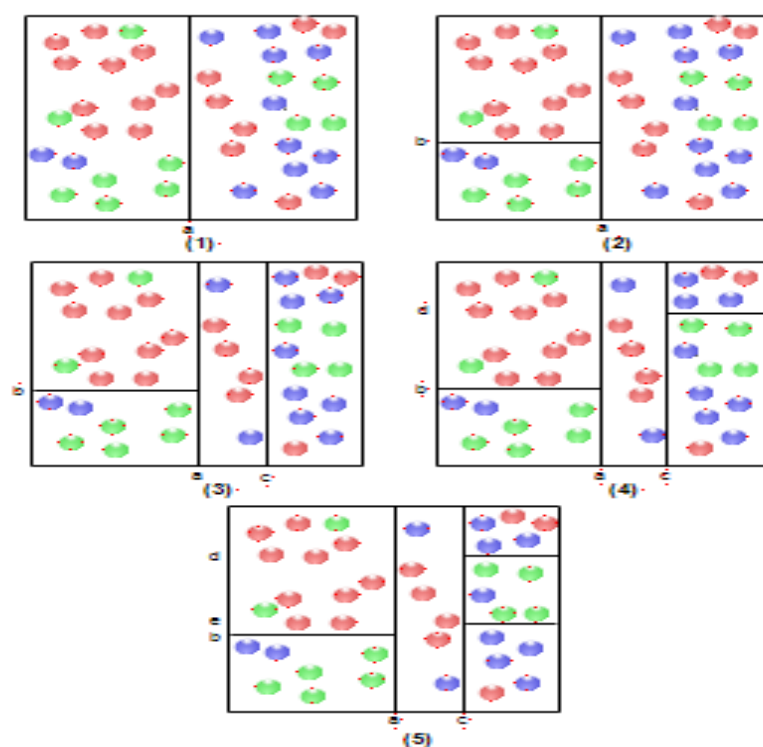
با انجام سلسله‌مراتبی افراز فضاهای ایجادشده، درخت تصمیم رشد خواهد کرد. نمایش شماتیکی افراز فضا یا رشد درخت در شکل‌های ۴.۱ و ۵.۱ آمده است.



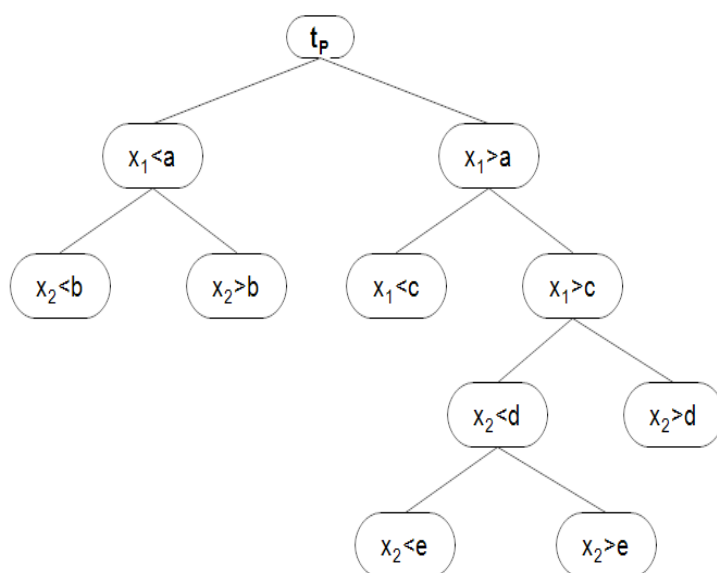
شکل ۲.۱: نمودار پراکنش مشاهدات



شکل ۳.۱: نمودار حاصل از انتخاب اولین افراز



شکل ۴.۱: مراحل افراز فضای مشاهدات



شکل ۵.۱: ساختار درختی مشاهدات

۳.۳.۱ پیش‌بینی درخت تصمیم

هر مشاهده‌ی جدید x_0 ، بر حسب مختصات خود یعنی $(x_{01}, \dots, x_{0p})$ در یکی از زیرفضاهای افراز شده (برگ‌ها)، واقع می‌شود. در این زیرفضا، مشاهداتی وجود دارند که مقدار پاسخ هر یک از آن‌ها مشخص است. رده‌ی تخمین‌شده برای مشاهده‌ی x_0 همان رده‌ی غالب آن زیرفضا، یا به عبارت دیگر مُد رده‌های آن زیرفضا است.

۴.۳.۱ دقت درخت تصمیم

بدیهی است که در تمام تکنیک‌های رده‌بندی، دستیابی به الگویی که دارای خطا نباشد، غیر ممکن است ولی هدف این است که الگویی با کمترین خطا معرفی گردد. از این رو یکی از مهم‌ترین نکات مورد توجه در معرفی یک الگوی درختی، بررسی دقت و کارایی آن است. هدف از معرفی یک درخت رده‌بندی، دستیابی به یک الگوی رده‌بندی براساس مشاهدات موجود است به طوری که این الگوی معرفی‌شده از دقت مناسبی برای مشاهدات جدید برخوردار باشد، یعنی اینکه بتواند در مورد یک مشاهده جدید به خوبی و با کمترین اشتباه ممکن تصمیم‌گیری نماید. بنابراین برای محاسبه‌ی میزان خطای درخت رده‌بندی از معیار نرخ خطا^{۱۹} (ER) استفاده می‌شود و به معنای نسبت مشاهداتی است که رده‌ی تخمینی آن‌ها با رده‌ی واقعی متفاوت است که به صورت (۶.۱) است:

$$ER = \frac{1}{n} \sum_{i=1}^n I(Y_i \neq \hat{Y}_i) \quad (6.1)$$

در اینجا $I(\cdot)$ تابع نشانگر، Y_i رده‌ی واقعی و $\hat{Y}_i$ رده‌ی پیش‌بینی‌شده‌ی مشاهدات است. درختی که مقدار ER کمتری داشته باشد از دقت بالایی برای رده‌بندی برخوردار است.

۵.۳.۱ توقف ساخت درخت

فرآیند ساختن درخت تا جایی ادامه می‌یابد که تنها یک مشاهده در هر گره وجود داشته باشد یا ضابطه افراز بوسیله حدود مشخص‌شده توسط تحلیل گر، متوقف گردد. البته گاهی این حدود مشخص‌شده به اندازه کافی کوچک نیست که این امر موجب می‌شود داده‌ها به خوبی افراز نگردند و درخت، اطلاعات موجود در داده‌ها را به خوبی برازش ندهد. به عبارت دیگر گره‌های پایانی بصورت بالقوه قابلیت افراز را خواهند داشت. بطور معمول با رسیدن به بزرگترین درخت، فرآیند افراز متوقف می‌شود. در این حالت پدیده‌ی بیش‌برازش رخ می‌دهد. راهکارهای متعددی برای توقف رشد یک درخت رده‌بندی به کار برده می‌شود که می‌توان به موارد ذیل اشاره نمود:

^{۱۹} Error Rate

- هنگامی که در گره، تمامی مشاهدات آن به یک رده تعلق داشته باشند.
 - در حالتی که تعداد مشاهدات درون یک گره از یک مقدار آستانه کمتر شود، تقسیمات دیگری بر روی گره انجام نخواهد شد. وقتی تعداد مشاهدات در یک گره بسیار اندک گردد بهتر است از ادامه رده‌بندی صرف‌نظر نماییم.
 - هنگامی که افزایش اندازه درخت در گره مربوطه باعث افزایش چشم‌گیری در دقت آن درخت نشود. در صورتی که تقسیمات بیشتر در یک گره موجب نشود که نرخ خطا تغییر چندانی نماید، آنگاه در نظر گرفتن این تقسیمات فقط موجب معرفی یک درخت بزرگتر می‌شود نه درختی با دقت بیشتر.
- با توجه به روش‌های ذکرشده درمی‌یابیم این روش‌ها به گونه‌ای عمل می‌نمایند که از یک نقطه به بعد، دیگر تقسیمات بر روی گره‌ها انجام نخواهد شد. استفاده از این روش‌ها دو مزیت دارد، اول اینکه مانع از انجام محاسبات اضافی شده و دوم اینکه دقت درخت معرفی‌شده بهتر از درخت کامل خواهد بود [۳۸، ۱۳].

۶.۳.۱ روش هرس درخت تصمیم

در درخت تصمیم به منظور ممانعت از بیش‌برازشی و دستیابی به درختی با اندازه مناسب، پس از رشد کامل درخت، عمل هرس انجام می‌شود. بدین منظور از روشی به نام هزینه‌ی پیچیدگی هرس^{۲۰} استفاده می‌شود [۳۱، ۳۳، ۱۳]. روش اعتبارسنجی متقابل رابطه‌ی بین پیچیدگی درخت و خطای رده‌بندی است به عبارت دیگر با افزایش اندازه‌ی درخت خطای رده‌بندی کاهش یافته، در نتیجه دقت رده‌بندی بالا می‌رود. بنابراین راهکار آن دستیابی به درخت بهینه است. هزینه‌ی پیچیدگی هرس براساس رابطه (۷.۱) محاسبه می‌شود:

$$ER_{\alpha}(T) = ER(T) + \alpha(\tilde{T}) \quad (7.1)$$

که در آن $ER(T)$ خطای رده‌بندی درخت T ، $\tilde{T}$ تعداد گره‌های ترمینال درخت T و α عامل جریمه است. الگوریتم هرس براساس هزینه‌ی پیچیدگی به صورت زیر انجام می‌شود:

- گره‌ی t از آخرین گره‌ی داخلی انتخاب می‌شود.
- هزینه‌ی پیچیدگی برای هر دو حالت زیر محاسبه می‌شود:
 ۱. حالتی که درخت شامل گره‌ی t است.
 ۲. حالتی که گره‌ی t از درخت حذف شده است.
- اگر هزینه‌ی محاسبه‌شده در حالت ۱ بیشتر از حالت ۲ باشد، گره‌ی t از درخت حذف می‌شود، در غیر این صورت درخت رده‌بندی هرس نمی‌شود.

^{۲۰}Cost complexity pruning

۷.۳.۱ مزایا و معایب درخت تصمیم

استفاده از الگوهای درختی در مقایسه با روش‌های دیگر دارای مزیت‌های متنوعی هستند و در کنار این مزیت‌ها، معایبی را نیز دارند [۱۳]. در اینجا به طور مختصر به بحث در این مورد خواهیم پرداخت.

● مزایا

۱. درخت تصمیم یک روش ناپارامتری است و نیازی به تعیین فرم تابعی خاصی برای مدل ندارد.
۲. زمانی که تعداد متغیرها زیاد باشد، این روش بسیار مناسب است.
۳. درخت تصمیم یک نمودار قابل درک و ساده برای رده‌بندی معرفی می‌کند و نیازی به معرفی یک ساختار تصمیم‌گیری ندارد.
۴. درخت تصمیم نسبت به داده‌های پرت حساس نیست و در اصطلاح تنومند است.
۵. نتایج CART نسبت به تغییرات یکنوا در متغیرهای توضیحی، تغییرناپذیر است.

● معایب

۱. برای متغیر پاسخ رتبه‌ای مناسب نیست، زیرا ماهیت "رتبه" را در نظر نمی‌گیرد و ممکن است دقت رده‌بندی کاهش یابد.
۲. این روش معمولاً برای داده‌هایی با حجم کم مناسب نیست.
۳. تولید درخت و رشد آن، هزینه محاسباتی بالایی دارد.
۴. روش CART، نسبت به تغییر داده‌های آموزشی پایدار نیست به‌طوری‌که حذف چند مشاهده از داده‌ها، درخت را تغییر داده و ممکن است منجر به تغییر انتخاب در متغیری شود که در راستای آن، قرار است فضا افزاز شود.

۴.۱ درخت استنباط شرطی

در حالتی که متغیر پاسخ از نوع اسمی است، درخت تصمیم می‌تواند رهیافتی مناسب برای برازش مجموعه داده‌ها باشد. اما هنگامی که متغیر پاسخ از ساختار رتبه‌ای تبعیت می‌کند، رتبه‌ی متغیرها در مدل‌سازی لحاظ نمی‌شود و آن‌ها را به عنوان متغیر اسمی در نظر گرفته که باعث نادیده گرفتن اطلاعات رتبه‌ها می‌شود. در چنین وضعیتی، از درخت استنباط شرطی که مبنای آن، میزان همبستگی متغیر پاسخ رتبه‌ای با متغیرهای توضیحی است استفاده می‌شود [۱۶، ۱۷، ۳۵].

یادآوری می‌شود که در موضوع رشد درخت همواره با دو مسئله مواجه هستیم، که در راستای کدام متغیر و در چه مقداری، فضای متغیرها به دو قسمت تقسیم شود. در درخت کلاسیک، برای ایجاد شاخه جدید روی هر شاخه از درخت، ابتدا جهت راستا و نقطه‌ای افراز، همه‌ی متغیرهای توضیحی مورد بررسی قرار می‌گیرند. متغیری از بین متغیرهای توضیحی $X_1, \dots, X_p$ انتخاب می‌شود که دارای بیشترین میزان تغییرات تابع ناخالصی باشد. تفاوت عمده درخت استنباط شرطی با درخت تصمیم کلاسیک در نحوه‌ی انتخاب متغیر برای افراز فضا است (بخش ۲.۳.۱ را ببینید).

۱.۴.۱ انتخاب متغیر برای افراز فضا

در درخت استنباط شرطی که متغیر پاسخ از نوع رتبه‌ای است، نحوه‌ی افراز فضای متغیرهای توضیحی، متفاوت از درخت کلاسیک است. در این روش برای افراز فضا، متغیری از بین متغیرهای توضیحی انتخاب می‌شود که دارای بیشترین همبستگی با متغیر پاسخ رتبه‌ای Y باشد. به عبارت دیگر برای انتخاب متغیر، فرضیه‌ی استقلال متغیرهای توضیحی و متغیر پاسخ رتبه‌ای مورد آزمون قرار داده می‌شود که در آن رابطه بین هر یک از متغیرهای توضیحی و متغیر پاسخ مورد بررسی قرار می‌گیرد و متغیری که بیشترین همبستگی را با متغیر پاسخ داشته باشد، جهت افراز فضا انتخاب می‌شود (بخش ۲.۱ را ببینید). در هر گره که با یک W نشان داده می‌شود، فرض کلی $H_0 = \bigcap_{j=1}^m H_0^j$ استقلال بر حسب m فرض جزئی $H_0^j = D(Y|X_j) = D(Y)$ آزمون می‌شود که آماره‌ی آزمون یا میزان رابطه‌ی بین Y و X_j که $j = 1, \dots, m$ توسط آماره‌ی خطی به فرم (۸.۱) تعریف می‌شود

$$T_j = \sum_{i=1}^n w_i g_j(X_{ij}) s(Y_i) \quad (۸.۱)$$

در اینجا X_{ij} معرف مشاهده i ام متغیر j ام، $g_j : X_j \rightarrow \mathbb{R}^{p_j}$ یک تبدیل غیرتصادفی از متغیر j ام، $s(Y_i)$ نمرات سطوح پاسخ و w_i وزن مشاهدات است. هر گره از درخت با یک بردار وزنی نشان داده می‌شود، که در آن وزن‌ها متناظر با مشاهدات عضو گره هستند، به‌طوری که $w_i = 1$ اگر مشاهده عضو گره باشد و در غیراینصورت $w_i = 0$.

مقادیر بزرگ آماره‌ی T مبین همبستگی بیشتر متغیر پاسخ با متغیر توضیحی و مقادیر کوچک آن نشان از همبستگی اندک هستند. از آنجا که مقادیر بزرگ T موجب رد فرض H_0 می‌شوند، می‌توان نتیجه گرفت که هر چقدر p - مقدار کمتر باشد همبستگی پاسخ با X_j بیشتر است. بنابراین براساس آزمون استنباط شرطی رابطه‌ی بین متغیرهای توضیحی و متغیر پاسخ رتبه‌ای مورد آزمون قرار می‌گیرند و مقدار آماره‌ی آزمون و همچنین p - مقدار به ازای متغیرهای توضیحی حاصل می‌شود و در نهایت متغیری که p - مقدار کمتری دارد، در واقع بیشترین ارتباط را با Y خواهد داشت و به عنوان راستای افراز انتخاب می‌شود [۱۶، ۱۷].

۲.۴.۱ انتخاب نقطه افراز

پس از تعیین متغیری که باید فضای متغیرها در راستای آن به دو قسمت تقسیم شود، سوال بعدی این است که از چه نقطه‌ای، این افراز باید صورت گیرد. طبق اصول درخت استنباط شرطی، نقطه تقسیم باید به گونه‌ای انتخاب شود که تمایز بین مشاهدات در گروه ایجاد شده به حداکثر مقدار خود برسد. برای هر یک از مجموعه‌های ایجاد شده A_1 و A_2 متناظر با گره‌های t_L و t_R (شکل ۱.۱) مجموع نمرات اختصاص داده شده به مقادیر پاسخ هر گروه محاسبه شده و تفاوت آن‌ها محاسبه می‌شود و سرانجام نقطه بهینه برای افراز فضا نقطه‌ای است که تفاوت مذکور به حداکثر مقدار خود برسد. با فرض اینکه متغیر منتخب X_j برای افراز فضا تعیین شده باشد، فرض کنید $T_{A_1,c} = \sum_{i=1}^n s(Y_i)I_{A_1}(X_{ij})$ و $T_{A_2,c} = \sum_{i=1}^n s(Y_i)I_{A_2}(X_{ij})$ مجموع رتبه‌های پاسخ در دو مجموعه A_1 و A_2 برای هر نقطه تصمیمی چون c باشند. نقطه c^* را نقطه تقسیم بهینه نامند، اگر

$$c^* = \arg \max_c |T_{A_1,c} - T_{A_2,c}|.$$

بدین صورت در هر مرحله، افراز فضا صورت می‌گیرد و درخت رشد می‌کند [۱۶، ۱۷].

۳.۴.۱ ضابطه توقف

در روش پیشنهادی در هر مرحله، با انجام آزمون، افراز فضا صورت می‌گیرد و درخت رشد می‌کند. رشد درخت و مرحله‌ای افراز، زمانی متوقف می‌شود که نتوان فرض استقلال بین متغیر پاسخ رتبه‌ای و هر متغیر پیش‌گو را رد کرد. نکته مهمی که در این روش آماری وجود دارد، اطمینان از ساخت یک درخت با اندازه مناسب است، بنابراین نیازی به فرایند هرس یا هر روش دیگری برای انتخاب اندازه مناسب و درست درخت وجود ندارد.

۴.۴.۱ پیش‌بینی درخت استنباط شرطی

پیش‌بینی مشاهدات جدید در درخت استنباط شرطی همانند پیش‌بینی درخت تصمیم انجام می‌گیرد. همانطور که گفته شد مشاهداتی که در هر یک از افرازهای ایجاد شده (برگ)، قرار دارند طبق فراوانی رده‌های پاسخ آن‌ها، رده‌ای به این فضاها نسبت داده می‌شود. مشاهده جدید بر حسب مختصات خود، در یکی از این نواحی قرار می‌گیرد. بنابراین رده‌ی نسبت داده شده برای این ناحیه، رده‌ی مشاهده‌ی جدید می‌باشد.

۵.۴.۱ الگوریتم درخت استنباط شرطی

برای تشکیل درخت استنباط شرطی برای مجموعه مشاهدات به حجم n در هر مرحله از افراز، گام‌های زیر انجام می‌شوند:

- انجام آزمون استقلال متغیرهای توضیحی و پاسخ.
 - در صورت پذیرش فرض استقلال، الگوریتم متوقف شده، در غیر اینصورت محاسبه‌ی p -مقدار به ازای همه‌ی متغیرهای توضیحی.
 - انتخاب راستای افراز (متغیر توضیحی دارای کمترین p -مقدار به عنوان راستای افراز انتخاب می‌شود).
 - محاسبه‌ی مجموع رتبه‌ی پاسخ مشاهدات برای هر دو زیرگروه‌ی ممکن.
 - انتخاب نقطه‌ی افراز (نقطه‌ای که منجر به بیشترین اختلاف ممکن مجموع رتبه‌ها، در دو زیرگروه شود).
- افراز سلسله‌مراتبی فضا تا زمانی ادامه می‌یابد که متغیر توضیحی مستقل از متغیر پاسخ باشد. بدین صورت درخت با اندازه‌ی مناسب رشد می‌کند و نیازی به هرس ندارد.

فصل ۲

جنگل تصادفی

در فصل قبل موضوع رده‌بندی را در قالب درخت تصمیم، برای متغیر پاسخ اسمی و رتبه‌ای بررسی کردیم. اکنون این موضوع را با در نظر گرفتن اجماع درختان، تحت عنوان جنگل‌های تصادفی مورد بحث قرار خواهیم داد. جنگل‌های تصادفی مصداقی از روش رده‌بندی اجماعی و اجماع‌سازی خودگردان^۱ (Bagging) هستند. در این روش دقت و عملکرد پیش‌بینی افزایش یافته و همچنین متغیرهایی که نقش بسزایی در رده‌بندی دارند، مشخص می‌شوند. مجموعه‌ی مشاهدات (داده‌ها) شامل $\mathbf{Z} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ که $x_i = (x_{i1}, \dots, x_{ip})$ و $i = 1, \dots, n$ است. در این فصل ابتدا مفاهیم روش رده‌بندی اجماعی و اجماع‌سازی خودگردان را شرح داده و سپس مبحث اصلی جنگل تصادفی را در حالت کلاسیک و برای متغیر پاسخ رتبه‌ای، مطرح می‌کنیم.

۱.۲ روش رده‌بندی اجماعی

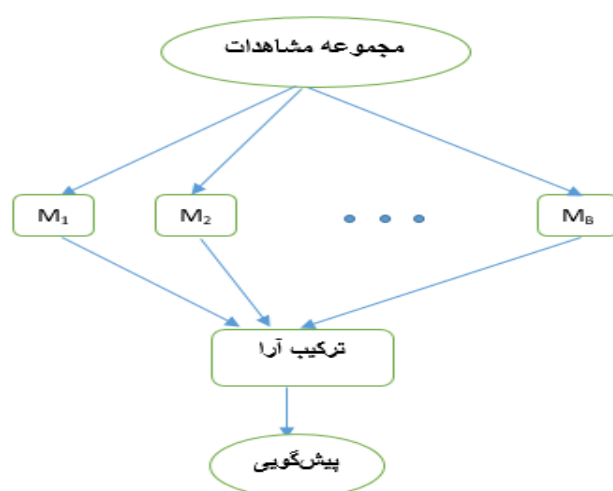
رده‌بندی اجماعی شامل مجموعه‌ای از مدل‌های رده‌بندی است که هدف آن افزایش دقت رده‌بندی است. در این روش مدلی برای رده‌بندی ارائه می‌شود که ترکیبی از چندین روش رده‌بندی است و روش کار آن، بدین صورت است که هر مدل مبنای طبق‌فرآیندش رده‌ای را به مشاهده‌ی جدید نسبت می‌دهد [۱۳]. سپس مُد رده‌ها، رده‌ی موردنظر برای مشاهده‌ی

^۱Bootstrap aggregating

جدید می باشد که این عمل باعث افزایش دقت رده بندی می شود.

شکل ۱.۲ ساختار رده بندی اجماعی را نشان می دهد که در آن از B مدل پایه $M_1, \dots, M_B$ برای ایجاد مدل اجماعی استفاده می شود که هر یک از این مدل های مبنا روش های مختلفی هستند که هر چه تنوع میان مدل های مبنای روش های اجماعی بیشتر باشد نتایج بهتری از خود نشان می دهند.

در رده بندی مشاهده ی جدید، تمامی مدل های مبنا شرکت خواهند داشت به طوری که هر یک از مدل ها، رده ای را به مشاهده ی جدید نسبت می دهند و در نهایت مُد رده های تعیین شده، رده ی مشاهده ی جدید است.



شکل ۱.۲: ساختار رده بندی اجماعی

روش های اجماع سازی خودگردان و جنگل های تصادفی، نمونه هایی از روش اجماعی محسوب می شوند [۱۳].

۱.۱.۲ اجماع سازی خودگردان

روش اجماع سازی خودگردان یا Bagging رهیافتی برای کاهش واریانس تابع پیش گوی برآورد شده است. این روش همانند روش رده بندی اجماعی، ترکیبی از چند مدل مبنای رده بندی است با این تفاوت که همه ی مدل های مبنای آن، یکسان هستند و داده های آموزشی هر یک از آن ها، یک نمونه ی خودگردان از داده های اصلی است [۱۳]. در رده بندی مشاهده ی جدید، هر یک از مدل ها، رده ای را به مشاهده ی جدید نسبت می دهد و در نهایت، مُد رده های تعیین شده، رده ی مشاهده ی جدید است.

روش جنگل تصادفی یک روش رده بندی اجماعی است که در آن مدل های مبنا توسط درخت رده بندی و رگرسیون تشکیل می شوند.

۲.۲ جنگل تصادفی

جنگل تصادفی نقش برجسته‌ای از اجماع‌سازی خودگردان را به نمایش می‌گذارد که براساس مجموعه‌ی بزرگی از درخت‌های غیروابسته ساخته شده و متوسط‌گیری می‌شود. درخت‌ها، مدل‌های مبنای مناسبی برای اجماع‌سازی خودگردان هستند زیرا می‌توانند بر ساختارهای پیچیده در داده‌ها فائق آیند و در صورت رشد مناسب درخت، اریبی نسبتاً کمی داشته باشند [۵، ۱۳].

ایده‌ی جنگل‌های تصادفی آن است که بتواند به وسیله کاهش همبستگی بین درخت‌ها، واریانس روش اجماع‌سازی خودگردان را کاهش دهد. این مهم از طریق انتخاب تصادفی متغیرهای توضیحی در هر مرحله از رشد درخت به دست می‌آید و نحوه آن بدین صورت است که “قبل از هر بار افراز فضا، مجموعه‌ای m تایی ($m = \sqrt{p} < p$) از متغیرهای توضیحی انتخاب شود و راستای بهینه برای افراز، فقط در بین این m متغیر مورد جستجو واقع شود”.

لازم به یادآوری است که بر اساس ماهیت اجماع‌سازی خودگردان، هر درخت دارای داده‌های آموزشی مخصوص خود است که این داده‌ها یک نمونه تصادفی با جایگذاری از داده‌های آموزشی اصلی است. بنابراین تعدادی از مشاهدات، در داده‌های آموزشی هر درخت غایب هستند که آن‌ها را داده‌های خارج از کیسه^۲ OOB نامند و لذا هر درخت دارای داده‌های OOB خاص خود است که نقش داده‌های آزمون را ایفا کرده و برای تعیین میزان اهمیت متغیرها به کار می‌روند [۵، ۱۲].

۱.۲.۲ الگوریتم جنگل تصادفی کلاسیک

با فرض اینکه جنگل تصادفی شامل تعداد B درخت است، برای $b = 1, \dots, B$ مراحل زیر انجام می‌شوند:

۱. یک نمونه خودگردان $data_b$ به اندازه n از داده‌های اصلی انتخاب می‌شود. مشاهدات انتخاب نشده را، OOB_b می‌نامیم.

۲. درخت T_b تحت نمونه‌ی خودگردان مذکور، با شرایط زیر رشد می‌کند. در هر مرحله از افراز یا در هر گره

- تعداد m متغیر از کل متغیرها (p) به تصادف انتخاب می‌شود.
- جستجو برای افراز بهینه و نقطه بهینه فقط در بین این m متغیر موجود و براساس اصول درخت کلاسیک صورت می‌گیرد.

^۲ Out Of Bag

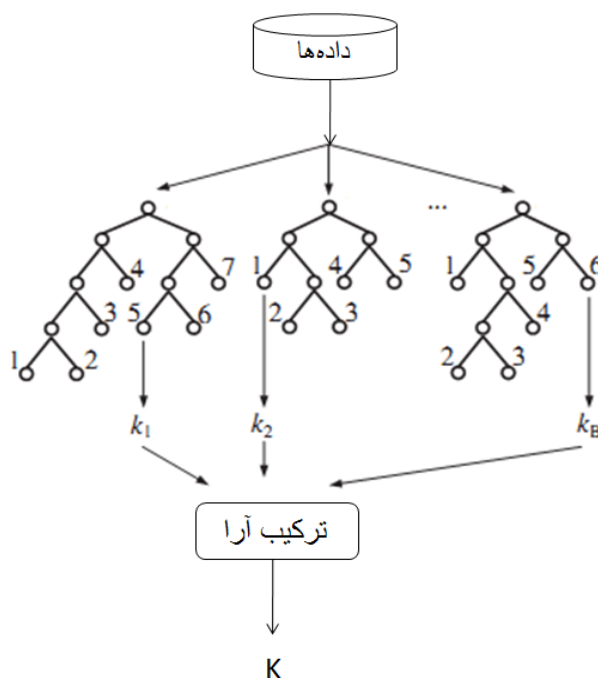
• گرهی مذکور به دو قسمت تقسیم می‌شود.

افراز سلسله‌مراتبی فضا تا زمانی ادامه می‌یابد که در هر برگ، تعداد از قبل تعیین شده n_{min} مشاهده قرار گیرد. با انجام این مراحل، اجماعی از درختان $T_1, T_2, \dots, T_B$ ایجاد می‌شود که به دلیل تفاوت آن‌ها در حوزه‌ی جستجوی متغیرها برای افراز فضا، بسیار متنوع بوده و نتایج بهتری مطابق اصول رده‌بندی اجماعی از خود نشان خواهد داد.

۲.۲.۲ رده‌بندی با جنگل تصادفی

هر مشاهده جدید بر حسب مختصات خود، در یکی از برگ‌های هر یک از درختان جنگل واقع می‌شود و هر درخت یکی از رده‌های متغیر پاسخ را برای آن مشاهده در نظر می‌گیرد. در نهایت مشاهده جدید به رده‌ای نسبت داده می‌شود که حداکثر آرا (مُد) را در بین تمام درخت‌ها کسب کرده باشد. شایان ذکر است که تعداد درخت‌های جنگل تصادفی، یک پارامتر است که با تغییر آن می‌توان مقدار بهینه را بر اساس مینیمم سازی خطای رده‌بندی، در نظر گرفت [۵، ۱۳].

شکل ۲.۲ ساختار جنگل تصادفی را برای تعداد B درخت تصمیم نشان می‌دهد که هر یک از درخت‌ها، رده‌ای برای مشاهده جدید نسبت می‌دهند و در نهایت نتایج تمامی درخت‌ها با هم ترکیب شده و رده‌ای با حداکثر فراوانی، به عنوان رده نهایی آن مشاهده، تعیین می‌شود.



شکل ۲.۲: ساختار جنگل تصادفی

۳.۲.۲ دقت رده‌بندی و پیش‌بینی جنگل تصادفی کلاسیک

مشاهدات مجموعه‌ی آزمون که در مدل شرکت ندارند، بر حسب مختصات خود در یکی از رده‌های هر یک از درختان جنگل واقع می‌شوند و هر درخت رده‌ای را به آن‌ها نسبت می‌دهد و با ترکیب آن‌ها درختان جنگل، رده‌ی این مشاهدات تخمین زده می‌شود. رده‌ی تخمین زده با رده‌ی اصلی مشاهدات مورد مقایسه قرار می‌گیرد و در نهایت خطای رده‌بندی با استفاده از معیار ER طبق رابطه‌ی (۶.۱) مورد ارزیابی قرار می‌گیرد. بنابراین هرچه مقدار خطا کمتر باشد، دقت رده‌بندی بیشتر است.

۴.۲.۲ اهمیت متغیرها در جنگل تصادفی

در رده‌بندی مشاهدات، تاثیر همه متغیرها به یک اندازه نیست. بعضی از آن‌ها تاثیر بسزایی در تعیین رده‌ی مورد نظر دارند و بعضی دیگر تاثیر بسیار اندک یا حتی بدون اثر هستند. بنابراین دانستن اینکه چه متغیرهایی در رده‌بندی موثر بوده‌اند، موجب تفسیر بهتر نتایج می‌شود. در روش جنگل تصادفی اندازه اهمیت متغیرها^۳ (VIM) را می‌توان با استفاده از داده‌های $OOB_b; b = 1, \dots, B$ در هر یک از درخت‌های جنگل تصادفی، به صورت فرآیند زیر محاسبه کرد [۲۳، ۵].

برای تعیین اهمیت متغیر j ام، (X_j) و برای درخت b ام ($b = 1, \dots, B$)، داده‌های OOB_b در نظر گرفته می‌شوند.

- خطای داده‌های آزمون OOB_b در درخت b ام، محاسبه می‌شود ($error_b$).
- جایگشت تصادفی از مقادیر متغیر X_j در داده‌های OOB_b ایجاد می‌شود.
- این جایگشت تصادفی، جایگزین مقادیر X_j شده و بدین صورت داده‌های جدیدی به نام OOB_{bp} ایجاد می‌شود.
- خطای داده‌های جدید OOB_{bp} محاسبه می‌شود ($error_{bp}$).
- میزان تاثیر تعویض مقادیر متغیر X_j برای هر یک از درختان جنگل، به صورت زیر محاسبه می‌شود

$$(error_b - error_{bp}) \quad b = 1, \dots, B$$

در نهایت، اهمیت متغیر X_j عبارت از

$$VI_j = \frac{1}{B} \sum_{b=1}^B (error_b - error_{bp})$$

^۳ Variable Importance Measures

واضح است که هرچقدر تفاوت خطاهای مذکور بزرگتر باشد، بدین معنا است که انجام جایگشت مقادیر X_j توانسته است نتیجه رده‌بندی را بیشتر تحت تاثیر قرار دهد و لذا X_j نقش بیشتری در رده‌بندی خواهد داشت و بالعکس، کوچک بودن این مقدار حاکی از تاثیر ناچیز X_j در رده‌بندی است.

۳.۲ جنگل تصادفی رتبه‌ای

همانطور که گفته شد، روش جنگل تصادفی یک ابزار معمول برای رده‌بندی داده‌ها در ابعاد بالا است به‌طوری‌که قادر به پیش‌بینی رده مشاهدات و تعیین رتبه متغیرهای پیش‌گو براساس میزان اهمیت آن متغیرها بوده و قابل استفاده برای متغیر پاسخ از نوع کمی، اسمی یا بقا است. اما وقتی متغیر پاسخ از نوع رتبه‌ای است، بطور معمول آن را به صورت اسمی در نظر می‌گیرد که باعث نادیده گرفتن اطلاعات رتبه‌ای و کاهش دقت پیش‌بینی می‌شود. بنابراین چگونگی مدل‌سازی متغیر پاسخ رتبه‌ای با رویکرد استفاده از روش مدرن و کارای جنگل‌های تصادفی مورد سوال است.

جنگل تصادفی رتبه‌ای با حفظ ساختار کلی حالت کلاسیک جنگل تصادفی، مدل پیش‌گویی مناسبی را برای متغیر پاسخ رتبه‌ای ارائه می‌دهد. در این روش، مدل‌های مبنا براساس درخت استنباط شرطی تشکیل می‌شوند (بخش ۴.۱ را ببینید) [۱۶]. یادآوری می‌شود که در موضوع رشد درخت همواره با دو مسئله مواجه هستیم، که در راستای کدام متغیر و در چه مقداری، فضای متغیرها به دو قسمت تقسیم شود. بدین منظور، برای ایجاد شاخه جدید روی هر شاخه از یک درخت جنگل تصادفی، ابتدا زیرمجموعه‌ای m تایی ($m < p$) از کل متغیرها به تصادف انتخاب شده و متغیر منتخب برای افراز فضا، متغیری است که دارای بیشترین $\Delta(x_j, a_i)$ باشد. تفاوت عمده جنگل تصادفی رتبه‌ای با جنگل تصادفی کلاسیک در همین جا یا به عبارتی نحوه انتخاب متغیر برای افراز فضا است [۲۰].

در جنگل تصادفی رتبه‌ای، برای افراز فضا در هر مرحله از رشد هر درخت، میزان همبستگی هر یک از m متغیر نامزد با متغیر پاسخ رتبه‌ای، اندازه‌گیری شده و متغیر منتخب برای تعیین راستای افراز فضا، متغیری است که دارای بیشترین همبستگی باشد. با انجام مکرر افراز در زیرفضاهای تولیدشده، همه درخت‌های جنگل تصادفی جدید رشد می‌کنند و زمانی این رشد متوقف می‌شود که نتوان فرض استقلال بین متغیر پاسخ و هر متغیر پیش‌گو را رد کرد [۲۳، ۲۷، ۳۷، ۴۰].

روند انتخاب داده‌های آموزشی در جنگل تصادفی رتبه‌ای متفاوت از جنگل تصادفی کلاسیک است. در جنگل تصادفی کلاسیک، در هر یک از مدل‌های مبنا، یک نمونه‌ی تصادفی با جایگذاری و با حجم n از مشاهدات اصلی انتخاب می‌شود. اما در جنگل تصادفی رتبه‌ای، یک نمونه‌ی تصادفی به روش بدون جایگذاری و با حجم n^* ($n^* < n$) از مشاهدات

اصلی انتخاب می‌شود [۲۲، ۳۶].

۱.۳.۲ الگوریتم جنگل تصادفی رتبه‌ای

همانطور که گفته شد، جنگل تصادفی رتبه‌ای، ساختار کلی جنگل تصادفی کلاسیک را حفظ نموده و تفاوت آن در فرآیند رشد درخت‌ها یا به عبارت دیگر افراز فضای متغیرهای توضیحی است که به شرح زیر خلاصه می‌شود. با فرض اینکه جنگل تصادفی شامل تعداد B درخت است، برای $b = 1, \dots, B$ مراحل زیر انجام می‌شود:

۱. یک نمونه‌ی تصادفی به روش بدون جایگذاری و با حجم n^* ($n^* < n$) از مشاهدات اصلی انتخاب می‌شود. مشاهدات انتخاب نشده را، OOB_b می‌نامیم.

۲. درخت T_b تحت نمونه‌ی خودگردان مذکور، با شرایط زیر رشد می‌کند. در هر مرحله از افراز یا در هر گره

- تعداد m متغیر از کل متغیرها (p) به تصادف انتخاب می‌شود.
- جستجو برای افراز بهینه و نقطه بهینه فقط در بین این m متغیر موجود و براساس اصول درخت استنباط شرطی صورت می‌گیرد.
- گره‌ی مذکور به دو قسمت تقسیم می‌شود.

افراز سلسله‌مراتبی فضا تا زمانی ادامه می‌یابد که متغیر توضیحی مستقل از متغیر پاسخ باشد. با انجام این مراحل، اجماعی از درختان $T_1, T_2, \dots, T_B$ ایجاد شده و رده‌بندی مشاهده‌ی جدید x همانند روش جنگل تصادفی کلاسیک اعمال می‌شود. لازم به ذکر است که حجم داده‌های آموزشی در هر درخت (n^*) نسبتی از حجم داده‌های اصلی است. استرابل و همکاران در سال ۲۰۰۷ نشان دادند که $63/2$ درصد نمونه‌ی خودگردان، جزء داده‌های اصلی هستند و لذا می‌توان $n^* = 0/632n$ فرض کرد [۲۰، ۳۷].

۲.۳.۲ معیارهای عملکرد

در این قسمت به معرفی معیارهای اندازه‌گیری خطا به منظور ارزیابی دقت پیش‌بینی جنگل تصادفی برای رده‌بندی متغیر پاسخ رتبه‌ای و استفاده از آن‌ها برای تعیین متغیرهای با اهمیت می‌پردازیم.

خطای رده‌بندی

با فرض اینکه متغیر پاسخ Y دارای رتبه‌های $1, \dots, K$ باشد، ماتریس زیر را می‌توان در خصوص نتایج رده‌بندی تشکیل داد که در آن درآیه (i, j) این ماتریس، فراوانی مشاهداتی

است که در اصل متعلق به رده i بوده‌اند ولی مدل جنگل تصادفی آن‌ها را در سطح j رده‌بندی کرده است.

$$\begin{matrix} & 1 & \dots & j & \dots & K \\ \begin{matrix} 1 \\ \vdots \\ i \\ \vdots \\ K \end{matrix} & \begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1K} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & a_{ij} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{K1} & \dots & a_{Kj} & \dots & a_{KK} \end{pmatrix} \end{matrix}$$

عناصر غیر از قطر اصلی، خطاهای رده‌بندی را نشان می‌دهند. خطای رده‌بندی (ER) عبارت است از:

$$ER = \frac{1}{n} \sum_{i=1}^K \sum_{j=1}^K a_{ij(i \neq j)} = \sum_{i=1}^n I(Y_i \neq \hat{Y}_i)$$

که در آن $I(\cdot)$ تابع نشانگر، Y_i رده‌ی واقعی و $\hat{Y}_i$ رده‌ی پیش‌بینی شده‌ی مشاهدات است.

مثال ۱.۳.۲. جدول ۱.۲ حاوی تعداد ۵ نمونه با دو متغیر توضیحی با رده‌های واقعی Y_i و پیش‌بینی شده $\hat{Y}_i$ است که در آن متغیر پاسخ می‌تواند از نوع رتبه‌ای یا اسمی فرض شود. برای بررسی عملکرد رده‌بندی آن با توجه معیار میزان خطا داریم:

$$ER = \frac{1}{n} \sum_{i=1}^n I(Y_i \neq \hat{Y}_i)$$

$$= \frac{1}{5}(2) = 0.4$$

همان‌طور که ملاحظه می‌شود این معیار برای متغیر پاسخ از نوع اسمی و رتبه‌ای، کارا است زیرا فقط به تعداد رده‌های متغیر پاسخ نیاز داریم و به همین دلیل، نیازی به تغییر ماهیت کیفی بودن مقادیر به ماهیت کمی، نیست.

جدول ۱.۲: مشاهدات با متغیر پاسخ اسمی

n	X_1	X_2	$\hat{Y}_i$	Y_i
۱	x_{11}	x_{12}	خوب	خوب
۲	x_{21}	x_{22}	ضعیف	خوب
۳	x_{31}	x_{32}	ضعیف	ضعیف
۴	x_{41}	x_{42}	خوب	ضعیف
۵	x_{51}	x_{52}	ضعیف	ضعیف

میانگین توان‌های دوم خطا

میانگین توان‌های دوم خطا (MSE) که بطور معمول به عنوان یک ابزار مناسب سنجش خطاهای رگرسیونی به کار می‌رود را می‌توان برای متغیر پاسخ رتبه‌ای نیز تعریف کرد [۱۴]. هنگامی که سطوح متغیر پاسخ رتبه‌ای $r = 1, \dots, K$ باشد، با در نظر گرفتن نمره s برای این رده‌ها به طوری که $s(1) < s(2) < \dots < s(K)$ ، میانگین توان‌های دوم خطاها به صورت (۱.۲) محاسبه می‌شود:

$$MSE = \frac{1}{n} \sum_{i=1}^n (s(Y_i) - s(\hat{Y}_i))^2 \quad (1.2)$$

که $s(Y_i)$ نمره‌ی رده‌ی واقعی مشاهدات و $s(\hat{Y}_i)$ نمره‌ی رده‌ی پیش‌بینی شده‌ی آن‌ها است. هر چه مقدار MSE بیشتر باشد، دقت رده‌بندی کاهش یافته و در نتیجه عملکرد رده‌بندی ضعیف است. واضح است که اگر $s(r) = r$ آنگاه

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

مثال ۲.۳.۲. همان‌طور که در جدول ۱.۲ در مثال قبل دیدیم، مشاهدات دارای متغیر پاسخ رتبه‌ای (ضعیف، خوب) هستند. بنابراین مقدار خطای آن‌ها براساس معیار MSE معمول، قابل محاسبه نیست. مطابق جدول ۲.۲ با تبدیل ماهیت کیفی رده‌های خوب و ضعیف به ماهیت کمی به ترتیب ۳ و ۵ یعنی $s(\text{ضعیف})=3$ و $s(\text{خوب})=5$ ، مقدار MSE عبارت است از:

$$\begin{aligned} MSE &= \frac{1}{n} \sum_{i=1}^n (s(Y_i) - s(\hat{Y}_i))^2 \\ &= \frac{1}{5} ((5-5)^2 + (5-3)^2 + (3-3)^2 \\ &\quad + (3-5)^2 + (3-3)^2) = 1/6 \end{aligned}$$

جدول ۲.۲: مشاهدات دارای متغیر پاسخ رتبه‌ای

n	X_1	X_2	$\hat{Y}_i$	Y_i	$s(\hat{Y}_i)$	$s(Y_i)$
۱	$x_{۱۱}$	$x_{۱۲}$	خوب	خوب	۵	۵
۲	$x_{۲۱}$	$x_{۲۲}$	ضعیف	خوب	۳	۵
۳	$x_{۳۱}$	$x_{۳۲}$	ضعیف	ضعیف	۳	۳
۴	$x_{۴۱}$	$x_{۴۲}$	خوب	ضعیف	۵	۳
۵	$x_{۵۱}$	$x_{۵۲}$	ضعیف	ضعیف	۳	۳

میانگین قدرمطلق خطا

میانگین قدرمطلق خطا^۴ (MAE) بسیار شبیه MSE است با این تفاوت که مشاهداتی که رده‌های پیش‌بینی شده‌ی آن‌ها، با رده‌ی واقعی خود زیاد اختلاف دارند، این تفاوت به اندازه معیار MSE قابل توجه نمی‌باشد [۱۴]. میانگین قدرمطلق خطاها به صورت (۲.۲) محاسبه می‌شود:

$$MAE = \frac{1}{n} \sum_{i=1}^n |s(Y_i) - s(\hat{Y}_i)|. \quad (2.2)$$

با کاهش مقدار MAE دقت رده‌بندی افزایش و در نتیجه عملکرد پیش‌بینی بهتر می‌شود.

امتیاز احتمال رتبه‌ای

امتیاز احتمال رتبه‌ای^۵ (RPS) احتمال دقت رده‌بندی را اندازه‌گیری می‌کند که برای ارزیابی احتمال پیش‌بینی متغیرهای رتبه‌ای، مورد استفاده قرار می‌گیرد و احتمال اینکه مشاهده i ام به رده‌ی r ام تعلق داشته باشد را محاسبه می‌کند. این معیار براساس رابطه زیر اختلاف بین تابع توزیع تجمعی پیش‌بینی شده و تابع توزیع تجمعی واقعی را به صورت (۳.۲) محاسبه می‌کند:

$$RPS = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^K (\hat{H}_i(r) - I(Y_i \leq r))^2 \quad (3.2)$$

^۴ Mean Absolute Error

^۵ Ranked probability score

که در آن K تعداد سطوح پاسخ، $\hat{I}_i(r)$ نشان‌دهنده‌ی احتمال پیش‌بینی مشاهده‌ی i ام در کلاس $r = 1, \dots, K$ و $I(\cdot)$ تابع نشانگر است. هر چه مقدار RPS کوچکتر باشد، نشان‌دهنده پیش‌بینی بهتر یا در واقع پیش‌بینی به کلاس واقعی نزدیکتر است [۲۲، ۱۰].

۳.۳.۲ اهمیت متغیر جنگل تصادفی رتبه‌ای

در جنگل تصادفی کلاسیک، اهمیت متغیرهای توضیحی به وسیله جایگشت متغیرهای توضیحی، برای داده‌های $B, b = 1, \dots, B$ تعیین می‌شود، به طوری که اگر متغیرهای توضیحی و پاسخ با هم ارتباط نداشته باشند، جایگشت متغیرها بر روی رده‌بندی تاثیر نمی‌گذارد. اما اگر متغیرهای توضیحی و پاسخ با هم ارتباط داشته باشند، جایگشت متغیرهای توضیحی باعث از بین رفتن این ارتباط و ایجاد اختلاف خطای رده‌بندی قبل و بعد از جایگشت می‌شود. بنابراین هرچه این اختلاف خطای رده‌بندی بیشتر باشد متغیر از اهمیت بالاتری برخوردار است.

در جنگل تصادفی کلاسیک اهمیت متغیرها به وسیله جایگشت متغیرها با استفاده از معیارهای خطای رده‌بندی و MSE ارزیابی می‌شود. اهمیت متغیرها در جنگل تصادفی رتبه‌ای نیز همانند جنگل تصادفی کلاسیک مورد ارزیابی قرار می‌گیرد با این تفاوت که از معیارهای متنوعی برای اندازه‌گیری خطا نظیر ER, MSE, MAE و RPS که در بخش ۲.۳.۲ به آن پرداختیم، می‌توان استفاده کرد [۲۳، ۲۲، ۴].

۴.۳.۲ ارزیابی کیفیت رتبه‌بندی متغیرها

همانطور که در ۳.۳.۲ گفته شد، اهمیت متغیرها براساس معیارهای عملکرد مورد ارزیابی قرار می‌گیرند.

برای ارزیابی کیفیت رتبه‌ی اهمیت متغیرها، از مساحت زیر منحنی AUC ^۶ استفاده می‌شود که کیفیت مدل استفاده شده را، مورد ارزیابی قرار می‌دهد به این صورت که متغیرهای توضیحی، به دو مجموعه مجزا $B = B_0 \cup B_1$ تقسیم می‌شوند که B_0 نشان‌دهنده‌ی متغیرهای بدون اثر^۷ و B_1 نشان‌دهنده‌ی متغیرهای موثر^۸ هستند. کمیّت AUC به صورت (۴.۲) محاسبه می‌شود:

$$AUC = \frac{1}{|B_0| |B_1|} \sum_{j \in B_0} \sum_{i \in B_1} I(VI_j < VI_i) + 0.5 I(VI_j = VI_i) \quad (4.2)$$

که $|B_0|$ و $|B_1|$ به ترتیب نشان‌دهنده‌ی تعداد متغیرهای بدون اثر و موثر، $I(\cdot)$ تابع نشانگر و $VI(\cdot)$ نشان‌دهنده میزان اهمیت متغیرها است. این کمیّت برای ارزیابی توانایی روش

^۶ Area Under the Curve

^۷ Noise

^۸ Signal

اهمیت متغیرها، در تمایز بین متغیرهای بدون اثر و موثر مورد استفاده قرار می‌گیرد. هنگامی که $AUC = 1$ بدین معنا است که همه‌ی متغیرهای موثر اهمیت بالاتری نسبت به متغیرهای بدون اثر دارند، همچنین $AUC = \frac{1}{4}$ نشان‌دهنده‌ی این است که در نیمی از حالات، اهمیت متغیرهای بدون اثر از اهمیت متغیرهای موثر، بالاتر است [۱۹].

فصل ۳

مطالعات شبیه‌سازی و داده‌های واقعی

در فصل قبل الگوریتم جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای شرح داده شد. در این فصل عملکرد جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای را توسط داده‌های شبیه‌سازی‌شده با هم مورد مقایسه قرار داده و سپس کارایی عملکرد جنگل‌های تصادفی کلاسیک و رتبه‌ای را برای رده‌بندی مجموعه داده‌های واقعی مورد ارزیابی قرار می‌دهیم.

۱.۳ داده‌های شبیه‌سازی شده

داده‌های شبیه‌سازی شده برای الگوریتم مورد استفاده، شامل $p = 65$ متغیر پیش‌گو با اندازه نمونه $n = 200$ هستند و داده‌های متغیر پاسخ با ترکیب دو مدل احتمالی از رابطه (۱.۳) شبیه‌سازی شده‌اند:

$$P(Y \leq r | X) = \xi P_1(Y \leq r | X) + (1 - \xi) P_2(Y \leq r | X) \quad (1.3)$$

که در آن $P_1(Y \leq r | X)$ و $P_2(Y \leq r | X)$ نشان‌دهنده احتمال تجمعی رخداد رتبه‌های مذکور است که به صورت (۲.۳) تعریف می‌شوند:

$$P_g(Y \leq r | X) = \frac{\exp(\gamma_{\circ rg} + X^T \gamma_g)}{1 + \exp(\gamma_{\circ rg} + X^T \gamma_g)} \quad (2.3)$$

که در آن $r = 1, \dots, K$ و $g \in \{1, 2\}$ در اینجا به منظور بررسی تحلیل حساسیت روش پیشنهادی نسبت به تعداد سطوح متغیر پاسخ رتبه‌ای، سه مقدار متفاوت $K = 3, 6, 9$ در نظر گرفته شده است. نسبت ترکیب این دو مدل احتمال با اندازه‌های مختلف $\xi = 0, \xi = 1$ در نظر گرفته شده است. همچنین $X^T = (X_1, X_2, \dots, X_{65})$ نشان‌دهنده بردار متغیرهای پیش‌گو، $\gamma_{01g} \leq \dots \leq \gamma_{0Kg}$ ضرایب ثابت که مقدار آن‌ها در جدول ۱.۳ نمایش داده شده و مقادیر آن‌ها به سطوح متغیر پاسخ r وابسته است. ضرایب γ_g برخلاف ضرایب ثابت به r وابسته نبوده و نسبت به سطوح مختلف رتبه‌ها، تغییر نمی‌کنند.

جدول ۱.۳: ضرایب ثابت γ_0

K	γ_{01}	γ_{02}	γ_{03}	γ_{04}	γ_{05}	γ_{06}	γ_{07}	γ_{08}	γ_{09}
۳	$-1/80$	۰	$1/80$	∞	—	—	—	—	—
۶	$-4/50$	$-1/50$	۰	$1/50$	$4/50$	∞	—	—	—
۹	$-5/90$	$-3/41$	$-1/55$	$-0/31$	$0/31$	$1/55$	$3/41$	$5/90$	∞

در این شبیه‌سازی بعضی از متغیرها در هر دو مدل و برخی دیگر فقط بر یک مدل اثر می‌گذارند و تعدادی از آن‌ها برای هر دو مدل بدون اثر هستند که در جدول ۲.۳ مشخص شده‌اند. ضرایب متغیرهای فاقد اثر برابر صفر و متغیرهای تاثیرگذار دارای ضرایب غیر صفر هستند.

جدول ۲.۳: ضرایب متغیرهای پیش‌گو موثر و بدون اثر

متغیر مدل	متغیرهای تاثیرگذار			متغیرهای فاقد اثر
۱	$(X_1, \dots, X_5)$ $\gamma_1 = \dots = \gamma_5 = 1$	$(X_6, \dots, X_{10})$ $\gamma_6 = \dots = \gamma_{10} = 0/75$	$(X_{11}, \dots, X_{15})$ $\gamma_{11} = \dots = \gamma_{15} = 0/5$	$(X_{16}, \dots, X_{65})$ $\gamma_{16} = \dots = \gamma_{65} = 0$
۲	$(X_1, X_2, X_6, X_7, X_{11}, X_{12})$ $\gamma_1 = \gamma_2 = \gamma_6 = \gamma_7 = \gamma_{11} = \gamma_{12} = 1$	$(X_3, X_4, X_8, X_9, X_{13}, X_{14})$ $\gamma_3 = \gamma_4 = \gamma_8 = \gamma_9 = \gamma_{13} = \gamma_{14} = -1$	$(X_5, X_{10}, X_{15}, X_{16}, \dots, X_{65})$ $\gamma_5 = \gamma_{10} = \gamma_{15} = \gamma_{16} = \dots = \gamma_{65} = 0$	

داده‌های مجموعه متغیر $X^T = (X_1, X_2, \dots, X_{65})$ از توزیع نرمال چندمتغیره تولید شده‌اند که به منظور تنوع بخشیدن به بررسی عملکرد روش مورد مطالعه، این متغیرها یک بار بدون همبستگی درونی و از توزیع نرمال $N(0_p, I_p)$ ، که در آن I_p نشان‌دهنده ماتریس واحد، و بار دیگر با در نظر گرفتن همبستگی و از توزیع نرمال $N(0_p, \Sigma_p)$ در نظر گرفته

شده‌اند که

$$\Sigma_p = \begin{bmatrix} A_{\text{signal}} & \circ & \dots & \circ \\ \circ & A_{\text{noise}_1} & \dots & \circ \\ \vdots & \vdots & \ddots & \vdots \\ \circ & \circ & \dots & A_{\text{noise}_\delta} \end{bmatrix}$$

$A_{\text{signal}_j(15 \times 15)} = (a_{ij})$ نشان‌دهنده همبستگی بین متغیرهای پیش‌گوی موثر $X_1, \dots, X_{15}$ است که

$$a_{ij} = \begin{cases} 1 & i = j \\ 0/8 & i \neq j; i, j \in \{1, 3, 6, 8, 11, 13\} \\ 0 & O.W. \end{cases}$$

یعنی متغیرهای توضیحی موثر $X_1, X_3, X_6, X_8, X_{11}, X_{13}$ دارای هم‌خطی شدید هستند. همچنین همبستگی 50 متغیر توضیحی غیرموثر $X_{16}, \dots, X_{65}$ به صورت 5 بلوک 10 تایی و با نمادهای $A_{\text{noise}_1}, \dots, A_{\text{noise}_5}$ در نظر گرفته شده‌اند

$$A_{\text{noise}_j} = \begin{bmatrix} 1 & \rho_j & \dots & \rho_j \\ \rho_j & 1 & \dots & \rho_j \\ \vdots & \vdots & \ddots & \vdots \\ \rho_j & \rho_j & \dots & 1 \end{bmatrix}_{10 \times 10}$$

که $j = 1, \dots, 5$ و $\rho_1 = 0/8, \rho_2 = 0/6, \rho_3 = 0/4, \rho_4 = 0/2, \rho_5 = 0$

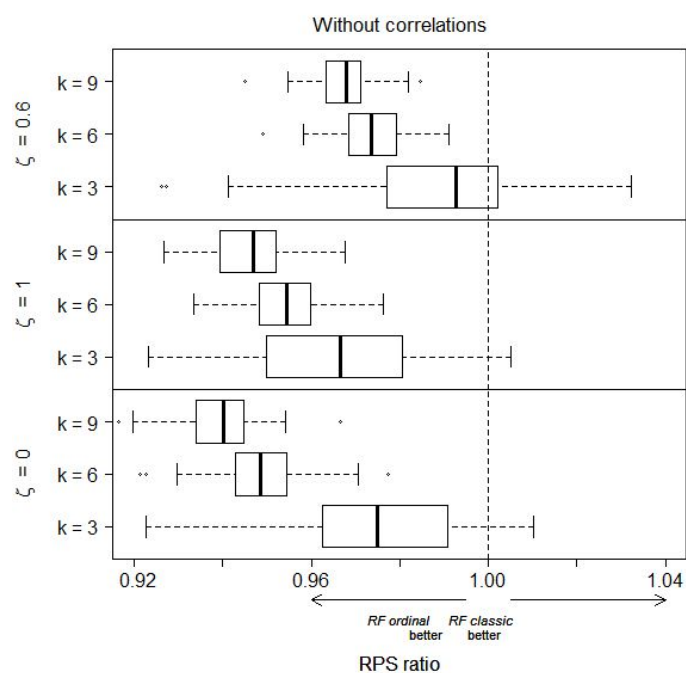
۱.۱.۳ بررسی عملکرد جنگل‌های تصادفی

به منظور بررسی الگوریتم جنگل تصادفی رتبه‌ای، عملکرد آن را با روش کلاسیک، که در آن متغیر پاسخ رتبه‌ای از نوع اسمی فرض می‌شود، مقایسه می‌کنیم. در شبیه‌سازی انجام شده، تعداد درخت‌ها یا مدل‌های مبنا برای هر دو روش جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای 1000 در نظر گرفته شده است. پس از اجرای دو مدل مختلف جنگل‌های تصادفی کلاسیک (متغیر پاسخ اسمی) و جنگل تصادفی رتبه‌ای برای داده‌های شبیه‌سازی شده، خطای مدل براساس معیار RPS در هر دو حالت محاسبه شده و نسبت خطای مدل رتبه‌ای به مدل کلاسیک در نظر گرفته شده است. هر چه این نسبت کوچکتر از یک باشد به معنای برتری مدل جنگل تصادفی رتبه‌ای و در نتیجه افزایش دقت رده‌بندی و پیش‌بینی آن نسبت به جنگل تصادفی کلاسیک است و هر چه این نسبت بیشتر از یک باشد حاکی از برتری جنگل تصادفی کلاسیک است. برای پایش اثرات تصادفی داده‌های شبیه‌سازی شده، عمل تولید داده‌ها به تعداد 100 بار تکرار شده است و برای هر دسته داده تولیدشده در حالت‌های مختلف همبستگی و عدم همبستگی متغیرهای پیش‌گو، و همچنین به ازای

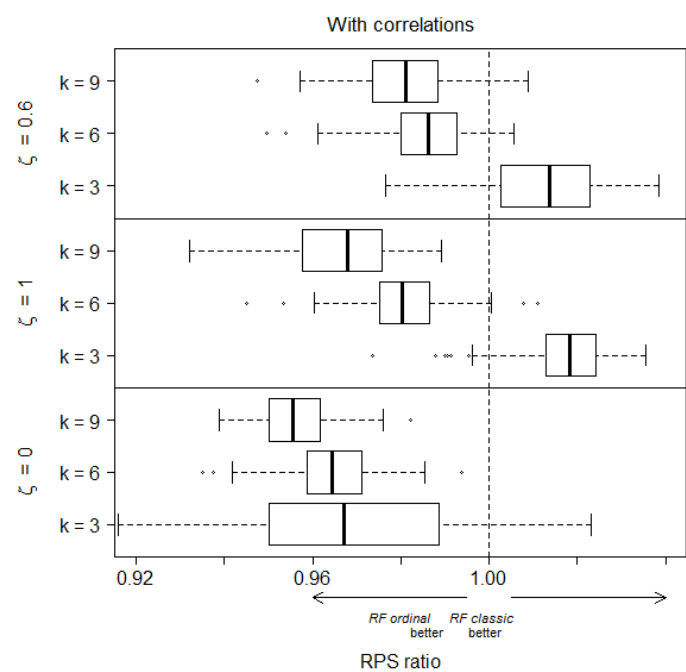
نسبت‌های مختلف ترکیب دو مدل احتمال ($\xi = 0, \xi = 0.06, \xi = 1$)، خطای مدل‌سازی محاسبه شده است. همچنین به منظور بررسی تاثیر تعداد رتبه‌ها (سطوح متغیر پاسخ) در عملکرد روش معرفی‌شده، سه حالت مختلف برای تعداد رتبه‌ها در نظر گرفته شده است یعنی $K = 3, 6, 9$ و نتایج آن به صورت نمودارهای جعبه‌ای نشان داده شده است. شکل‌های ۱.۳ و ۲.۳ نمودار جعبه‌ای خطاها را برای متغیرهای پیش‌گوی ناهمبسته و همبسته نشان می‌دهد.

همانگونه که در نمودارها ملاحظه می‌شود، حساسیت عملکرد روش مورد مطالعه نسبت به نوع مدل، مقدار ξ ، تعداد سطوح پاسخ و همبستگی متغیرهای پیش‌گو مشخص شده است. در حالتی که متغیرهای توضیحی ناهمبسته باشند، نسبت به حالتی که متغیرها، دارای همبستگی درونی هستند خطای جنگل تصادفی رتبه‌ای نسبت به جنگل تصادفی کلاسیک کاهش یافته و در نتیجه دقت پیش‌بینی افزایش می‌یابد که بهبود عملکرد جنگل تصادفی رتبه‌ای را موجب می‌شود. در حالتی که داده‌ها از هر دو مدل تولید می‌شوند ($\xi = 0.06$) خطای جنگل تصادفی رتبه‌ای نسبت به کلاسیک کمتر است. برای $\xi = 0$ ، یعنی وقتی داده‌ها فقط از مدل دوم ($g = 2$) تولید می‌شوند، خطای جنگل تصادفی رتبه‌ای از جنگل تصادفی کلاسیک کمتر بوده و در نتیجه دقت رده‌بندی افزایش یافته و عملکرد آن نسبت به حالت‌های $\xi = 0.06$ و $\xi = 1$ بهتر است.

می‌توان روند بهبودی عملکرد جنگل تصادفی رتبه‌ای را نسبت به جنگل تصادفی کلاسیک، هنگامی که تعداد سطوح متغیر پاسخ رو به افزایش است، مشاهده نمود. در شکل ۱.۳ و به ازای تمام مدل‌ها ($\xi = 0, \xi = 0.06, \xi = 1$)، به طور کلی می‌توان گفت که در حالت عدم همبستگی متغیرهای توضیحی، همواره جنگل تصادفی رتبه‌ای عملکرد بهتری نسبت به جنگل تصادفی کلاسیک داشته است و این برتری با افزایش مقادیر سطوح (K) شدت بیشتری به خود می‌گیرد. در شکل ۲.۳ به جز حالت $K = 3$ برای ($\xi = 0.06, \xi = 1$)، نیز نتیجه مشابهی را می‌توان گرفت.



شکل ۱.۳: نمودار جعبه‌ای نسبت خطای جنگل تصادفی رتبه‌ای به جنگل تصادفی کلاسیک وقتی متغیرهای توضیحی فاقد همبستگی هستند



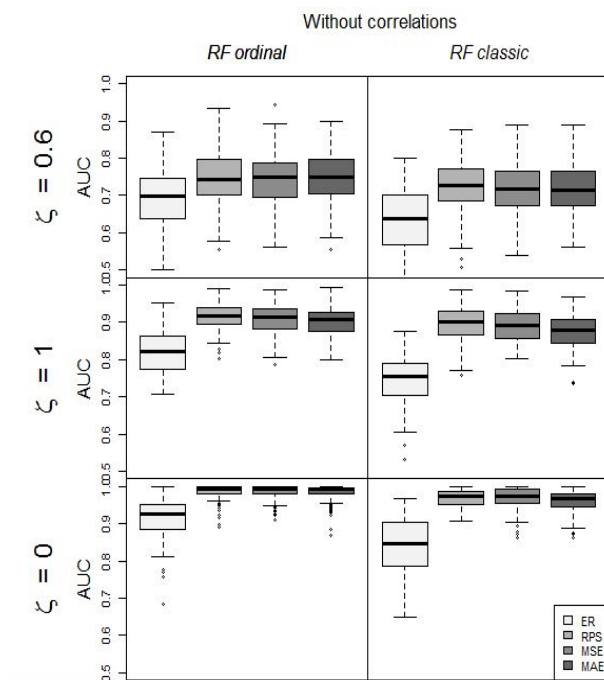
شکل ۲.۳: نمودار جعبه‌ای نسبت خطای جنگل تصادفی رتبه‌ای به جنگل تصادفی کلاسیک با فرض همبستگی درونی متغیرها

۲.۱.۳ بررسی اهمیت متغیرها

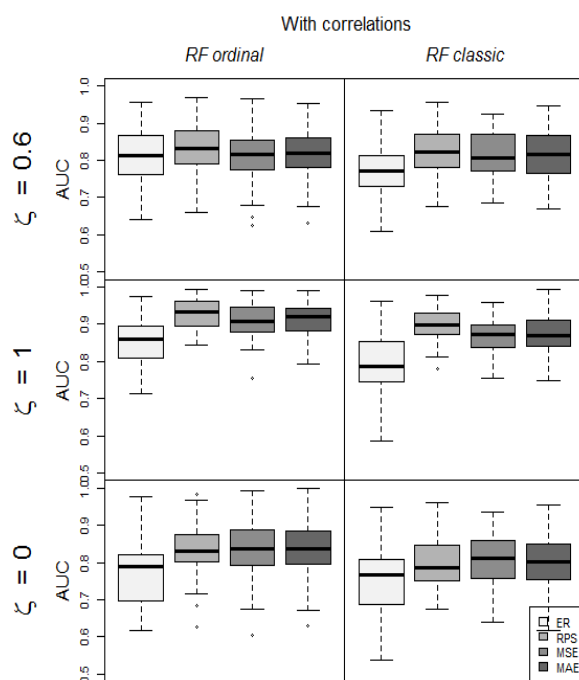
همانطور که در بخش ۳.۳.۲ گفته شد از کمیت AUC می‌توان برای کیفیت الگوریتم تعیین اهمیت متغیرها در خصوص تشخیص متغیر با اهمیت از سایر متغیرها استفاده کرد. باید توجه داشت که AUC مقادیر بین صفر و یک را می‌گیرد و هر چه این مقدار نزدیک به یک باشد، نشان‌دهنده‌ی بهبود روش مورد بررسی است یا به عبارت دیگر بدین معناست که در اکثر حالات مقدار اهمیت متغیرهای موثر از مقدار اهمیت متغیرهای غیرموثر بیشتر است. اگر $AUC = 1$ آنگاه مقدار VIM برای همه متغیرهای موثر بیشتر از مقدار VIM برای متغیرهای بدون اثر است. اما هر چه مقدار AUC به صفر نزدیک باشد نشان‌دهنده‌ی آن است که در اکثر حالات مقدار VIM برای متغیرهای موثر کمتر از مقدار آن برای متغیرهای بدون اثر است، بنابراین روش مورد مطالعه مناسب نمی‌باشد.

به منظور بررسی کیفیت الگوریتم رتبه‌بندی اهمیت متغیرها، مقدار AUC را در جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای، مورد مقایسه قرار می‌دهیم. در شبیه‌سازی انجام شده، متغیرهای پیش‌گو $B = X_1, \dots, X_{65}$ ، به دو مجموعه مجزا $B_1 = X_1, \dots, X_{15}$ (متغیرهای موثر) و $B_0 = X_{16}, \dots, X_{65}$ (متغیرهای بدون اثر) تقسیم می‌شوند. برای هر یک از این دسته متغیرها، مقدار VIM براساس معیارهای عملکرد MAE ، MSE ، ER و RPS برای هر دو روش جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای، محاسبه می‌شود و با اعمال این مقادیر در رابطه ۴.۲، مقدار AUC برای هریک از معیارهای عملکرد حاصل می‌شود.

شکل‌های ۳.۳ و ۴.۳ نمودار جعبه‌ای AUC را برای حالتی که تعداد سطوح پاسخ $K = 9$ است به ترتیب برای متغیرهای پیش‌گوی ناهمبسته و همبسته نشان می‌دهد. همانگونه که در نمودارها مشخص است، برای هر دو روش جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای، VIM های مبتنی بر MAE ، MSE و RPS عملکرد بهتری نسبت به VIM مبتنی بر ER دارند. در شکل ۳.۳ و در حالتی که $\xi = 1$ و $\xi = 0$ است یعنی داده‌ها فقط از مدل اول یا مدل دوم تولید شده‌اند، مقدار AUC نسبت به حالت $\xi = 0/6$ افزایش یافته است. همچنین در شکل ۴.۳ در حالتی که متغیرهای توضیحی ناهمبسته باشند، نسبت به حالتی که متغیرها دارای همبستگی درونی هستند، مقادیر AUC افزایش یافته و موجب بهبودی روند عملکرد می‌شود. بنابراین در اکثر حالات، مقادیر AUC برای جنگل تصادفی رتبه‌ای، بیشتر از جنگل تصادفی کلاسیک است، که این روند، با در نظر گرفتن رتبه‌ی متغیرها، هنگام محاسبه‌ی AUC ، اهمیت متغیرهای موثر را بهتر نشان می‌دهد.



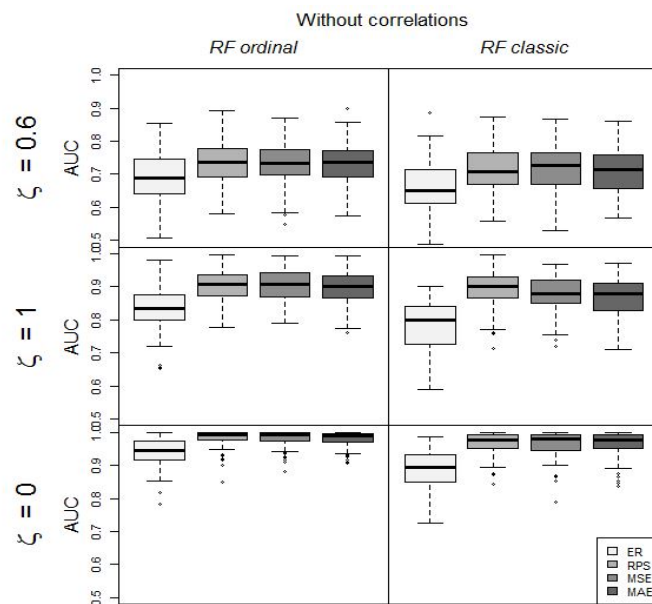
شکل ۳.۳: نمودار جعبه‌ای AUC برای $K = 9$ با استفاده از معیارهای عملکرد برای مقایسه جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک وقتی متغیرهای توضیحی فاقد همبستگی هستند



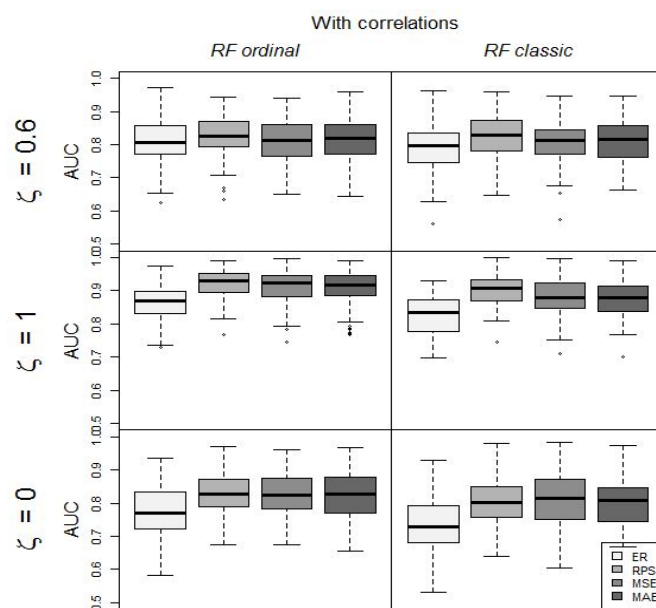
شکل ۴.۳: نمودار جعبه‌ای AUC برای $K = 9$ با استفاده از معیارهای عملکرد برای مقایسه جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک با فرض همبستگی درونی متغیرها

شکل‌های ۵.۳ و ۶.۳ نمودار جعبه‌ای AUC را برای حالتی که تعداد سطوح پاسخ $K = 6$ است به ترتیب برای متغیرهای پیش‌گوی ناهمبسته و همبسته نشان می‌دهد. همانگونه که در نمودارها مشخص است، برای هر دو روش جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای، نتایج تقریباً شبیه حالت قبل $K = 9$ است که در آن VIM ‌های مبتنی بر MSE ، MAE و RPS عملکرد بهتری نسبت به VIM مبنی بر ER دارند. در حالتی که $\xi = 1$ و $\xi = 0$ است یعنی داده‌ها فقط از مدل اول یا مدل دوم تولید شده‌اند، مقدار AUC افزایش یافته است. همچنین در حالتی که متغیرهای توضیحی ناهمبسته باشند، نسبت به حالتی که متغیرها دارای همبستگی درونی هستند، مقادیر AUC افزایش یافته و موجب بهبودی روند عملکرد می‌شود، با این تفاوت که مقادیر AUC برای حالتی که تعداد سطوح پاسخ $K = 6$ کمتر از مقادیر آن، برای حالتی که تعداد سطوح پاسخ $K = 9$ است. بنابراین می‌توان گفت تعداد سطوح، در رتبه اهمیت متغیرها مهم است و با افزایش سطوح پاسخ، دقت الگوریتم افزایش می‌یابد.

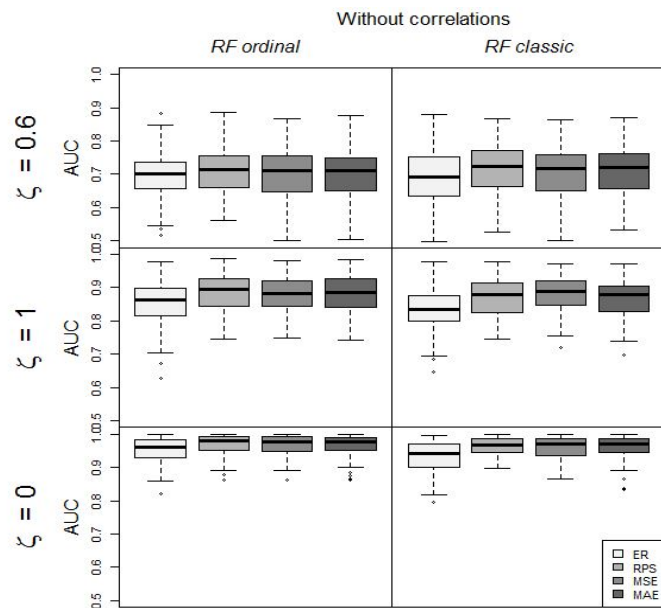
شکل‌های ۷.۳ و ۸.۳ نتایج AUC را برای $K = 3$ نشان می‌دهد. همانطور که ملاحظه می‌شود همانند دو حالت قبلی، VIM ‌های مبتنی بر MSE ، MAE و RPS عملکرد بهتری نسبت به VIM مبنی بر ER دارند. در حالتی که $\xi = 1$ و $\xi = 0$ است یعنی داده‌ها فقط از مدل اول یا مدل دوم (به ویژه در حالت ناهمبسته بودن متغیرهای توضیحی) تولید شده‌اند، مقدار AUC افزایش یافته است. همچنین در حالتی که متغیرهای توضیحی ناهمبسته باشند، نسبت به حالتی که متغیرها دارای همبستگی درونی هستند، مقادیر AUC افزایش یافته و موجب بهبودی روند عملکرد می‌شود. در حالت کلی هنگامی که $K = 3$ است نسبت به حالت‌های قبلی که $K = 9$ و $K = 6$ ، مقادیر AUC کاهش یافته است.



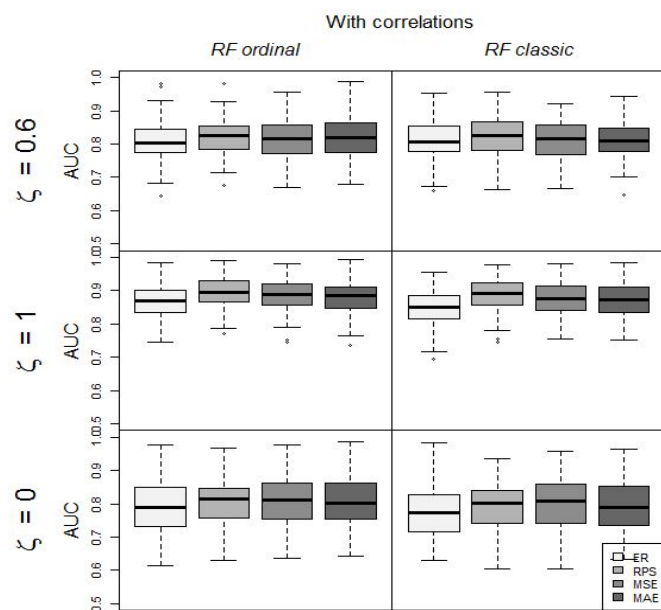
شکل ۵.۳: نمودار جعبه‌ای AUC برای $K = 6$ با استفاده از معیارهای عملکرد برای مقایسه جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک وقتی متغیرهای توضیحی فاقد همبستگی باشند



شکل ۶.۳: نمودار جعبه‌ای AUC برای $K = 6$ با استفاده از معیارهای عملکرد برای مقایسه جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک با فرض همبستگی درونی متغیرها



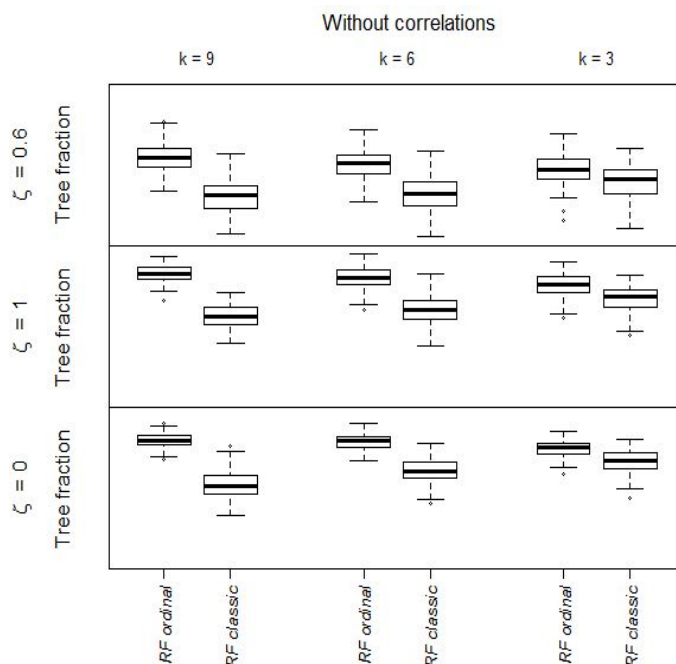
شکل ۷.۳: نمودار جعبه‌ای AUC برای $K = 3$ با استفاده از معیارهای عملکرد برای مقایسه جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک وقتی متغیرهای توضیحی فاقد همبستگی هستند



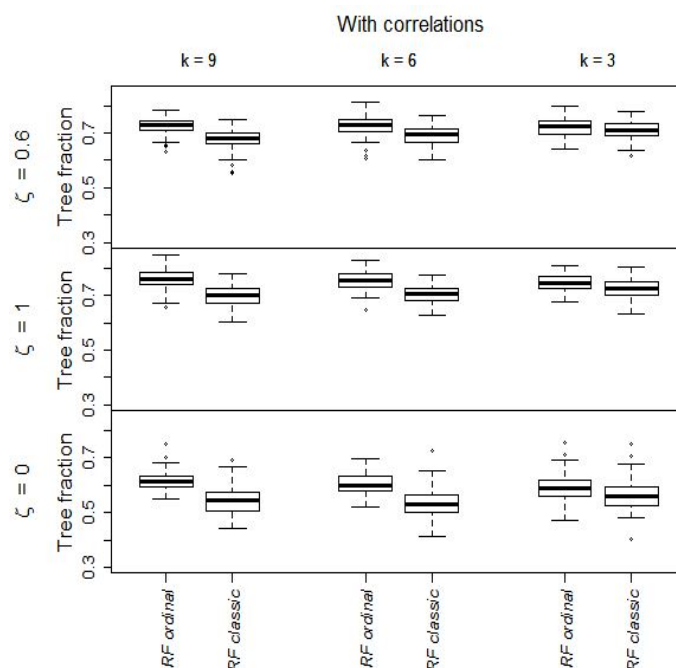
شکل ۸.۳: نمودار جعبه‌ای AUC برای $K = 3$ با استفاده از معیارهای عملکرد برای مقایسه جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک با فرض همبستگی درونی متغیرها

در درخت استنباط شرطی، که در آن مرتبه‌ی متغیرهای پاسخ لحاظ می‌گردد، متغیرهایی برای افراز انتخاب می‌شوند که رابطه‌ی قوی‌تری با متغیر پاسخ داشته باشند. مزیت درخت استنباط شرطی آن است که قدرت بالایی در تشخیص ارتباط بین متغیر پیش‌گو و متغیر پاسخ رتبه‌ای دارد. همچنین متغیرهایی که به عنوان اولین افراز انتخاب می‌شوند و نزدیک ریشه هستند از اهمیت بالاتری برخوردارند. بنابراین انتظار می‌رود متغیرهای منتخب برای اولین افراز، جزو متغیرهای تاثیرگذار مدل باشند. در جنگل تصادفی رتبه‌ای که در آن رتبه‌ی متغیر پاسخ لحاظ می‌گردد، اگر متغیرهای موثر در اولین افراز آن‌ها انتخاب شود، می‌توان نتیجه گرفت که درخت استنباط شرطی نسبت به درخت کلاسیک دقیق‌تر هستند.

به منظور پایش تاثیر رتبه‌ها در مراحل افراز، برای مجموعه مشاهدات به حجم $n = 200$ با متغیرهای شبیه‌سازی شده $X_1, \dots, X_{65}$ که $X_1, \dots, X_{15}$ متغیرهای موثر و $X_{16}, \dots, X_{65}$ متغیرهای بدون اثر هستند، ساخت درخت ۵۰۰ بار تکرار شده و تعداد درختانی که در اولین افراز آن‌ها، متغیرهای موثر انتخاب می‌شوند مورد ارزیابی قرار می‌گیرند. شکل‌های ۹.۳ و ۱۰.۳ نسبت تعداد درختان جنگل تصادفی که در اولین افراز آن‌ها، متغیرهای تاثیرگذار انتخاب شده‌اند را به ترتیب برای متغیرهای ناهمبسته و همبسته نمایش می‌دهد. همانطور که ملاحظه می‌شود در حالت ناهمبسته بودن متغیرها، نسبت دقت جنگل تصادفی رتبه‌ای به جنگل تصادفی کلاسیک برای انتخاب متغیرهای موثر در اولین افراز بیشتر است. همچنین هنگامی که متغیرهای توضیحی همبسته هستند دقت جنگل تصادفی رتبه‌ای بیشتر است. هنگامی که تعداد سطوح پاسخ رو به افزایش است، نسبت تعداد درختان نیز زیاد می‌شود، که این نسبت در حالت رتبه‌ای بیشتر از کلاسیک است.



شکل ۹.۳: نسبت تعداد درختان برای انتخاب متغیر با اهمیت در اولین افراز، وقتی متغیرهای توضیحی فاقد همبستگی هستند



شکل ۱۰.۳: نسبت تعداد درختان برای انتخاب متغیر با اهمیت در اولین افراز، با فرض همبستگی درونی متغیرها

شبیه‌سازی داده‌ها و مدل‌سازی جنگل تصادفی توسط بسته‌های `mvtnorm`، `snowfall`،

ROCR و party در نرم افزار R انجام گرفته است [۱۵].

۲.۳ معرفی مجموعه داده‌های واقعی

در این بخش مجموعه داده‌ی نوزادان دارای وزن کم قبل از تولد که از طریق درگاه <http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets> در دسترس عموم قرار دارد، استفاده شده است. داده‌ها مربوط به نوزادانی است که قبل از تولد وزن آن‌ها کمتر از ۱۶۰۰ گرم است و در سال ۱۹۸۱ تا ۱۹۸۷ متولد شده‌اند. این داده‌ها توسط شی در سال ۱۹۹۸ مورد تحلیل قرار گرفته است. [۲۸]. هدف از تحلیل، پیش‌بینی نمره‌ی آپگار^۱ است که وضعیت جسمانی نوزاد را اندازه‌گیری می‌کند و امتیاز ۱ تا ۹ را به آن‌ها نسبت می‌دهد [۳۰].

در این بررسی عواملی چون تلاش تنفسی، ضربان قلب، توان عضلانی، بازتاب‌های عصبی و رنگ پوست، مورد ارزیابی قرار می‌گیرند. به هر کدام از معیارهای ذکر شده، سه نمره داده می‌شود که نمره ۲ نشان‌دهنده‌ی مناسب بودن معیار موردنظر است. سپس این نمرات با هم جمع می‌شوند تا نمره‌ی کلی آپگار مشخص شود. نمره‌های این آزمون از ۱ تا ۱۰ متفاوت است و نمره‌ی ۱۰ سالم‌ترین حالت محسوب می‌شود، با این حال امتیاز ۸ یا ۹ نیز رضایت بخش است. اما به طور معمول امتیاز بین ۸ تا ۱۰ خوب یا عالی است و غالباً نیازی به اقدامات حمایتی ندارد. امتیاز بین ۵ تا ۷ در موقعیت نسبتاً عادی قرار دارد و ممکن است نیاز به حمایت تنفسی داشته باشد. نوزادانی که امتیاز زیر ۵ دارند، در موقعیت بدی قرار می‌گیرند و نیاز به کمک سریع خواهند داشت.

اکثر نوزادان، نمره آپگار ۷ یا بالاتر دارند و تنها تعداد کمی ۱۰ نمره کامل را به خود اختصاص می‌دهند. آزمون آپگار سنجش میزان سلامتی نوزاد است که در دقایق ۱، ۵، ۱۵ و ۲۰ پس از تولد از نوزاد گرفته می‌شود. اگر آزمون نوزاد در دقیقه ۱ مشکلی نداشته باشد، در دقایق بعدی انجام نمی‌شود، در غیر اینصورت در دقایق بعدی نیز بعمل می‌آید [۱، ۳۰]. عوامل مختلفی موجب تغییرات حقیقی و کاذب در نمره آپگار می‌گردند. از علل کاهش نمره آپگار، می‌توان به نارسایی، مصرف بعضی داروها توسط مادر، تجویز سولفات منیزیم، میوپاتی و نوروپاتی مادرزادی و ترومای مغزی و نخاعی، نژاد، دیابت و اسیدوز مادر باردار، جنسیت نوزاد و سایر عوامل اشاره نمود.

در این داده‌ها مواردی از قبیل جنسیت، وزن، نژاد، نوع حمل، نوع تولد، مدت بارداری، بارداری چندم، مصرف دارو توسط مادر چون بتامتازون، منیزیم سولفات و بتادرینرژیک اندازه‌گیری شده و به عنوان متغیرهای توضیحی مورد بررسی قرار گرفته شده‌اند.

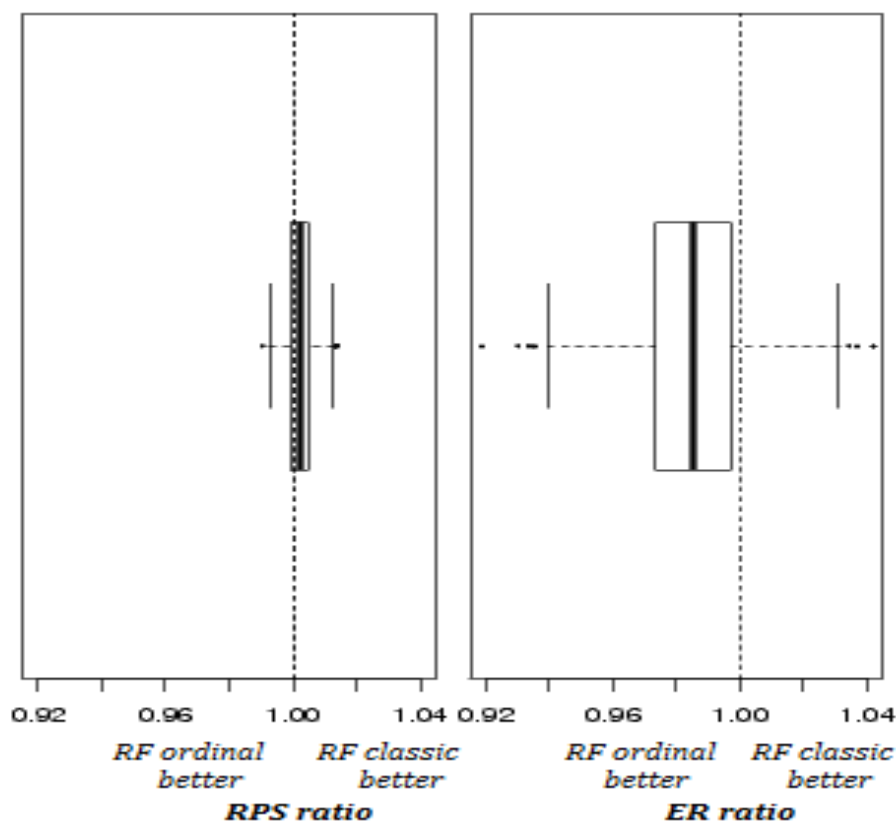
^۱Apgar

۱.۲.۳ اجرای الگوریتم بر روی مجموعه داده‌ها

به منظور بررسی الگوریتم جنگل تصادفی رتبه‌ای، عملکرد آن را با روش کلاسیک، که در آن متغیر پاسخ رتبه‌ای از نوع اسمی فرض می‌شود، مقایسه می‌کنیم. برای این منظور داده‌های واقعی مربوط به نوزادانی که قبل از تولد وزن کمتری دارند مورد بررسی قرار گرفته و نمره‌ی آپگار آن‌ها پیش‌بینی می‌شود با این هدف که آیا این نوزادان، هنگام تولد از سلامت جسمانی برخوردارند یا خیر. در این الگوریتم تعداد مشاهدات مورد بررسی $n = 218$ ، تعداد متغیرهای پیش‌گو $p = 10$ و تعداد سطوح پاسخ $K = 9$ است که همان تعداد رده‌های نمره‌ی آپگار می‌باشد. همچنین به منظور افزایش دقت پیش‌بینی تعداد درخت‌ها یا مدل‌های مبنا برای هر دو روش جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای ۱۰۰۰ در نظر گرفته شده است.

برای بررسی اثرات تصادفی داده‌ها از روش اعتبارسنجی 10-Fold استفاده شده (به پیوست «آ» مراجعه کنید) و به منظور ارائه نتایج پایدارتر، این عمل ۵۰۰ بار تکرار می‌شود. پس از اجرای دو مدل مختلف جنگل‌های تصادفی کلاسیک و جنگل تصادفی رتبه‌ای برای مجموعه داده‌ها، خطای مدل براساس معیار RPS و ER در هر دو حالت به‌دست آمده و نسبت خطای مدل رتبه‌ای به مدل کلاسیک محاسبه می‌شود به‌طوری که هر چه این نسبت کوچکتر از یک باشد به معنای برتری مدل جنگل تصادفی رتبه‌ای و در نتیجه افزایش دقت رده‌بندی و پیش‌بینی آن نسبت به جنگل تصادفی کلاسیک است.

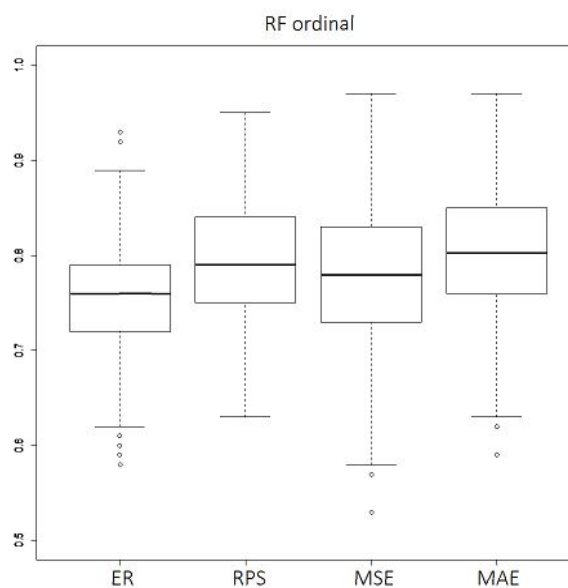
نمودار ۱۱.۳ میزان خطا را براساس معیارهای RPS و ER برای جنگل تصادفی رتبه‌ای نسبت به جنگل تصادفی کلاسیک نمایش می‌دهد. همانطور که ملاحظه می‌شود هنگامی که از ER برای محاسبه خطا استفاده شده، نسبت خطای جنگل تصادفی رتبه‌ای به کلاسیک کمتر از یک است و بدین معنا است که دقت جنگل تصادفی رتبه‌ای نسبت به کلاسیک بهتر است. اما هنگامی که از RPS برای محاسبه خطا استفاده می‌شود، نسبت خطای جنگل تصادفی رتبه‌ای به کلاسیک افزایش یافته یا حداقل در یک اندازه هستند، که این به نوع ماهیت متغیرهای پاسخ برمی‌گردد. بنابراین می‌توان گفت دقت پیش‌بینی جنگل تصادفی رتبه‌ای بهتر یا حداقل در سطح جنگل تصادفی کلاسیک است.



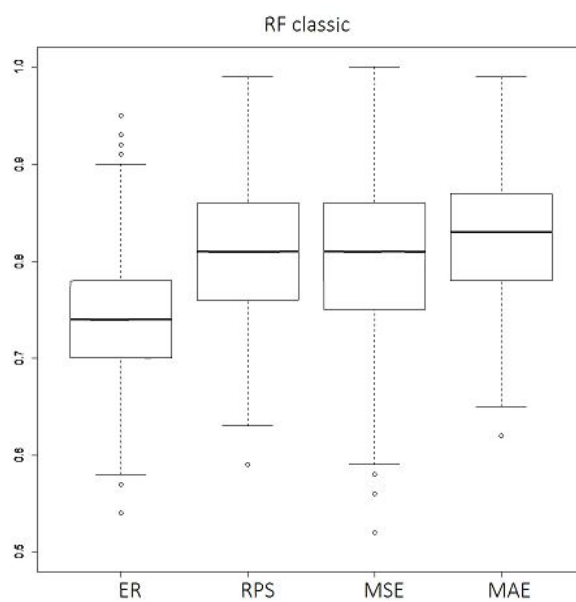
شکل ۱۱.۳: نمودار جعبه‌ای نسبت خطای جنگل تصادفی رتبه‌ای به جنگل تصادفی کلاسیک برحسب معیار ER (نمودار سمت چپ) و معیار RPS (نمودار سمت راست)

۲.۲.۳ بررسی اهمیت متغیرها

همانطور که در بخش ۳.۳.۲ گفته شد از کمیت AUC می‌توان برای کیفیت الگوریتم تعیین اهمیت متغیرها در خصوص تشخیص متغیر با اهمیت از سایر متغیرها استفاده کرد. هر چه این مقدار نزدیک به یک باشد، نشان‌دهنده‌ی بهبود روش مورد بررسی است. شکل‌های ۱۲.۳ و ۱۳.۳ معیارهای عملکرد را برای اهمیت متغیرها نشان می‌دهد. کیفیت VIM براساس معیارهای عملکرد ER ، MSE ، MAE و RPS برای هر دو روش جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای، مورد ارزیابی قرار گرفته است. همانطور که ملاحظه می‌شود در هر دو روش جنگل تصادفی رتبه‌ای و جنگل تصادفی کلاسیک معیارهای RPS ، MSE و MAE عملکرد بهتری نسبت به معیار ER دارند. همچنین می‌توان نتیجه گرفت که اندازه‌ی کمیت AUC در هر دو روش جنگل تصادفی کلاسیک و جنگل تصادفی رتبه‌ای در یک سطح هستند.



شکل ۱۲.۳: مقایسه معیارهای خطا برای ارزیابی اهمیت متغیرها در جنگل تصادفی رتبه‌ای



شکل ۱۳.۳: مقایسه معیارهای خطا برای ارزیابی اهمیت متغیرها در جنگل تصادفی کلاسیک

نتیجه‌گیری و پیشنهادات

نتیجه‌گیری

در مدل‌سازی جنگل‌های تصادفی کلاسیک برای داده‌های پاسخ رتبه‌ای، معمولاً این داده‌ها به صورت اسمی در نظر گرفته می‌شوند و ماهیت رتبه‌ای بودن داده‌ها تحت‌الشعاع قرار می‌گیرد. در این پایان‌نامه صورت دیگری از درخت رگرسیونی معرفی شد که در آن ماهیت ذاتی رتبه‌ها در نظر گرفته شده و نحوه‌ی افراز فضای متغیرها که منجر به رشد درخت می‌شود، دارای ساختار کاملاً متمایزی نسبت به جنگل تصادفی کلاسیک است. داده‌های شبیه‌سازی شده و واقعی نشان دادند که در اغلب حالت‌های مختلف، جنگل تصادفی رتبه‌ای دارای دقت پیش‌گویی بالاتری نسبت به جنگل تصادفی کلاسیک رده‌بندی است و هرچه تعداد رده‌های پاسخ بیشتر باشد، عملکرد آن بهتر است. همچنین می‌توان نتیجه گرفت که استقلال متغیرهای توضیحی باعث بهبود عملکرد جنگل تصادفی رتبه‌ای شده که منجر به افزایش دقت رده‌بندی و پیش‌گویی می‌شود. در نهایت می‌توان نتیجه گرفت که میزان دقت پیش‌بینی و محاسبه اهمیت متغیرها به ماهیت متغیر پاسخ وابسته است زیرا نتایج در داده‌های مختلف متفاوت هستند.

همچنین در بررسی کیفیت اهمیت متغیرها که براساس معیارهای عملکرد انجام می‌شود، می‌توان نتیجه گرفت که معیارهای RPS ، MSE و MAE به دلیل اینکه در محاسبات، ماهیت متغیر پاسخ را لحاظ می‌کنند، نسبت به معیار ER عملکرد بهتری دارند.

پیشنهادهات

به عنوان پیشنهاد برای آینده تحقیق می‌توان الگوریتم تحقیق را در حالت چندمتغیره پیاده‌سازی کرد و کارایی الگوریتم را مورد بررسی قرار داد.

پیوست آ

۱.آ مجموعه‌های آموزشی و آزمون

معمولا برای ارزیابی یک مدل، مشاهدات موجود به صورت تصادفی به دو دسته نمونه آموزشی^۱ و نمونه آزمون^۲ تقسیم‌بندی می‌شوند. نمونه‌ی آموزشی شامل زیرمجموعه‌ای تصادفی از مشاهدات که در برآزش مدل شرکت می‌کنند و مورد تحلیل قرار می‌گیرند. مشاهداتی که به عنوان نمونه‌ی آزمون انتخاب می‌شوند، دقت و عملکرد مدل را مورد ارزیابی قرار می‌دهند [۱۳].

۲.آ اعتبارسنجی متقابل

اعتبارسنجی متقابل^۳ روشی برای ارزیابی نتایج یک تجزیه و تحلیل آماری به یک مجموعه داده مستقل است. این روش به صورت تصادفی، نمونه اصلی را به یک مجموعه آموزشی، برای آموزش مدل و یک مجموعه آزمون، برای ارزیابی آن، تقسیم می‌کند و شامل روش‌های

^۱ Training or Learning sample

^۲ Test sample

^۳ Cross Validation

زیر است:

• اعتبارسنجی Leave-one-out

• اعتبارسنجی Leave-k-out

• اعتبارسنجی K-Fold

در این پایان نامه مفهوم اعتبارسنجی K-Fold را مورد بحث قرار می دهیم.

۱.۲.۲ اعتبارسنجی K-Fold

در روش اعتبارسنجی K-Fold، نمونه اصلی به طور تصادفی به K زیرنمونه مساوی تقسیم می شود. از زیرنمونه ها، یک زیرنمونه ی واحد به عنوان داده های آزمون، برای آزمایش مدل حفظ می شود و زیرنمونه های باقی مانده $(K - 1)$ به عنوان داده های آموزشی مورد استفاده قرار می گیرند. فرآیند اعتبارسنجی متقابل K بار تکرار می شود. سپس میانگین K نتایج حاصل، می تواند به عنوان یک برآوردگر واحد باشد. مزیت این روش بر نمونه گیری تصادفی این است که تمام مشاهدات برای آموزش و آزمون مورد استفاده قرار می گیرند و هر یک از مشاهدات دقیقاً یک بار برای نمونه آزمون استفاده می شود.

فرض کنید یک مجموعه مشاهده به طور تصادفی به K زیرمجموعه ی $d_1, \dots, d_K$ که دو به دو ناسازگار هستند و اندازه های تقریباً یکسانی دارند، تقسیم می شود. فرآیند آموزش و آزمون K بار اجرا می شود. به طوری که در تکرار i ام، زیرمجموعه ی d_i به عنوان مجموعه ی آزمون و زیرمجموعه های دیگر به عنوان مجموعه ی آموزشی مدل، در نظر گرفته می شوند. بنابراین در تکرار اول با زیرمجموعه های $d_2, \dots, d_K$ مدل ساخته شده و با زیرمجموعه ی d_1 ، مدل مورد آزمون و ارزیابی قرار می گیرد [۱۳].

پیوست ب

دستورات نرم افزار R

```
path <- ""
setwd(paste(path, "reproducible_files/
Simulation_Studies/R_Objects/", sep = ""))
require(snowfall)
sfInit(parallel = TRUE, cpus = 100, type = "MPI")
sfLibrary(party)
sfLibrary(mvtnorm)
sfLibrary(ROCR)

perform_simulation <- function(seed,
ncat          = 6,
mixing        = 0.6,
correlations  = TRUE,
nobs          = 200,
scores        = NULL,
nnoise        = 50, # must be more than 40
effect11      = 0.5,
```

```

effect12      = 0.75,
effect13      = 1,
effect2       = 1,
corstrength   = 0.8,
ncor          = 6,
get.tree.fractions.only = FALSE,
mtry          = NULL){
  if(!is.null(scores)){
    if(!length(scores) == ncat)
      stop("Score vector
must have the same
length as number of
      categories.")
    else scores <- list(y = scores)
  }else{
    scores <- list(y = 1:ncat)
  }
  if(is.null(mtry)) mtry <- floor(sqrt(nnoise + 15))
  traindata <- generate_data
  (seed = seed, n = nobs, ncat = ncat,
  mixing = mixing, correlations = correlations,
  nnoise = nnoise, ncor = ncor,
  effect11 = effect11, effect12 = effect12,
  effect13 = effect13, effect2 = effect2,
  corstrength = corstrength)
  if(get.tree.fractions.only){
    result <- view_first_split(seed = seed,
    traindata = traindata, mixing = mixing,
    scores = scores, mtry = mtry)
  }else{
    testdata <- generate_data(seed = seed * 10000,
    n = 10000, ncat = ncat, mixing = mixing,
    correlations = correlations, nnoise = nnoise,
    ncor = ncor, effect11 = effect11,

```

```

effect12 = effect12, effect13 = effect13,
effect2 = effect2, corstrength = corstrength)
result <- get_PA_VI(seed = seed,
traindata = traindata, testdata = testdata,
scores = scores, mtry = mtry, mixing = mixing)
}
return(result)
}

generate_data <- function(seed, n = nob,
ncat, mixing, correlations, nnoise, ncor,
effect11, effect12, effect13, effect2, corstrength){
ncat_half <- ceiling(ncat/2)
if(ncat %% 2 == 1){
grid <- ncat_half*(ncat_half+1)/2-0.5
thres <- 9/grid * c(0.5, 2:ncat_half)
thresholds <- c(-rev(cumsum(thres)),
cumsum(thres))
}else{
grid <- ncat_half*(ncat_half+1)/2
thres <- 9/grid * c(1:ncat_half)
thresholds <- c(-rev(cumsum(thres)), 0,
cumsum(thres))
}
thresholds[1] <- -Inf
thresholds[ncat+1] <- Inf
beta1 <- rep(c(effect13, effect12, effect11, 0),
c(5, 5, 5, nnoise))
beta2 <- c(rep(c(effect2, effect2, - effect2,
- effect2, 0), length = 15), rep(0, nnoise))
set.seed(seed)
tooless <- TRUE
while(tooless){
if(correlations == TRUE)

```

```

signal_sigma08 <- matrix(corstrength, nrow = ncor, ncol = ncor)
noise_sigma08  <- matrix(0.8, nrow = 10, ncol = 10)
noise_sigma06  <- matrix(0.6, nrow = 10, ncol = 10)
noise_sigma04  <- matrix(0.4, nrow = 10, ncol = 10)
noise_sigma02  <- matrix(0.2, nrow = 10, ncol = 10)

diag(signal_sigma08) <-
diag(noise_sigma08) <-
diag(noise_sigma06) <-
diag(noise_sigma04) <-
diag(noise_sigma02) <- 1

X_cor_signal    <- rmvnorm(n = n, sigma = signal_sigma08)
X_cor_noise08   <- rmvnorm(n = n, sigma = noise_sigma08)
X_cor_noise06   <- rmvnorm(n = n, sigma = noise_sigma06)
X_cor_noise04   <- rmvnorm(n = n, sigma = noise_sigma04)
X_cor_noise02   <- rmvnorm(n = n, sigma = noise_sigma02)

X_uncor_noise   <- rmvnorm(n = n, mean = rep(0, nnoise - 40))
X_uncor_signal  <- rmvnorm(n = n, mean = rep(0, 15 - ncor))

X <- cbind(X_cor_signal[,1],
X_uncor_signal[,1],
X_cor_signal[,2],
X_uncor_signal[,c(2, 3)],
X_cor_signal[,3],
X_uncor_signal[,4],
      X_cor_signal[,4],
X_uncor_signal[,c(5, 6)],
X_cor_signal[,5],
X_uncor_signal[,7],
X_cor_signal[,6],
X_uncor_signal[,c(8, 9)],
X_cor_noise08,

```

```

X_cor_noise06,
X_cor_noise04,
X_cor_noise02,
X_uncor_noise)
}else{
X <- matrix(rnorm(n =
  n * (nnoise + 15)), nrow = n)
}

eta1 <- apply(X %*% beta1, 1,
  function(z) z +
  thresholds[-(ncat + 1)])
eta2 <- apply(X %*% beta2,
  1, function(z) z + thresholds
  [-(ncat + 1)])
p_1 <- exp(eta1)/(1 + exp(eta1))
p_2 <- exp(eta2)/(1 + exp(eta2))
cumprob <- mixing *
p_1 + (1 - mixing) * p_2

cumprob <- rbind(cumprob, 1)
prob <- diff(cumprob)
y_mat <- apply(prob, 2,
  function(z) rmultinom
  (n = 1, size = 1, prob = z))
y <- apply(y_mat, 2, which.max)

tooleess <- min(table(y)) < 10
}

return(data.frame(y = y, X = X))
}

get_PA_VI <- function
(seed, traindata,
```

```

        testdata,scores,
        mtry, mixing){
set.seed(seed)
traindata$y <- factor
(traindata$y,
ordered = TRUE)
RF_ordinal <- cforest(y ~ .,
data      = traindata,
scores    = scores,
controls  =
cforest_control
(teststat  = "quad",
testtype  = "Univ",
mtry      = mtry,
minbucket = 0,
minsplit  = 0,
mincriterion = 0,
replace   = FALSE,
ntree     = 1000))
traindata$y <- factor
(traindata$y, ordered = FALSE)

RF_classification <- cforest(y ~ .,
data      = traindata,
controls  =
cforest_control(teststat = "quad",
testtype      = "Univ",
mtry          = mtry,
minbucket     = 0,
minsplit      = 0,
mincriterion  = 0,
replace       = FALSE,
ntree         = 1000))

```

```

PA_ordinal      <- get_PA(
RF = RF_ordinal,
testdata = testdata)
PA_classification <- get_PA(RF =
RF_classification,
testdata = testdata)

AUC_ordinal    <- get_AUC
(RF = RF_ordinal,
scores = scores$y, mixing = mixing)
AUC_classification <-
get_AUC(RF = RF_classification,
scores = scores$y, mixing = mixing)

result <- c(PA_ordinal,
PA_classification,
AUC_ordinal,
AUC_classification)
names(result) <- c("RPS_RF_ordinal",
"ER_RF_ordinal",
"RPS_RF_classification",
"ER_RF_classification",
"AUC_ERVIM_RF_ordinal",
"AUC_RPSVIM_RF_ordinal",
"AUC_MSEVIM_RF_ordinal",
"AUC_MAEVIM_RF_ordinal",
"AUC_ERVIM_RF_classification",
"AUC_RPSVIM_RF_classification",
"AUC_MSEVIM_RF_classification",
"AUC_MAEVIM_RF_classification")
return(result)
}

pred_class <- as.numeric(sapply
(pred_prob, which.max))

```

```

ER <- mean(as.numeric(pred_class)
!= as.numeric
(testdata$y))
return(c(RPS = RPS, ER = ER))
}

view_first_split <- function(seed, traindata,
mixing, scores, mtry){
set.seed(seed)
if(mixing == 0) effect.pred <-
c(1:4, 6:9, 11:14)
else effect.pred <- 1:15
index <- lapply(1:1000,
function(x)
list(sample(1:(ncol(traindata)-1),
size = mtry),
sample(1:nrow(traindata),
size = floor(0.632 *
nrow(traindata))))))
traindata$y <- factor
(traindata$y,
ordered = TRUE)
tree_ordinal <- ctree(y ~ .,
data = traindata[obs.index,
c(1, (pred.index + 1))],
controls = ctree_control
(teststat      = "quad",
testtype      = "Univ",
minbucket     = 0,
minsplit      = 0,
mincriterion   = 0,
savesplitstats = FALSE,
stump         = TRUE))

traindata$y <- factor(traindata$y,

```

```

ordered = FALSE)
tree_classification <- ctree(y ~ .,
data = traindata[obs.index,
c(1, (pred.index + 1))],
controls = ctree_control
(teststat      = "quad",
testtype      = "Univ",
minbucket     = 0,
minsplit      = 0,
mincriterion  = 0,
savesplitstats = FALSE,
stump         = TRUE))
return(c(tree_ordinal@tree@.
Data[[5]]$variableName,
tree_classification@tree@.
Data[[5]]$variableName))
}

sel.pred <- sapply(index,
function(z)grow_trees
(pred.index = z[[1]],
obs.index = z[[2]]))

obtain_tree_frac <-
function(pred)
mean(pred %in% paste
("X.", effect.pred, sep = ""))

fraction <- apply(sel.pred,
1, obtain_tree_frac)
return(fraction)
}

prediction_performance <- function
(seed, data, yname,
scores = NULL,
mtry = NULL, ncv = 10){
set.seed(seed)

```

```

ncat <- length(levels(data[, yname]))
if(!is.null(scores)){
  if(!length(scores) == ncat) stop
  ("Score vector must have
the same length as
number of categories.")
else scores <- eval
(parse(text = paste("list
(", yname, " = scores)")))
}else{
  scores <- eval(parse(text = paste
("list(", yname, " = 1:ncat)")))
}
if(is.null(mtry))
mtry <- floor(sqrt(ncol(data) - 1))
cvid <- rep(NA, length = nrow(data))
for(i in levels(data[, yname]))
cvid[data[, yname] == i] <-
  sample(rep(1:ncv, length =
sum(data[, yname] == i)))

result <- sapply(1:ncv, function(cv){
  tryCatch({
    RFs <- fit_RF(seed = seed,
traindata = data[cvid != cv,],
yname = yname,
scores = scores,
mtry = mtry)
    PA_ordinal
    <- get_PA(RF = RFs$RF_ordinal,
      testdata = data[cvid == cv,],
      yname = yname)
    PA_classification <-
    get_PA(RF = RFs$RF_classification,
```

```
testdata = data[cvid == cv,],
yname = yname)

result <- c(PA_ordinal, PA_classification)

names(result) <- c("RPS_RF_ordinal",
"ER_RF_ordinal",
"RPS_RF_classification",
"ER_RF_classification")
return(result)
}

, error = function(e)
{return(c(RPS_RF_ordinal = NA,
ER_RF_ordinal = NA,
RPS_RF_classification = NA,
ER_RF_classification = NA))})
})

return(rowMeans(result, na.rm = TRUE))
}

variable_importance <-
function(seed, data, yname,
scores = NULL, mtry = NULL){
set.seed(seed)

ncat <- length(levels(data[, yname]))

if(!is.null(scores)){
if(!length(scores) == ncat)
stop("Score vector must
have the same length as
number of categories.")
else scores <-
eval(parse(text =
paste("list(", yname, " =
```

```

scores)))))
}else{ scores <-
eval(parse(text =
paste("list(", yname, " =
1:ncat)"))))
}

p <- ncol(data) - 1
data <- data[,
c(which(names(data) == yname),
rep(which(names(data)
!= yname), 2))]
perm <- sample(1:nrow(data))
data[, (p + 2) : (2*p + 1)] <-
data[perm, (p + 2) : (2*p + 1)]
if(is.null(mtry)) mtry <-
floor(sqrt(ncol(data) - 1))
RFs <- fit_RF(seed = seed,
traindata = data,
yname = yname,
scores = scores,
mtry = mtry)

AUC_ordinal <-
get_AUC(RF = RFs$RF_ordinal,
scores = scores[[1]])
AUC_classification <-
get_AUC(RF = RFs$RF
_classification,
scores = scores[[1]])

result <- c(AUC_ordinal,
AUC_classification)
names(result) <- c(

```

```

    "AUC_ERVIM_RF_ordinal",
    "AUC_RPSVIM_RF_ordinal",
    "AUC_MSEVIM_RF_ordinal",
    "AUC_MAEVIM_RF_ordinal",
    "AUC_ERVIM_RF_classification",
    "AUC_RPSVIM_RF_classification",
    "AUC_MSEVIM_RF_classification",
    "AUC_MAEVIM_RF_classification")

return(result)
}

get_AUC <- function(RF, scores){
VIM_ER  <- varimp(RF)
VIM_RPS <- varimpRPS(RF)
VIM_MSE <- varimpMSE(RF, scores = scores)
VIM_MAE <- varimpMAE(RF, scores = scores)
effect_ind <- rep(c(1, 0),
each = length(VIM_ER)/2)
return(sapply(list(VIM_ER, VIM_RPS,
VIM_MSE, VIM_MAE), function(z)
unlist(performance(prediction(z, effect_ind),
measure = "auc")@y.values)))
}

RF_ordinal <- cforest(formula,
data      = traindata,
scores    = scores,
controls  =
cforest_control(teststat = "quad",
testtype     = "Univ",
mtry         = mtry,
minbucket    = 0,
minsplit     = 0,
mincriterion = 0,
```

```

replace      = FALSE,
ntree        = 1000))

traindata[, yname] <- factor(traindata
[, yname], ordered = FALSE)

RF_classification <- cforest(formula,
data      = traindata,
controls =
cforest_control(teststat      = "quad",
testtype      = "Univ",
mtry          = mtry,
minbucket     = 0,
minsplit      = 0,
mincriterion  = 0,
replace       = FALSE,
ntree         = 1000))

return(list(RF_ordinal = RF_ordinal,
  RF_classification = RF_classification))
}

pred_prob <- predict(RF,
newdata = testdata,
type = "prob")

hat_F      <- sapply(pred_prob,
function(x) cumsum(x))

indicator_mat <- matrix(as.numeric
(rep(testdata[, yname], each = ncat)),
  nrow = ncat, byrow = FALSE)

T_F          <- apply(indicator_mat, 2,
function(x) as.numeric(x <= 1:ncat))

RPS <- sum((hat_F - T_F)^2)/
nrow(testdata)

pred_class
<- as.numeric(sapply(pred_prob
, which.max))

```

```
ER <- mean(as.numeric(pred_class) !=  
  as.numeric(testdata[, yname]))  
  
return(c(RPS = RPS, ER = ER))  
}
```


مراجع

- [۱] مهدوی س، حسینی ر، (۱۳۹۴)، «ارایه یک سیستم استنتاج فازی جهت تشخیص سلامتی نوزادان»، سومین کنفرانس بین‌المللی پژوهش‌های کاربردی در مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی مالک اشتر، تهران
- [2] Agresti A. (2002), "Categorical Data Analysis", John Wiley, New York.
- [3] Boulesteix, A.L., Bender, A., Lorenzo Bermejo, J. and Strobl, C. (2011), "Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations", **Briefings in Bioinformatics**, 13(3), 292-304.
- [4] Boulesteix, A.L., Janitza, S., Kruppa, J. and König, I.R. (2012), "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics", **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, 2(6), 493-507.
- [5] Breiman, L. (2001) "Random forests", **Machine learning**, 45(1), 5-32.
- [6] Breiman, L.F., Olshen, J.H. and Stone, C.J. (1984), "Classification and Regression Trees, Belmont, CA: Wadsworth".
- [7] Cappelli, C., Mola, F., Siciliano, R. (2002), "A statistical approach to growing a reliable honest tree", **Computational statistics data analysis**, 38(3), 285-299.
- [8] Casey B.M., McIntire D.D. and Leveno K.J. (2001), "The continuing value of the Apgar score for the assessment of newborn infants", **New England Journal of Medicine**, 344(7), 467-471.
- [9] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009), "Modeling wine preferences by data mining from physicochemical properties", **Decision Support Systems**, 47(4), 547-553.

-
- [10] Epstein, E. S. (1969), "A scoring system for probability forecasts of ranked categories", **Journal of Applied Meteorology**, 8(6), 985-987.
- [11] Frank, E. and Hall, M. (2001), "A simple approach to ordinal classification", **In European Conference on Machine Learning**, 145-156.
- [12] Franklin, J. (2005), "The elements of statistical learning: data mining, inference and prediction", **The Mathematical Intelligencer**, 27(2), 83-85.
- [13] Friedman, J., Hastie, T. and Tibshirani, R. (2001), "**The elements of statistical learning**", Vol. 1, New York: Springer series in statistics.
- [14] Fürnkranz, J. and Hüllermeier, E. (2011), "**Preference learning**", In Encyclopedia of Machine Learning.
- Hechenbichler, K. and Schliep, K. (2004), "**Weighted k-nearest-neighbor techniques and ordinal classification**".
- [15] Hothorn, T., Hornik, K. and Zeileis A. (2012), "Party: a laboratory for recursive partytioning", R package version 10-3, URL <http://cranr-project.org/package=party>.
- [16] Hothorn, T., Hornik, K. and Zeileis, A. (2006), "Unbiased recursive partitioning: A conditional inference framework", **Journal of Computational and Graphical statistics**, 15(3), 651-674.
- [17] Hothorn, T., Hornik, K., Van De Wiel, M. A. and Zeileis, A. (2006), "A Lego system for conditional inference", **The American Statistician**, 60(3), 257-263.
- [18] Janitza, S., Binder, H. and Boulesteix, A. L. (2016), "Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications", **Biometrical Journal**, 58(3), 447-473.
- [19] Janitza, S., Strobl, C. and Boulesteix, A. L. (2013), "An AUC-based permutation variable importance measure for random forests", **BMC bioinformatics**, 14(1), 119.
- [20] Janitza S., Tutz G. and Boulesteix A. L. (2016), "Random forest for ordinal responses: prediction and variable selection", **Computational Statistics & Data Analysis**, 96, 57-73.
- [21] Karamanian V. A., Harhay, M., Grant, G. R., Palevsky, H. I., Grizzle, W. E., Zamanian, R. T., Ihida-Stansbury, K., Taichman, D. B., Kawut, S. M. and Jones, P. L. (2014), "Erythropoietin upregulation in pulmonary arterial hypertension", **Pulmonary circulation**, 4(2), 269-279.

- [22] Murphy A.H. (1970), "The ranked probability score and the probability score: A comparison", **Mon. Weather Rev.**, 98(12), 917–924.
- [23] Nicodemus, K.K. (2011), "Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures", **Briefings in bioinformatics**, 12(4), 369-373.
- [24] Nicodemus, K.K., Callicott, J.H., Higier, R. G., Luna, A., Nixon, D.C., Lipska, B. and et al. (2010), "Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging", **Human genetics**, 127(4), 441-452.
- [25] Liu, C., Ackerman, H.H. and Carulli, J.P. (2011), "A genome-wide screen of gene–gene interactions for rheumatoid arthritis susceptibility", **Human genetics**, 129(5), 473-485.
- [26] Loh, W.Y. (2011), "Classification and regression trees", **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, 1(1), 14-23.
- [27] Louppe G. (2014), PhD thesis, "Understanding random forests: From theory to practice", Department of Electrical Engineering & Computer Science, University of Liège.
- [28] O'Shea T.M., Kothadia J.M., Roberts D.D. and Dillard R.G. (1998), "Perinatal events and the risk of intraparenchymal echodensity in very-low-birthweight neonates", **Paediatric and perinatal epidemiology**, 12, 408-421.
- [29] Pepe, M.S. (2003), "**The statistical evaluation of medical tests for classification and prediction**", Medicine.
- [30] SCORE, A. (1986), "Use and abuse of the Apgar score". **Pediatrics**, 78(6).
- [31] Shih, Y.S. (1999), "Families of splitting criteria for classification trees", **Statistics and Computing**, 9(4), 309-315.
- [32] Shih, Y.S. (2004), "A note on split selection bias in classification trees", **Computational statistics & data analysis**, 45(3), 457-466.
- [33] Shin Y and Tsai H. (2004), "variable selection bias in regression trees with constant fits", **Computational Statistics & Data Analysis**, 45(3), 595-607.
- [34] Sun Y.V., Cai Z., Desai K., Lawrance R., Leff R., Jawaid A., Kardia S. and Yang H. (2007), "Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests", **BMC Proc**, 1(Suppl 1), S62.

-
- [35] Strasser, H. and Weber, C. (1999), "**On the asymptotic theory of permutation statistics**".
- [36] Strobl C., Boulesteix A.L. and Augustin T. (2007), "Unbiased split selection for classification trees based on the Gini index", **Computational Statistics Data Analysis**, 52(1), 483-501.
- [37] Strobl, C., Boulesteix, A.L., Zeileis, A., and Hothorn, T. (2007), "Bias in random forest variable importance measures: Illustrations, sources and a solution", **Briefings in Bioinformatics**, 8(1), 25.
- [38] Timofeev, R. (2004), "**Classification and regression trees (CART) theory and applications**", Humboldt University, Berlin.
- [39] Tutz G. (2011), "**Regression for categorical data**", vol. 34, Cambridge University Press.

واژه‌نامه فارسی به انگلیسی

Bootstrap aggregating	اجماع‌سازی خودگردان
Cross Validation	اعتبارسنجی متقابل
Proportional odds	بخت‌های متناسب
Impurity Function	تابع ناخالصی
Random Forest	جنگل تصادفی
Adjacent category	رده‌ی همسایه
Logistic regression	رگرسیون لجستیک
Data mining	داده‌کاوی
Decision tree	درخت تصمیم
Terminal Node	گره خارجی
Mean Absolute Error	میانگین قدرمطلق خطا
Mean Squared Error	میانگین مربعات خطا
Error Rate	نرخ خطا
Ranked Probability Score	امتیاز احتمال رتبه‌ای
Test sample	نمونه آزمون
Traning sample	نمونه آموزش
Bootstrap Sample	نمونه خودگردان
Node	گره

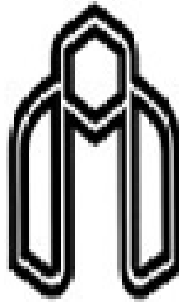
واژه‌نامه انگلیسی به فارسی

Adjacent category	رده‌ی همسایه
Bootstrap aggregating	اجماع‌سازی خودگردان
Bootstrap Sample	نمونه خودگردان
Cross Validation	اعتبارسنجی متقابل
Data mining	داده‌کاوی
Decision tree	درخت تصمیم
Error Rate	نرخ خطا
Impurity Fancion	تابع ناخالصی
Logistic regression	رگرسیون لجستیک
Mean Absolute Error	میانگین قدرمطلق خطا
Mean Squared Error	میانگین مربعات خطا
Node	گره
Random Forest	جنگل تصادفی
Ranked Probability Score	امتیاز احتمال رتبه‌ای
Proportional odds	بخت‌های متناسب
Terminal Node	گره خارجی
Test sample	نمونه آزمون
Traning sample	نمونه آموزشی

Aabstract

Random forest, as a combination of decision trees, is one of the most comprehensive and modern methods of data mining which is placed in the category of consensus bootstrap or bagging. This method is a common tool for classification in high dimensional data, so that it is able to predict the class of observations and determine the rank of predictor variables based on the importance of those variables. Random forests can be used for the quantitative, nominal or survival response variables. But when the response variable is ordinal, it is usually considered the response as a nominal one, which ignores the ordinal information and reduces the prediction accuracy. In this thesis, we use a modification of classical decision tree to construct a different type of random forest method for ordinal response variable. The results of simulation data and real data imply that by applying of these changes, the predictive performance has been improved in comparison with the classical random forest method.

Keywords: Decision tree, Conditional inference tree, Random forest, Ordinal random forest.



Shahrood University of Technology

Faculty of Mathematical Sciences

M.Sc. Thesis in Statistics

Modified Random Forests for Ordinal Response

By: Maryam Alibaygi

Supervisor

Dr. Davood Shamsavani

Jan 2018