دوماهنامه طبّ جنوب
پژوهشکده زیست- پزشکی خلیج فارس
دانشگاه علوم پزشکی و خدمات بهداشتی درمانی بوشهر
سال بیست و چهارم، شماره ۵، صفحه ۴۶۸-۴۵۴ (آذر و دی ۱۴۰۰)

Iran South Med J 2021; 24(5): 454-468

مقایسه روش‌های مدل‌بندی پاسخ ترتیبی از قبیل درخت تصمیم، انباشت تصادفی ترتیبی و رگرسیون نسبت پیوسته جریمه شده در داده‌های با ابعاد بالا

زهرا ترکاشوند (MSc)[۱*]، حسین محجوب (PhD)[۱و۲]، علیرضا سلطانیان (PhD)[۱و۳]،
مریم فرهادیان (PhD)[۱و۲**]

[۱] گروه آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی همدان، همدان، ایران
[۲] مرکز تحقیقات علوم بهداشتی، دانشکده بهداشت، دانشگاه علوم پزشکی همدان، همدان، ایران
[۳] مرکز تحقیقات مدلسازی بیماری‌های غیرواگیر، دانشکده بهداشت، دانشگاه علوم پزشکی همدان، همدان، ایران

(دریافت مقاله: ۱۴۰۰/۴/۱۴- پذیرش مقاله: ۱۴۰۰/۶/۱۴)

چکیده

زمینه: در بسیاری از تحقیقات در حوزه‌های پزشکی و بهداشتی متغیر پاسخ ماهیت ترتیبی دارد. روش‌های مرسوم مبتنی بر فرض استقلال میان متغیرهای پیشگو و همچنین زیاد بودن تعداد نمونه‌ها (**n**) در مقایسه با تعداد کوواریت‌ها (**p**) هستند. لذا برای داده‌های ژنتیکی با ابعاد بالا که در آن‌ها **p>n** می‌باشد، استفاده از مدل‌های مرسوم امکان‌پذیر نیست. در پژوهش حاضر از روش‌های رگرسیون نسبت پیوسته جریمه شده، درخت تصمیم و انباشت ترتیبی برای پیش‌بینی پاسخ‌های ترتیبی استفاده خواهد شد.

مواد و روش‌ها: در مطالعه حاضر از سه دیتاست استفاده شد. مجموعه داده **B-cell** حاوی اطلاعات ۱۲۶۲۵ ژن در ۱۲۸ بیمار که پاسخ در چهار سطح ترتیبی قرار داشت، داده **HCC** مرتبط با سرطان کبد شامل ۱۴۶۹ ژن در ۵۶ بیمار که پاسخ در سه سطح ترتیبی قرار داشت و همچنین داده قلب شامل اطلاعات پنج متغیر در ۲۹۴ بیمار تحت آنژیوگرافی که پاسخ در ۵ سطح قرار داشت. عملکرد روش‌های مدنظر با استفاده از مجموعه داده یکسان آموزش و آزمون براساس شاخص‌هایی از قبیل دقت، گاما و کاپا مورد مقایسه قرار گرفت.

یافته‌ها: در دو مجموعه داده با ابعاد بالا مدل انباشت ترتیبی از توانایی پیش‌بینی بالاتری برخوردار بود. در حالی که برای مجموعه داده با ابعاد پایین مدل رگرسیون نسبت پیوسته جریمه شده عملکرد پیش‌بینی بهتری داشت.

نتیجه‌گیری: انتخاب بهترین مدل پیش‌بینی از بین مدل‌های بکار رفته بستگی به مجموعه داده مورد استفاده دارد و برای هر مجموعه داده بایستی روش‌های مختلف را مورد بررسی قرار داد تا به بهترین مدل دست یافت.

واژگان کلیدی: پاسخ ترتیبی، روش رگرسیون نسبت پیوسته جریمه شده، روش انباشت ترتیبی، داده‌های بیان ژن

**همدان، گروه آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی همدان

Email: maryam_farhadian80@yahoo.com
***ORCID: 0000-0003-4321-3842**
****ORCID: 0000-0002-6054-9850**

مقدمه

در بسیاری از مطالعات معمولاً یک متغیر پاسخ برای تحقیق مدنظر است. در مدل‌سازی آماری، هدف پیش‌بینی یا درک عمیق‌تر نسبت به رابطه بین متغیر وابسته یا پاسخ و مجموعه‌ای از متغیرهای مستقل یا پیشگو است. هدف بسیاری از روش‌های مدل‌سازی، تعیین رابطه بین دو متغیر (مستقل و وابسته) است تا بتوان یک متغیر پاسخ را پیش‌بینی کرد و یا درک بهتری نسبت به رابطه بین دو متغیر بدست آورد. طیف گسترده‌ای از مدل‌ها برای اندازه‌گیری و ارزیابی روابط بین متغیرهای وابسته و مستقل وجود دارد. انتخاب مدل اغلب توسط مقیاس اندازه‌گیری متغیر پاسخ حاصل می‌شود. مقیاس متغیر پاسخ می‌تواند یکی از چهار نوع مختلف اسمی، ترتیبی، فاصله‌ای یا نسبتی باشد (۱).

در بسیاری از تحقیقات در حوزه‌های پزشکی و بهداشتی متغیر پاسخی که قرار است پیش‌بینی شود و یا مورد مقایسه قرار بگیرد، ماهیت ترتیبی دارد (۲). برخی از مثال‌های مرتبط با متغیرهای رتبه‌ای عبارتند از: سطوح سمیت مواد که با سطح خفیف، متوسط یا شدید ارزیابی می‌شوند. سیستم طبقه‌بندی گزارشات تصویربرداری پستان بر اساس اختصاص نمره ذهنی وضعیت بافت پستان در تصاویر ماموگرافی شامل دسته‌بندی‌های ناقص؛ منفی؛ خوش‌خیم؛ احتمالاً خوش‌خیم؛ اختلال مشکوک؛ بسیار مشکوک به بدخیمی؛ و رده بدخیمی، همچنین کمیت دارویی که با سطوح خفیف، متوسط و شدید ارزیابی می‌شود و نیز نمونه‌های بافت که می‌تواند در طبقات نرمال، بدخیمی خفیف و بدخیمی شدید دسته‌بندی شوند (۱).

در سال‌های اخیر با ظهور فناوری به نام ریزآرایه[1] دی ان ای، امکان بررسی و مطالعه فعالیت هزاران ژن به‌طور همزمان فراهم آمده است. استفاده از فناوری ریزآرایه حجم انبوهی از داده‌ها را مرسوم به داده‌های بیان ژنی[2] تولید می‌کند که مشخصه بارز آن‌ها تعداد بسیار زیاد متغیرها (ژن‌ها) نسبت به تعداد داده‌ها (بیماران) است. تجزیه و تحلیل داده‌های حاصل از ریزآرایه دریچه‌های جدید را به روی محققان گشوده است و از این حیث، توجه بسیاری از آن‌ها بر روی تجزیه و تحلیل این گونه داده‌ها معطوف گشته است. به عنوان مثال در مراحل سرطان به‌عنوان پاسخ ترتیبی، ویژگی‌های مولکولی به‌طور یکنواختی با پاسخ در ارتباط هستند، به عبارتی افزایش در سطوح فنوتیپ ترتیبی در یک ارتباط یکنواخت افزایشی یا کاهشی با سطح بیان ژن می‌باشد (۳). اگرچه می‌توان روش‌های طبقه‌بندی پاسخ اسمی را برای طبقه‌بندی پاسخ‌های ترتیبی بکار برد، در این شرایط بخشی از اطلاعات مرتبط با ماهیت ترتیبی نادیده گرفته می‌شود که می‌توان عملکرد روش طبقه‌بندی را تحت تأثیر قرار دهد، بنابراین استفاده از طبقه‌بندی کننده‌های اسمی نمی‌تواند این ارتباط یکنواخت را به خوبی شناسایی کند (۲).

روش‌های بسیاری برای مدل‌سازی پاسخ ترتیبی از قبیل مدل‌های لوجیت تجمعی، مدل‌های بخت متناسب و مدل‌های ربط تجمعی یا مدل‌های پروبیت تجمعی، وجود دارد. اما در این روش‌ها فرض استقلال میان متغیرهای پیشگو و همچنین زیاد بودن تعداد نمونه‌ها (n) در مقایسه با تعداد کوواریت‌ها (p) بایستی در نظر گرفته شود. لذا برای داده‌های ژنتیکی با ابعاد بالا که در آن‌ها $p>n$ می‌باشد، استفاده از مدل‌های مرسوم امکان‌پذیر نیست (۲). مطالعات متعددی نشان داده است که روش‌های یادگیری ماشین از قبیل انباشت تصادفی در مقایسه با روش‌های مرسوم رگرسیونی عملکرد بهتری در پیش بینی پاسخ دارند (۴).

[1] Microarray
[2] Gene expression data

در پژوهش حاضر از روش‌های رگرسیون نسبت پیوسته جریمه شده[3]، درخت تصمیم، انباشت ترتیبی[4]، برای پیش‌بینی پاسخ‌های ترتیبی در داده‌هایی با ابعاد بالا و پایین استفاده خواهد شد. همچنین عملکرد این روش‌ها با مجموعه داده یکسان آموزش و آزمون مورد مقایسه قرار خواهد گرفت.

مواد و روش‌ها

داده‌ها: در مطالعه حاضر از سه مجموعه داده به شرح زیر استفاده شده است:

مجموعه داده B cell:

ALL یک مجموعه داده است که توسط جنتلمن (Gentleman) و همکاران تولید شده است (۵). شامل ۱۲۶۲۵ بیان ژن و داده‌های فنوتیپی مرتبط با ۱۲۸ بیمار (۹۵ نفر با سلول نوع B و ۳۳ بیمار با سلول نوع T) است. در مطالعه حاضر اطلاعات ۹۰ بیمار مبتلا به لوسمی سلول B که شامل رده‌های ترتیبی B1، B2، B3 یا B4 با تعداد ۳۸۴۱ متغیرهای بیان ژن، استفاده شده است. از ۹۰ بیمار موجود در این مجموعه داده به صورت تصادفی ۶۰ نفر در مجموعه دیتای آموزشی و ۳۰ نفر در مجموعه دیتای آزمایشی قرار گرفتند. از ۳۰ بیماری که در مجموعه دیتای آزمایشی قرار دارند ۶ نفر در رسته B1، ۱۲ نفر در رسته B2، ۸ نفر در رسته B3 و ۴ نفر در رسته B4 قرار گرفته‌اند. داده‌ها از آدرس زیر قابل دسترسی هستند:

https://bioconductor.org/packages/release/data/experiment/html/ALL.html

مجموعه داده سرطان کبد (HCC: Hcc Cancer Panel)

این مجموعه داده مرتبط با سرطان کبد شامل زیرمجموعه‌ای از افراد و سایت‌های CpG (جزایر هیپرمتیلاسیون CpG که در مناطق پروموتوری قرار دارند) است. مجموعه کامل این دیتاست با عنوان GSE۱۸۰۸۱ در سایت Gene Expression Omnibus موجود می‌باشد.

این مجموعه داده شامل اطلاعات ۵۶ بیمار و ۱۴۶۹ سایت CpG در سه رده از بیماری (Normal < Cirrhosis non-HCC < Tumor) بود. از این تعداد، ۲۰ بیمار در وضعیت نرمال، ۱۶ بیمار در رده سیروتیکی فاقد HCC و ۲۰ بیمار در رده سیروتیکی دارای HCC قرار دارند. به‌صورت تصادفی ۳۷ نفر از بیماران در مجموعه دیتای آموزشی و ۱۹ نفر در مجموعه دیتای آزمایشی قرار گرفتند. از ۱۹ بیمار که در مجموعه دیتای آزمایشی قرار دارند ۶ نفر در دسته نرمال، ۶ نفر در دسته سیروتیکی فاقد HCC و ۷ نفر در دسته دارای HCC قرار گرفته‌اند. داده از آدرس زیر قابل دسترسی هستند:

https://www.bioconductor.org/packages/ release/bioc/html/GEOquery.html

مجموعه داده بیماری قلبی (Heart)

داده‌های این مطالعه شامل ۲۹۴ بیمار تحت آنژیوگرافی طی سال‌های ۱۹۸۳ تا ۱۹۸۷ در مجارستان است. متغیر پاسخ شدت بیماری عروق کرونر است که با استفاده از آنژیو گرافی تعیین شده است و شامل ۵ طبقه ترتیبی می‌باشد. متغیرهای مستقل نیز شامل سن، فشارخون در حال استراحت، کلسترول سرم، حداکثر ضربان قلب، افسردگی ST ناشی از ورزش نسبت به استراحت

[3] L_1 penalized continuation ratio
[4] Ordinal forest

می‌باشد. در این مطالعه ۱۹۶ نفر در مجموعه دیتای آموزشی و ۹۸ نفر در مجموعه دیتای آزمایشی قرار گرفتند. از ۹۸ بیمار در مجموعه دیتای آزمایشی ۶۱ نفر در دسته (۱)، ۱۲ نفر در دسته (۲)، ۱۰ نفر در دسته (۳)، ۸ نفر در دسته (۴)، ۷ نفر در دسته (۵) قرار گرفته‌اند. داده از آدرس زیر قابل دسترسی هستند:
https://www.openml.org/d/1565/

در این مطالعه برای مدل‌بندی پاسخ ترتیبی از روش‌های رگرسیون نسبت پیوسته جریمه شده، درخت تصمیم و روش انباشت تصادفی ترتیبی استفاده شده است. در ادامه به توضیح هریک از روش‌ها پرداخته می‌شود:

رگرسیون نسبت پیوسته جریمه شده

مدل رگرسیونی نسبت پیوسته، لوجیت پاسخ را با طبقه‌های پاسخ ترتیبی به‌صورت دنباله‌ای مدل‌سازی می‌کند. در واقع این مدل بر اساس احتمال‌های تجمعی به‌دست می‌آید. فرض کنید بردار پاسخ y_i متعلق به کلاس رتبه‌ای kام است $k=(1,\ldots,K)$: تعداد طبقات متغیر پاسخ ترتیبی و n تعداد افراد است. xi نیز نشان‌دهنده یک بردار با اندازه p از متغیرها می‌باشد. در این روش مدل به صورت زیر ساخته می‌شود.

$$\operatorname{logit}\big(P(y=k|y\le k.X=x)\big)=\log\left(\frac{P(y=k|y\le k.X=x)}{P(y<k|y\le k.X=x)}\right)=\alpha_k+\beta_k^{T}x \quad (۱)$$

که در آن β بردار پارامترها برای (K-1) طبقه متغیر پاسخ می‌باشد. لازم به ذکر است در این مدل یک طبقه از پاسخ به‌عنوان طبقه مبنا یا رفرنس در نظر گرفته می‌شود و مقدار $(\alpha_2\ldots.\alpha_k)$ و بردار پارامتر $(\beta_1 \ldots. \beta_p)$ برای تمام طبقات یکسان در نظرگرفته می‌شود.

برای برازش چنین مدلی (K-1) زیرمجموعه داده با استفاده از اطلاعات مجموعه داده‌های اولیه به این صورت بازسازی می‌شوند: به‌عنوان مثال اگر تعداد طبقات پاسخ شامل ۳ طبقه باشد، تعداد ۲ زیرمجموعه (مجموعه اول شامل مشاهدات طبقه ۱ و ۲، همچنین مجموعه دوم شامل مشاهدات طبقات ۱ تا ۳) تشکیل می‌شود. در هر مجموعه بازسازی شده اگر پاسخ متعلق به طبقه مورد نظر باشد کد ۱ و در غیر اینصورت کد صفر را دریافت خواهد کرد. با به کارگیری مدل رگرسیون لوجستیک به این مجموعه داده‌های بازسازی شده، مدل رگرسیون نسبت پیوسته جریمه شده تشکیل می‌گردد (۵).

روش لاسو

یکی از راه‌های کنترل تعداد متغیرهای وارد شده به مدل، اضافه نمودن جریمه به تابع درستنمایی می‌باشد. روش مرسوم به لاسو که توسط تیب شیرانی (Tibshirani) پیشنهاد شد، یک روش حداقل مربعات جریمه شده است که یک جریمه L_1 (درجه یک) را به ضرایب رگرسیون تحمیل می‌نماید. بنابراین با اعمال جریمه لاسو به مدل رگرسیون نسبت پیوسته جریمه شده، برآورد ضرایب مدل جواب مسئله زیر می‌باشد:

$$\hat{\beta}=\arg\max_{\beta}\left\{L(\beta|y.x)-\text{lambda}\sum_{m=1}^{p}|\beta_m|\right\} \quad (۲)$$

که در آن $\beta=(\beta_1 \ldots \beta_P)^T$ پارامترهای مدل و $L(\beta|y,x)$ لگاریتم تابع درستنمایی است، $Lambda \geq 0$ پارامتر تنظیم کننده شدت انقباض متغیرهای پیشگو و m نیز تعداد متغیرهای توضیحی است (۶).

درخت رده‌بندی ترتیبی

درخت‌های تصمیم از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و هدف در این فرآیند افزایش فاصله بین گروه‌ها یا طبقات در هر جداسازی است. ریشه درخت شامل تمام نمونه آموزشی است که باید در طبقات مختلف تقسیم شود. در هر یک از گره‌های داخلی فضای نمونه بر اساس یک یا چند متغیر به دو یا چند قسمت تقسیم و بهترین تقسیم‌بندی به ازای کمینه کردن معیار ناخالصی انتخاب می‌شود (۷ و ۸).

فرض کنید بردارکوواریت‌ها با $\mathbf{x_i} = (x_{i1},\ldots,x_{ip})$ نشان داده می‌شود و هر فرد بر اساس بردار مشاهده x_1 در یکی از K کلاس قرار می‌گیرد. اگر متغیر رده‌بندی یا طبقه پاسخ را با w نشان دهیم، w_1 نشان‌دهنده مشاهدات کلاس ۱ می‌باشد و به همین ترتیب w_k نیز نشان‌دهنده مشاهدات کلاس K می‌باشند. برای استخراج یک درخت تصمیم تمام مشاهدات از گره ریشه شروع می‌شوند، بنابراین بهترین تقسیم‌بندی (تقسیم‌بندی بهینه) برای پیشگوهای اول، دوم، ...، pام براساس تقسیم‌بندی که منجر به کمینه کردن معیار ناخالصی است صورت می‌گیرد. برای هر گره t، این تقسیم‌بندی بهینه، مشاهدات را به گره‌های سمت راست (t_R) و سمت چپ (t_L) تقسیم می‌کند. نسبتی از مشاهدات (افراد) که به ازای این گره‌ها در هر کلاس قرار می‌گیرند را با $P(w_k|t)$ که در آن $k=1,\ldots,K$ است، نشان می‌دهیم و داریم:

$$p(\omega_1|t) + p(\omega_2|t)+\ldots+ p(\omega_k|t) = 1 \qquad (۴)$$

مهم‌ترین جزء متمایز یک درخت، چگونگی انتخاب یک ضابطه برای تقسیم‌بندی مجموعه داده‌ها در هر گره درخت است. انتخاب یک ضابطه افراز به معنای انتخاب یک متغیر پیش‌بین از میان متغیرها و انتخاب بهترین سطح افراز است. برای طبقه‌بندی پاسخ اسمی، معیار ناخالصی درون‌گره‌ای که غالباً مورد استفاده قرار می‌گیرد معیار جینی است که به صورت زیر تعریف می‌شود:

$$i(t) = \sum_{k}\sum_{k\neq l} p(\omega_k|t)p(\omega_l|t) \qquad (۳)$$

با این حال به کارگیری این معیار ناخالصی، اطلاعات اضافی موجود در پاسخ ترتیبی را نادیده می‌گیرد. برای حل این مشکل از تابع ناخالصی جینی تعمیم یافته[5] و روش تقسیم‌بندی دوتایی[6] معرفی شده توسط بریمن (Breiman)، در پاسخ‌های ترتیبی استفاده می‌شود (۸ و ۹). معیار تابع ناخالصی جینی تعمیم یافته[7] به صورت زیر تعریف می‌شود که در آن $C(w_K|w_L)$ هزینه

[5] Generalized Gini impurity function
[6] Ordered twoing method
[7] Generalized Gini impurity function

(جریمه) طبقه‌بندی نادرست مشاهدات کلاس L در کلاس K برای یک پاسخ ترتیبی است (۱۰).

$$iGG(t) = \sum_{k=1}^{J}\sum_{L=1}^{J} C(\omega_k|\omega_L)\, p(\omega_k|t)p(\omega_L|t) \quad (۴)$$

در روش تقسیم‌بندی دوتایی[8] با فرموله کردن پاسخ ترتیبی به شکل یک پاسخ دو حالته به صورت زیر که در آن برای هریک از مشاهدات iام، jامین پاسخ به شکل زیر است:

$$c_{ij} = \begin{cases} 1 & \text{if } \omega_i = 1.\ldots.j \\ 0 & \text{if } \omega_i = j+1.\ldots.J \end{cases} \quad (۵)$$

برای گره tام و پاسخ دوحالتی c_j، تقسیم‌بندی s که رابطه زیر را ماکزیمم می‌کند، برای تمام کووریت‌ها صورت می‌گیرد تا در نهایت بهترین تقسیم‌بندی یافت شود (۱۱).

$$\emptyset(s.t.\, c_j) = 2p_l p_R (p(c_j|t_l) - p(c_j|t_R))^2 \quad (۶)$$

هرس درخت[9] رده‌بندی ترتیبی

هرس روشی است که باعث کاهش اندازه درخت‌های تصمیم‌گیری با از بین بردن بخش‌هایی از درخت است که قدرت کمی برای طبقه‌بندی نمونه‌ها دارند. هرس باعث کاهش پیچیدگی طبقه‌بندی کننده نهایی می‌شود و از این رو باعث بهبود دقت پیش‌بینی و کاهش بیش برازش می‌شود. چندین روش برای به‌دست آوردن درختان با اندازه صحیح پیشنهاد شده است (۱۵-۱۲). یکی از محبوب‌ترین تکنیک‌های فرآیند هرس که توسط بریمن و همکاران پیشنهاد شده است، بر اساس معیار کمینه هزینه – پیچیدگی[10] است.

روش انباشت ترتیبی (OF)[11]

انباشت تصادفی یک الگوریتم یادگیری نظارت شده از نوع روش‌های گروهی است. این الگوریتم انباشتی را به صورت تصادفی متشکل از گروهی از درخت‌های تصمیم می‌سازد. روش ساخت انباشت با استفاده از درخت‌ها غالباً به روش کیسه‌گذاری انجام می‌شود. ایده اصلی در روش کیسه‌گذاری آن است که استفاده از ترکیبی از مدل‌های یادگیری، دقت کلی مدل را افزایش می‌دهد. به بیان ساده انباشت تصادفی چندین درخت تصمیم ساخته و آن‌ها را با یکدیگر ادغام می‌کند، تا پیش‌بینی‌های صحیح‌تر و پایدارتری حاصل شوند. الگوریتم انباشت تصادفی از تمام مشاهدات نمونه برای

[8] Ordered twoing method
[9] Pruning the tree
[10] Minimal cost-complexity pruning
[11] Ordinal forest (OF)

ساخت درخت استفاده نمی‌کند، بلکه یک نمونه تصادفی با جایگذاری به حجم n_1 (معمولاً برابر $2_n/3$ از مشاهدات انتخاب می‌شود. به مشاهدات انتخاب شده نمونه آزمایشی و به بقیه آن‌ها نمونه خارج کیسه گفته می‌شود. درخت‌های تصمیم با مشاهدات نمونه آزمایشی ساخته می‌شوند و از نمونه‌های خارج کیسه برای اندازه‌گیری ناخالصی درخت استفاده می‌شود (۱۴). روش انباشت ترتیبی در واقع یک روش مبتنی بر انباشت تصادفی است که برای پاسخ ترتیبی به‌کار می‌رود. در این شرایط امکان پیش‌بینی برای هر دو پاسخ با ابعاد کم و با ابعاد بالا وجود دارد (۱۶).

تشکیل درختان تصمیم درانباشت تصادفی با درخت کلاسیک تصمیم دارای تفاوت‌هایی است. در انباشت تصادفی هر درخت با یک نمونه خودگردان از داده‌های اصلی رشد می‌کند و به منظور انجام بهترین تقسیم فضا، تعداد m متغیر که به تصادف از بین متغیرها انتخاب شده‌اند، مورد جستجو قرار می‌گیرند. تعداد درختان و مقدار m که آن‌ها را به‌ترتیب ntree و mtry نشان می‌دهیم، می‌بایست توسط کاربر تعیین و بهینه شوند. هرچه تعداد درختان انباشت تصادفی بیشتر باشد پیش‌بینی از دقت بالاتری برخوردار است، بنابراین پارامتر ntree باید به قدر کافی بزرگ انتخاب شود اما گاهی هم اگر از یک تعداد بیشتر استفاده شود معادل این است که از همان داده‌های اصلی استفاده شده است و کارایی لازم را ندارد. درمورد پارامتر mtry معمولاً مقدار $\sqrt{p}$ پیشنهاد می‌شود که p تعداد متغیرها است. یک انباشت تصادفی آنقدر بزرگ است که تفسیر آن کار بسیار دشواری است، لذا نیازمند خلاصه کردن اطلاعات آن با استفاده از شاخص‌های کمی هستیم. یکی از این شاخص‌ها، شاخص اهمیت متغیر (Variable importance)[۱۲] می‌باشد که این شاخص برای رتبه‌بندی متغیرها برحسب اهمیت آن‌ها در اثرگذاری روی پاسخ است. معروف‌ترین شاخص‌های اهمیت متغیر، شاخص اهمیت جینی و شاخص اهمیت جایگشتی می‌باشد.

درطی ساخت درخت های مدل انباشت تصادفی برای تعیین اینکه گره براساس کدام متغیر افراز شود، از شاخص ناخالصی جینی استفاده می‌شود. اهمیت متغیر X_i در یک درخت، مجموع کاهش در شاخص ناخالصی جینی روی تمام گره‌هایی است که براساس X_i افراز شده‌اند. میانگین اندازه اهمیت متغیر X_i روی تمام درخت‌های انباشت، اندازه شاخص اهمیت جینی است.

برای محاسبه شاخص اهمیت جایگشتی نیز مقادیر X_i مشاهدات نمونه خارج کیسه به‌طور تصادفی جابجا می‌شوند و اندازه ناخالصی درخت روی مقادیر جابجا شده محاسبه می‌شود. اندازه اهمیت متغیر X_i در هر درخت، اختلاف بین این دو اندازه ناخالصی است و میانگین این مقادیر شاخص اهمیت جایگشتی است. هدف این روش این است که اگر X_i متغیر مهمی باشد جابجا شدن مقادیر آن به‌طور تصادفی منجر به افزایش ناخالصی درخت می‌شود در حالی‌که اگر متغیر تأثیرگذاری نباشد، تغییری در ناخالصی ایجاد نمی‌شود (۱۵ و ۱۶).

شاخص‌های ارزیابی مدل‌های طبقه‌بندی

برای مقایسه مدل‌های طبقه‌بندی از شاخص‌های: آماركاپا[۱۳]، آمار گاما[۱۴]، اندازه دی سامرز[۱۵] و دقت[۱۶] در مجموعه داده آزمون استفاده شد.

[12] Variable importance
[13] Kappa statistics
[14] Gamma statistic
[15] Somers'd
[16] accuracy

بسته‌های نرم‌افزاری مورد استفاده

از بسته‌های نرم‌افزاری rpartOrdinal، rpartScore، ordinalForest، glmnetcr، glmpathcr، در نرم‌افزار R ویرایش ۳.۶.۳ استفاده شد.

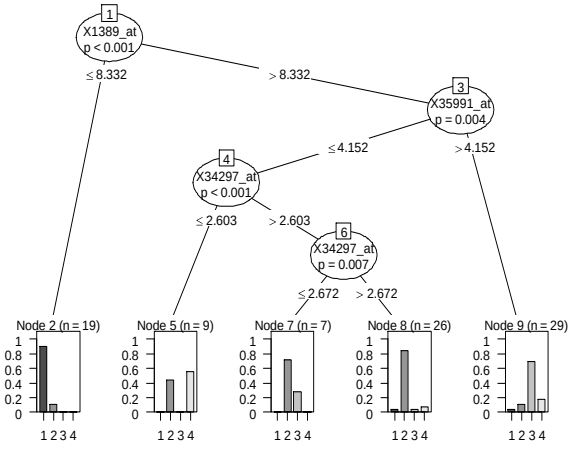
یافته‌ها

نتایج حاصل از بکارگیری مدل‌های مختلف برای مجموعه داده‌های B-cell، HCC و Heart در جدول ۱ ارائه شده است.

جدول ۱) معیارهای گاما، دی سامرز، کاپا و دقت برای مقایسه عملکرد پیش‌بینی مدل‌ها در مجموعه دیتای مختلف (بر اساس مجموعه آزمون)

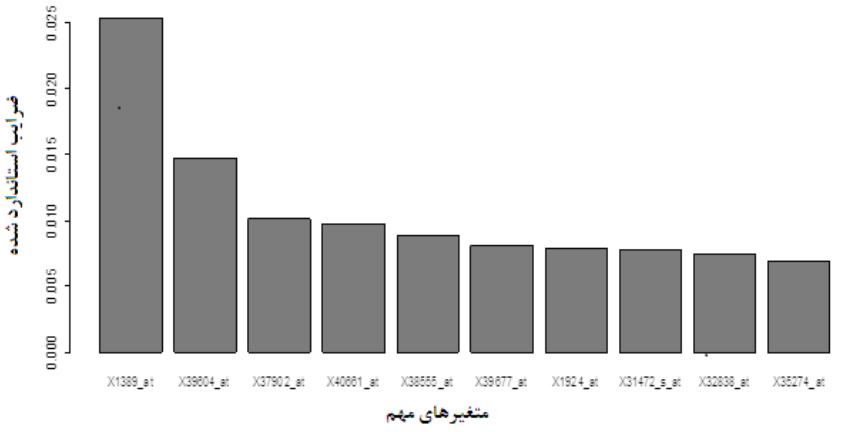
مجموعه دیتا (Data set)	روش‌های طبقه‌بندی	گاما (Gamma)	دی سامرز (Somers'd)	کاپا (Kappa)	دقت (Accuracy)
B cell	انباشت ترتیبی	۰/۸۰	۰/۵۹	۰/۴۸	۰/۶۶
	روش درخت تصمیم	۰/۵۸	۰/۴۱	۰/۴۴	۰/۴۶
	رگرسیون نسبت پیوسته جریمه شده	۰/۷۹	۰/۵۹	۰/۳۸	۰/۵۶
HCC	انباشت ترتیبی	۰/۸۶	۰/۷۴	۰/۷۶	۰/۸۹
	روش درخت تصمیم	۰/۸۵	۰/۷۰	۰/۶۸	۰/۷۸
	رگرسیون نسبت پیوسته جریمه شده	۰/۷۷	۰/۶۳	۰/۶۰	۰/۷۳
Heart	انباشت ترتیبی	۰/۶۵	۰/۳۸	۰/۲۴	۰/۶۶
	روش درخت تصمیم	۰/۵۹	۰/۳۴	۰/۱۹	۰/۶۰
	رگرسیون نسبت پیوسته جریمه شده	۰/۷۹	۰/۴۲	۰/۲۵	۰/۶۶

نتایج نشان دهنده آن است که در دو مجموعه داده با ابعاد بالا (B-cell و HCC) مدل انباشت ترتیبی از توانایی پیش‌بینی بالاتری در مجموعه آزمون برخوردار است. در حالی که برای مجموعه داده با ابعاد پایین (Heart) مدل رگرسیون نسبت پیوسته جریمه شده عملکرد پیش‌بینی بهتری در مجموعه آزمون داشته است. در نمودارهای ۱، ۳ و ۵ نمودار مربوط به مدل درخت تصمیم برای هریک از داده‌ها ارائه شده است.



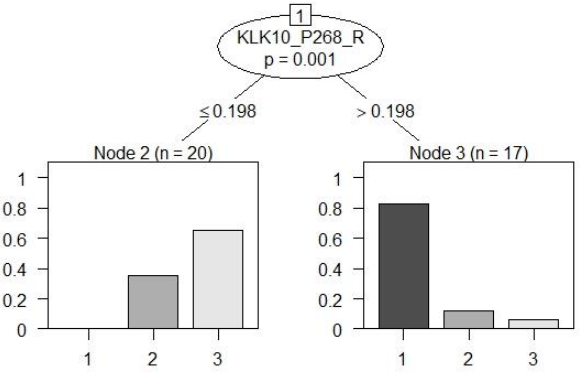
نمودار ۱) نمودار درختی مرتبط با روش درخت تصمیم در مجموعه دیتای B cell

Fig 1) Tree diagram related to decision tree method for B cell data set



نمودار ۲) نمودار میله‌ای متغیرهای مهم برای مدل انباشت ترتیبی در مجموعه دیتای B cell

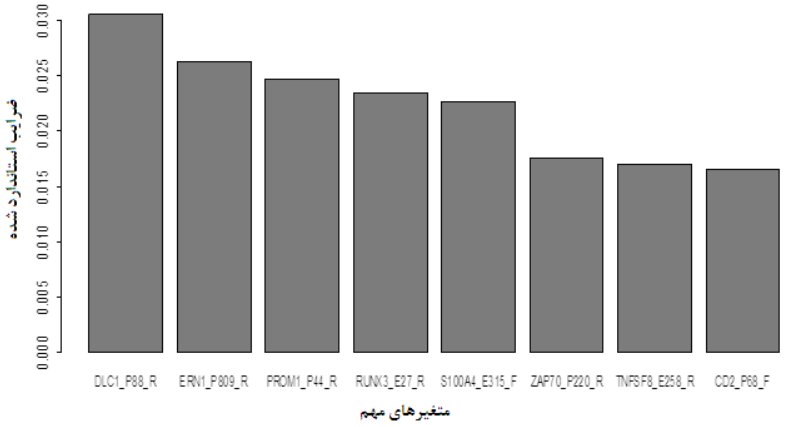
Fig 2) Bar chart of important variables for decision tree method for B cell data set



نمودار ۳) نمودار درختی روش درخت تصمیم در مجموعه دیتایHCC

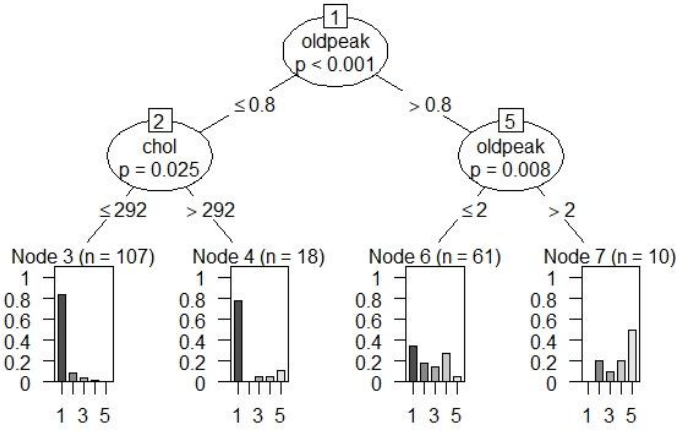
Fig 3) Tree diagram related to decision tree method for HCC data set

همچنین در نمودارهای ۲، ۴ و ۶ نمودار مربوط به شناسایی متغیرهای مهم بر اساس روش انباشت تصادفی ارائه شده است.



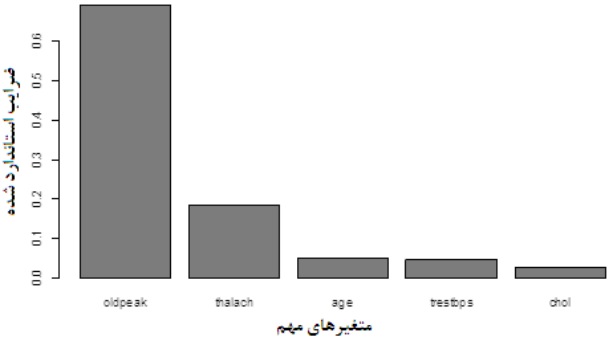
نمودار ۴) نمودار میله‌ای متغیرهای مهم برای مدل انباشت ترتیبی در مجموعه دیتای HCC

Fig 4) Bar chart of important variables for decision tree method for HCC data set



نمودار ۵) نمودار درختی روش درخت تصمیم در مجموعه دیتای Heart

Fig 5) Tree diagram related to decision tree method for Heart data set



نمودار ۶) نمودار میله‌ای متغیرهای مهم برای مدل انباشت ترتیبی در مجموعه دیتای Heart

Fig 6) Bar chart of important variables for decision tree method for Heart data set

جداول ۲، ۳ و ۴ نیز نشان‌دهنده متغیرهایی است که با ضرایب غیر صفر در مدل رگرسیون نسبت پیوسته جریمه شده شناسایی شده‌اند.

جدول ۲) ضرایب غیر صفر مدل انتخابی در روش رگرسیون نسبت پیوسته جریمه شده در مجموعه دیتای B cell

متغیرها (ژن)	ضرایب
X1044_s_at	۰/۱۶۳
X1389_at	۰/۵۴۲
X1990_g_at	۰/۱۱
X32838_at	۰/۰۲۴
X36478_at	۰/۰۱
X37265_at	۰/۰۱۴
X37981_at	۰/۰۶۳
X38266_at	۰/۵۶۸
X38555_at	-۰/۲۴۸
X39650_s_at	۰/۸۰۹
X39827_at	-۰/۰۳۶
X41837_at	۰/۱۹۳
X946_at	-۰/۰۱۳

جدول ۳) ضرایب غیر صفر مدل انتخابی در روش رگرسیون نسبت پیوسته جریمه شده در مجموعه دیتای HCC

متغیرها (ژن)	ضرایب
HLA.DPA1_P205_R	۰/۴۳۶
KLK10_P268_R	۰/۲۱۵
MPO_P883_R	۰/۸۹۸

جدول ۴) ضرایب غیر صفر برای مدل انتخابی در روش رگرسیون نسبت پیوسته جریمه شده در مجموعه دیتای Heart

متغیرها (ژن)	ضرایب
Intercept	-۳/۷۴۷۲
age - age in years	۰/۰۰۰۴
trestbps - resting blood pressure (in mm Hg on admission to the hospital)	۰/۰۰۰۴
chol - serum cholestoral in mg/dl	۰/۰۰۰۵
thalach - maximum heart rate achieved	-۰/۰۱۱۶
oldpeak - ST depression induced by exercise relative to rest	۰/۹۷۶۶

بحث

مطالعه حاضر با هدف مقایسه عملکرد سه روش رگرسیون نسبت پیوسته جریمه شده، درخت تصمیم، انباشت ترتیبی در پیش‌بینی پاسخ‌هایی با سطوح ترتیبی صورت گرفت. در این راستا عملکرد روش‌های مذکور بر روی دو مجموعه داده با ابعاد بالا (دیتای B cell و HCC) و یک مجموعه داده با ابعاد کم (Heart) در قالب شاخص‌های (گاما، دی سامرز، کاپا و دقت) مورد

مقایسه قرار گرفت. نتایج حاصل از مقایسه روش‌های به‌کار گرفته شده با استفاده از مجموعه آزمون و آموزش یکسان نشان داد که در هر دو مجموعه داده با ابعاد بالا روش انباشت ترتیبی از عملکرد بهتری برخوردار بوده است. همچنین برای مجموعه داده با ابعاد کم نیز روش رگرسیون نسبت پیوسته جریمه شده از عملکرد پیش‌بینی بهتری داشت. لازم به ذکر است داده آزمایشی در نظر گرفته شده در مجموعه داده سوم بسیار نامتعادل است و شاید بتوان دقت کمتر الگوریتم انباشت ترتیبی و درخت در مقایسه با رگرسیون نسبت پیوسته جریمه شده را به به دلیل عدم توانایی آن‌ها در مدیریت داده‌های نامتعادل نسبت داد. در این زمینه استفاده از تکنیک‌های oversampling و under sampling در فرایند تقسیم بوت استرپ داده‌ها در تشکیل درخت، در برخی از منابع توصیه شده است (۱۷). همچنین بایستی مشکل بیش برازش ایجاد شده در این مدل‌ها را نیز مد نظر قرار داد، که البته این مسأله را با افزایش تعداد درخت‌ها و تعداد نمونه‌های آموزشی می‌توان مرتفع نمود.

در مجموعه داده‌های ژنومی با ابعاد بالا، رویکرد رایج در تجزیه و تحلیل داده‌های پاسخ ترتیبی این است که مسئله را به یک یا چند مسئله با پاسخ دوحالتی تقسیم کنیم. رویکرد پاسخ دوحالتی از همه اطلاعات موجود استفاده نمی‌کند و بنابراین منجر به کاهش توان و افزایش تعداد خطاهای نوع اول می شود (۱۸). علاوه بر این، لحاظ نمودن پاسخ ترتیبی به عنوان یک پاسخ پیوسته و متعاقب آن استفاده از تکنیک‌های مدل‌سازی خطی نیز به دلیل آنکه پاسخ این مدل‌ها در قالب مقادیر برآورد شده کمی ارائه می‌شوند نیز رویکرد مناسبی نمی‌باشد (۱۹). بنابراین در ارتباط با پاسخ‌های ترتیبی استفاده از رویکردهایی از قبیل موارد مورد استفاده در این مطالعه توصیه می‌شود.

به عنوان مثال در مطالعه چن (Chen) و همکاران از مدل لجستیک تجمعی با کاهش بعد و ماشین‌بردار پشتیبان برای مدل‌سازی مرحله‌های ترتیبی سرطان در داده‌های بیان ژنی استفاده کردند و نتایج نشان داد که مدل طبقه‌بندی ترتیبی با روش اعتبار سنجی دقیق‌تر از طبقه‌بندی‌های مرسوم غیر ترتیبی است (۱).

مطالعات محدودی بر اساس روش‌های درخت تصمیم، انباشت ترتیبی و رگرسیون نسبت پیوسته جریمه شده بر روی داده‌های با ابعاد بالا صورت گرفته است در هر حالت تنها یکی از روش‌های مذکور بر روی داده‌های مورد نظر اعمال گردیده و حالت مقایسه عملکرد روش‌های مذکور مدنظر نبوده است. در حالی‌که در مطالعه حاضر عملکرد این سه روش در طبقه‌بندی مورد مقایسه قرار گرفت.

به عنوان مثال در مطالعه‌ای که توسط آرچر (Archer)، بر روی مجموعه دیتای با پاسخ ترتیبی B cell با به‌کارگیری روش درخت تصمیم و توابع تقسیم مختلف، صورت گرفت. تابع تقسیم ناخالصی ترتیبی با داشتن بیشترین مقدار five-fold gamma (۰/۷۶) در بین توابع تقسیم‌بندی دوتایی و تقسیم جینی تعمیم یافته بهترین پیش‌بینی را داشت. در مطالعه حاضر نیز علیرغم کاهش بعد تعداد متغیرهای بیان ژن به ۳۸۴۱ متغیر در ۹۰ بیمار مبتلا به لوسمی سلول B نیز تابع ناخالصی ترتیبی با استفاده از شاخص‌های ارزیابی بهترین پیش‌بینی را در بین توابع تقسیم درخت تصمیم داشت (۱۰).

همچنین در مطالعه جانیتزا (Janitza) و همکاران، روش انباشت تصادفی برای داده‌های با پاسخ ترتیبی برای پیش‌بینی و انتخاب متغیر استفاده کردند که باتوجه به دقت پیش‌بینی در این مطالعه، عملکرد درخت‌های رگرسیونی ترتیبی مشابه و در بیشتر مواقع حتی کمی بهتر از درخت طبقه‌بندی است (۲۰).

در مطالعه دیگر که توسط هورانگ (Hornung) صورت گرفت، روش انباشت ترتیبی برای پیش‌بینی پاسخ ترتیبی در داده‌های با ابعاد پایین و بالا ارائه شد، دراین روش متغیرهای کووریت براساس اهمیتشان در پیش‌بینی پاسخ، رتبه‌بندی شدند و پنج مجموعه دیتا در این مطالعه استفاده شده است. نتایج با استفاده از معیارهای ارزیابی شاخص کاپا و کاپا موزون نشان داده شد که روش مذکور در مقایسه با روش‌های انباشت تصادفی چندکلاسه عملکرد بهتری داشته است (۱۶).

نتیجه‌گیری

نتایج مطالعه حاضر نشان دهنده آن است که انتخاب بهترین مدل پیش‌بینی از بین مدل‌های به‌کار رفته بستگی به مجموعه داده مورد استفاده دارد و برای هر مجموعه داده بایستی روش‌های مختلف را مورد بررسی قرار داد تا به بهترین مدل پیش‌بینی دست یافت. همچنین در پیش‌بینی پاسخ‌های با ماهیت ترتیبی استفاده از رویکردهایی که ماهیت ترتیبی پاسخ را لحاظ کنند، توصیه می‌شود.

سپاس و قدردانی

این مقاله برگرفته از پایان نامه کارشناسی ارشد در رشته آمار زیستی بود، که نویسندگان مراتب تشکر و قدردانی خود از حمایت مالی معاونت پژوهشی دانشگاه علوم پزشکی همدان اعلام می‌دارند (کد طرح: ۹۹۰۷۲۹۵۳۵۶).

تضاد منافع

هیچ‌گونه تعارض منافع توسط نویسندگان مقاله بیان نشده است.

References:

1.Chen CK. The Classification Of Cancer Stage Microarray Data. Comput Meth Prog Bio 2012; 108(3): 1070-7.
2.Archer KJ, Hou J, Zhou Q, et al. Ordinalgmifs: An R Package For Ordinal Regression In High-Dimensional Data Settings. Cancer Inform 2014; 13: CIN.S20806.
3.Farhadi Z, Shahsavani D. Gene Expression Data Clustering With Random Forest Dissimilarity. Razi J Med Sci 2015; 22(136): 109-18. (Persian)
4.Safe M, Faradmal J, Mahjub H. A Comparison Between Cure Model And Recursive Partitioning: A Retrospective Cohort Study Of Iranian Female With Breast Cancer. Comput Math Methods Med 2016; 2016: 9425629.
5.Archer KJ, Williams AA. L1 Penalized Continuation Ratio Models For Ordinal Response Prediction Using High-Dimensional Datasets. Stat Med 2012; 31(14): 1464-74.
6.Tibshirani R. Regression Shrinkage And Selection Via The Lasso. J Royal Stat Soc Series B (Methodological) 1996; 58(1): 267-88.
7.Buntine W, Niblett T. A Further Comparison Of Splitting Rules For Decision-Tree Induction. Mach Learn 1992; 8: 75-85.
8.Zhang H, Singer B. Recursive Partitioning And Applications. New York: Springer Science & Business Media, 2010, 79-95.
9.Breiman L, Friedman J, Stone CJ, et al. Classification And Regression Trees. 1st ed. Chapman And Hall/CRC, 1984, 18-41.
10.Archer KJ. Rpartordinal: An R Package For Deriving A Classification Tree For Predicting An Ordinal Response J Stat Softw 2010; 34: 7.
11.Galimberti G, Soffritti G, Di Maso M. Classification Trees For Ordinal Responses In R: The Rpartscore Package. J Stat Softw 2012; 47(10): 1-25.
12.Cappelli C, Mola F, Siciliano R. A Statistical Approach To Growing A Reliable Honest Tree. Comput Stat Data Anal 2002; 38(3): 285-99.

13.Mingers J. Expert Systems—Rule Induction With Statistical Data. J Oper Res Soc 1987; 38(1): 39-47.
14.Niblett T, Bratko I. Learning Decision Rules In Noisy Domains. Proceedings Of Expert Systems' 86, The 6Th Annual Technical Conference On Research And Development In Expert Systems III. Brighton, United Kingdom: Cambridge University Press, 1987.
15.Genuer R, Poggi JM, Tuleau C. Random Forests: Some Methodological Insights. arXiv Preprint arXiv:0811.3619. 2008.
16.Hornung R. Ordinal Forests. J Classif 2020; 37: 4-17.
17.Drummond C, Holte RC. C4.5, Class Imbalance, And Cost Sensitivity: Why Under-Sampling Beats Over-Sampling. In Workshop On Learning From Imbalanced Datasets II. Washington DC: Citeseer, 2003; 11: 1-8.
18.Breiman L, Friedman J, Olshen R, et al. Classification And Regression Trees. Wadsworth Int Group 1984; 37(15): 237-51.
19.Gentry AE, Jackson-Cook CK, Lyon DE, et al. Penalized Ordinal Regression Methods For Predicting Stage Of Cancer In High-Dimensional Covariate Spaces. Cancer Inform 2015; 14(s2): CIN.S17277.
20.Janitza S, Tutz G, Boulesteix AL. Random Forest For Ordinal Responses: Prediction And Variable Selection. Comput Stat Data Anal 2016; 96(C): 57-73.

Original Article

Comparison of Ordinal Response Modeling Methods like Decision Trees, Ordinal Forest and L_1 Penalized Continuation Ratio Regression in High Dimensional Data

***Z. Torkashvand** (MSc)[1*], **H. Mahjub** (PhD)[1,2], **AR. Soltanian** (PhD)[1,3], **M. Farhadian** (PhD)[1,2**]*

[1] Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
[2] Research Center for Health Sciences, Hamadan University of Medical Sciences, Hamadan, Iran
[3] Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran

(Received 5 Jul, 2021 Accepted 5 Sep, 2021)

Abstract

Background: Response variables in most medical and health-related research have an ordinal nature. Conventional modeling methods assume predictor variables to be independent, and consider a large number of samples (n) compared to the number of covariates (p). Therefore, it is not possible to use conventional models for high dimensional genetic data in which p > n. The present study compared the predictive performance of decision trees, ordinal forest, and L1 penalized continuation ratio regression.

Materials and Methods: In the present study, three data sets were used. The B-cell data contained 12,625 gene expression data related to 128 patients with four ordinal levels of response variables. The HCC data related to liver cancer included 1469 genes of 56 patients with three ordinal levels of response variables. The Heart data contained information of five variables in 294 patients undergoing angiography with five ordinal levels of response variables. The performance of the methods was compared based on the same training and test datasets using indicators such as accuracy, gamma, and kappa.

Results: For two high-dimensional data sets, the ordinal forest model had a higher predictive ability while for the low-dimensional data set, the L1 penalized continuation ratio model had a better predictive performance.

Conclusion: The selection of the best prediction model depends on the data set, and for each data, different methods should be considered to achieve the best model.

Keywords: Ordinal response, Ordinal Forest, L_1 Penalized Continuation Ratio Regression, High dimensional data

Cite this article as: Torkashvand Z, Mahjub H, Soltanian AR, Farhadian M. Comparison of Ordinal Response Modeling Methods like Decision Trees, Ordinal Forest and L1 Penalized Continuation Ratio Regression in High Dimensional Data. Iran South Med J 2021; 24(5): 454-468

[**]*Address for correspondence:* Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran. Email: maryam_farhadian80@yahoo.com
*ORCID: 0000-0003-4321-3842
**ORCID: 0000-0002-6054-9850