تحلیل بقا با استفاده از رگرسیون درختی جمعی بیزی

فاطمه حسینی،[۱] امید کریمی،[۲] فاطمه حامدی،[۳]

تاریخ دریافت: ۱۳۹۷/۱۰/۱۸
تاریخ پذیرش: ۱۳۹۸/۶/۳۰

چکیده:
مدل‌های درختی یک روش جدید و ابتکاری را برای تحلیل مجموعه‌داده‌های بزرگ به‌وسیله تقسیم‌بندی فضای پیش‌بینی کننده‌ها به نواحی ساده‌تر به نمایش می‌گذارند. مدل رگرسیونی درختی جمعی بیزی، مدلی که در این مقاله به معرفی و توضیح آن می‌پردازیم، در ساختار خود از مدل جمع درختان استفاده می‌کند، زیرا ترکیب چند درخت از درخت تنها دقت بالاتری دارد. پس این مدل مبتنی بر درخت و جزء مدل‌های ناپارامتری است و در واقع تعمیمی از روش‌های رده‌بندی و رگرسیون درختی است، که در ساختار این روش‌ها درخت تصمیم وجود دارد. این روش‌ها تحلیلی قدرتمند برای کشف ساختار داده‌ها هستند و کاربرد آنها در علوم پزشکی بسیار وسیع است. در این روش، روی پارامترهای مدل جمع درختان پیشین‌هایی در نظر گرفته می‌شود و سپس با استفاده از الگوریتم‌های کمکی به تحلیل می‌پردازد. در این مقاله ابتدا مختصراً مدل رگرسیونی درختی جمعی بیزی را معرفی کرده و سپس کاربرد آن را در تحلیل بقا با بررسی داده‌های مربوط به بیماران سرطان ریه بیان می‌کنیم.

واژه‌های کلیدی: مدل‌های درختی، مدل جمعی بیزی، درخت تصمیم.

۱ مقدمه

مدل‌های درختی[۴] یک روش جدید و ابتکاری را برای تحلیل مجموعه‌داده‌های بزرگ به‌وسیله تقسیم‌بندی فضای پیش‌بینی کننده‌ها به نواحی ساده‌تر به نمایش می‌گذارند، [۱۰]. روش‌های مبتنی بر مدل‌های درختی که جزء خانواده رگرسیون ناپارامتری است، یکی از روش‌های انعطاف‌پذیر و شهودی و وسیله‌ای قدرتمند در تحلیل داده‌ها، برای کشف ساختار پیچیده داده‌ها است. روش‌های مبتنی بر مدل‌های درختی فضای متغیرهای کمکی را به‌طور بازگشتی به ناحیه‌های مجزا افراز می‌کند. برای این‌که هر گره افراز شود، تمام افرازهای ممکن برای هر متغیر کمکی ارزیابی می‌شود. متغیر و نقطه افراز متناظر با آن به‌گونه‌ای انتخاب می‌شود، که بهترین تفکیک بین دو گره حاصل شود. این روند به‌طور بازگشتی ادامه می‌یابد تا این‌که هر گره شامل تعداد محدودی از حالت‌ها باشد، برای مطالعه بیشتر به [۵، ۱۱] مراجعه شود.

روش رگرسیون درختی جمعی بیزی[۵] (BART) یک مدل مبتنی بر درخت و جزء مدل‌های ناپارامتری است که از روش‌های گردآوری به‌صورت کلّی و الگوریتم‌های کمکی به‌طور ویژه استفاده می‌کند و در واقع تعمیمی از روش بیزی رگرسیون درختی و رده‌بندی[۶](Bayesian CART) است، [۷]. در واقع این مدل‌ها تکنیک‌هایی هستند که زیرگروه‌های همگن را با استفاده از روش‌های ناپارامتری استخراج می‌کنند، [۴]. این مدل‌ها در

[۱] هیأت علمی گروه آمار، دانشگاه سمنان، سمنان، ایران
[۲] هیأت علمی گروه آمار، دانشگاه سمنان، سمنان، ایران
[۳] دانشجوی کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران

[۴] Tree moldels
[۵] Bayesian additive regression trees
[۶] Baysian classification and regression tree

دو دهه اخیر عمومیت یافته‌اند بر حسب متغیر پاسخ به دو دسته تقسیم می‌شوند:

۱. مدل‌های رده‌بندی درختی (اگر متغیر پاسخ رده‌بندی شده باشد)

۲. مدل‌های رگرسیونی درختی (اگر متغیر پاسخ پیوسته باشد)

روش‌های درختی در مقایسه با روش‌های خطی زمانی بهتر عمل می‌کنند که روابط بین پیش‌بینی کننده‌ها و پاسخ غیرخطی و پیچیده است. مدل BART نخستین بار توسط چیپمن و دیگران [۳] در سال ۲۰۰۷ معرفی شد. BART در ساختار خود از مدل جمع درختان استفاده می‌کند زیرا ترکیب چند درخت از درخت تنها دقت بالاتری دارد. در این روش، روی پارامترهای مدل جمع درختان پیشین‌هایی در نظر گرفته می‌شود و سپس با استفاده از الگوریتم پس برازش بیزی مونته‌کارلوی زنجیر مارکوفی (BBMCMC)[۷] به تحلیل می‌پردازد. در واقع ساختن مدل BART شامل دو مرحله است؛ مدل جمع درختان و تنظیمات پیشین[۸]. تنظیمات پیشین که برای پارامترهای مدل در نظر گرفته می‌شود، باعث می‌شود که اثر هر تک‌درخت مؤثر حفظ شود، به طوری که بدون این تنظیمات تعداد زیادی پارامتر در مدل وجود خواهد داشت که منجر به اعمال محدودیت‌های اضافی در محاسبات می‌شود. لذا می‌توان گفت روش BART یک روش کارامد برای کسب دانش از مجموعه‌داده‌ها تلقی کرد، به‌طوری‌که با استفاده از این روش می‌توان به زوایای مختلف مجموعه‌داده‌ها از جمله برآورد پارامترهای مربوط، یافتن مهمترین متغیرها، استنباط و پیش‌بینی آگاه گشت. کاربرد BART در تحلیل بقا توسط اسپارپانی و دیگران [۹] مطرح شد و با استفاده از تابع surv.bart در نرم‌افزار برای این مدل ارائه شد. در این مقاله هدف بررسی بیشتر کاربرد BART در مدل‌های بقا است.

ساختار مقاله به این صورت است که در بخش اول مقدمه و سپس رد بخش دوم مدل رگرسیونی درختی جمعی بیزی معرفی شده است. در بخش سوم تحلیل بقا با استفاده از مدل BART بیان و روی مجموعه‌ای از داده‌ها پیاده‌سازی شده است.

۲ مدل رگرسیونی درختی جمعی بیزی

اوّلین بار مدل رگرسیونی درختی جمعی بیزی (BART) در کنفرانسی با موضوع پیشروی وپیشرفت‌های تداخل در سیستم‌های پردازش اطلاعات عصبی در سال۲۰۰۶ بیان شد. BART یک روش رگرسیون ناپارامتری بیزی است که مبتنی بر درخت بوده و از روش‌های گردآوری به‌صورت کلّی و الگوریتم‌های کمکی به‌طور ویژه استفاده می‌کند و با یک مدل آماری تعریف می‌شود. در این روش بیزی از جمع درخت‌ها برای مدل کردن یا تقریب زدن $\mathbf{f(x)} = \mathbf{E(Y|x)}$ استفاده می‌شود. همان‌طور که اشاره شد ساختن مدل BART شامل دو مرحله است، مدل جمع درختان و تنظیمات پیشین که هر یک را در بخش‌های زیر توضیح می‌دهیم.

۱.۲ مدل جمع درختان

برای توضیح مدل جمع درخت‌ها، از مدل درخت تنها شروع می‌کنیم. T یک درخت دو حالتی است که شامل مجموعه‌ای از قوانین تصمیم‌گیری، گره‌های داخلی و مجموعهٔ گره‌های پایانی است. M برابر $\{\mu_۱, \mu_۲, \cdots, \mu_b\}$ نشان‌دهندهٔ مجموعه‌ای از پارامترهای مربوط به هر گرهٔ پایانی b از درخت T است. قوانین تصمیم‌گیری درخت دودویی بر اساس افراز فضای پیش‌بینی کننده‌ها به شکل $\{x \in A\}$ در مقابل $\{x \notin A\}$ است، که A زیرمجموعهٔ محدودهٔ x است. هر مقدارx مربوط به یک گره پایانی ازT است که بر اساس قوانین تصمیم‌گیری از بالا به پایین حاصل شده و سپس μ_i مربوط با آن گره پایانی مشخص می‌شود. برای هر T و M، از $g(x;T,M)$ برای نشان دادن تابعی که هر $\mu_i \in M$ به x اختصاص می‌دهد، استفاده می‌کنیم، بنا بر این

$$Y = g(x;T,M) + \varepsilon \qquad \varepsilon \sim N(۰, \sigma^۲) \qquad (۱)$$

[۷] Bayesian backfitting MCMC

[۸] A regularization prior

یک مدل درخت تنها است. تحت رابطهٔ (۱)، $E(Y|x)$ معادل پارامتر گرهٔ پایانی μ_i مشخص شده توسط $g(x;T,M)$ است. با این نمادگذاری، جمع درخت‌ها به‌صورت زیر قابل بیان است:

$$Y=\sum_{j=1}^{m} g(x;T_j,M_j)+\varepsilon \qquad \varepsilon\sim N(0,\sigma^2) \quad (۲)$$

که برای هر درخت رگرسیون دودویی T_j و گره‌های پایانی آن M_j، $g(x;T_j,M_j)$ تابعی است که $\mu_{ij}\in M_j$ را به x اختصاص می‌دهد. از رابطهٔ ((۲))، $E(Y|x)$ معادل جمع تمامی گره‌های پایانی μ_{ij} اختصاص داده شده به x توسط $g(x;T_j,M_j)$ است. علاوه براین، μ_{ij} وقتی اثر اصلی را نشان می‌دهد که $g(x;T_j,M_j)$ تنها به عنصر x وابسته باشد و وقتی اثر متقابل را نشان می‌دهد که $g(x;T_j,M_j)$ به یک یا چند عنصر x وابسته باشد. بنا بر این مدل جمع درخت‌ها می‌تواند با اثر اصلی و متقابل کار کند. در مدل جمع درختان، افزایش تعداد درختان منجر به افزایش انعطاف‌پذیری مدل می‌شود زیرا با افزایش تعداد پارامترهای مدل، انعطاف‌پذیری مدل افزایش می‌یابد.

۲.۲ تنظیمات پیشین

[۱، ۲] با در نظر گرفتن تنظیمات پیشین برای پارامترهای مدل (۲) باید اثر هر تک‌درخت مؤثر حفظ شود، علاوه بر این بدون این تنظیمات تعداد زیادی پارامتر در مدل فوق وجود خواهد داشت که منجر به اعمال محدودیت‌های اضافی در محاسبات می‌شود. با فرض استقلال پیشین‌ها تنها کافی است پیشین برای $p(T_j)$، $p(\mu_{ij}|T_j)$ و $p(\sigma)$ مشخص شود.

۱.۲.۲ پیشین T_j

پیشین برای $p(T_j)$، در سه بخش زیر بیان می‌شود:

(i) احتمال این‌که گره در عمق $d(=0,1,2,\cdots)$ بی‌پایان باشد (گره خارجی نباشد) به‌صورت زیر ست:

$$\alpha(1+d)^{-\beta}, \quad \alpha\in(0,1)\ ,\ \beta\in[0,\infty).$$

(ii) تخصیص دادن توزیع به متغیرهای جداکننده در هر گره داخلی.

(iii) تخصیص دادن توزیع به قانون جداکننده در هر گره داخلی.

که برای (ii) از پیشین یکنواخت و برای (iii) از پیشین یکنواخت گسسته برای مقادیر جداکننده موجود استفاده می‌شود.

۲.۲.۲ پیشین $\mu_{ij}|T_j$

برای$p(\mu_{ij}|T_j)$ از توزیع نرمال مزدوج $N(\mu_\mu,\sigma_\mu^2)$ که دارای مزایای محاسباتی بسیار است، استفاده می‌شود. برای مشخص کردن پارامترهای μ_μ و σ_μ توجه کنید که $E(Y|x)$ جمع m تا μ_{ij} تحت مدل جمع درختان است و چون μ_{ij}ها دارای پیشین‌های مستقل و هم‌توزیع هستند بنا بر این پیشین $E(Y|x)$ دارای توزیع $N(m\mu_\mu,m\sigma_\mu^2)$ است و به احتمال زیاد $E(Y|x)$ بین y_{min} و y_{max} است که مقادیر کمینه و بیشینه Y در داده هستند. بعد از انتخاب مقادیر μ_μ و σ_μ و در نتیجه مشخص شدن $N(m\mu_\mu,m\sigma_\mu^2)$، احتمال قابل توجهی به بازه (y_{min},y_{max}) اختصاص داده می‌شود. با انتخاب μ_μ و σ_μ و مقادیر از پیش انتخاب شده k، $y_{min}=m\mu_\mu-k\sqrt{m}\sigma_\mu$ و $y_{max}=m\mu_\mu+k\sqrt{m}\sigma_\mu$ است. برای مثال $k=2$، با احتمال پیشین ۹۵٪ مقدار $E(Y|x)$ در این بازه (y_{min},y_{max}) قرار دارد. معمولاً برای سادگی محاسبات ابتدا با تغییر مقیاس مقادیر Y مقدار مشاهدات تبدیل یافته در محدوده $y_{min}=-0/5$ تا$y_{max}=0/5$ قرار می‌گیرد، که به‌عنوان متغیر وابسته محسوب می‌شود. سپس پیشین μ_{ij} در $\mu_\mu=0$ مرکزی و σ_μ به‌گونه‌ای انتخاب می‌شود که برای مقادیر مناسب k،

$$\mu_{ij}\sim N(0,\sigma_\mu^2), \qquad \sigma_\mu=0/5/k\sqrt{m},$$

فرض می‌شود. نکته‌ای که در مورد k باید خاطر نشان کرد، این است که چنانچه مقادیر k بین ۱و۳ باشد استراتژی مورد بحث عملکرد بهتری خواهد داشت و به‌صورت پیش‌فرض مقدار k برابر ۲ است. البته مقدار k را می‌توان با استفاده از روش اعتبارسنجی متقابل محاسبه نمود.

۳.۲.۲ پیشین σ

برای $p(\sigma)$، از پیشین مزدوج استفاده می‌شود که در این‌جا توزیع خی‌دو وارون به‌صورت $\sigma^2\sim\nu\lambda/\chi_\nu^2$ است. برای یافتن مقادیر

ν وλ از اطلاعات داده‌ها استفاده می‌کنیم و احتمال قابل توجهی به ناحیهٔ مقادیر ممکن برای σ اختصاص داده می‌شود، به‌طوری که از بیش‌تمرکز[۹] و بیش‌پراکنش[۱۰] جلوگیری شود. دو انتخاب برای $\hat{\sigma}$ وجود دارد، یک روش ساده این است که $\hat{\sigma}$ انحراف معیار استاندارد نمونه‌ای[۱۱] Y باشد، یا می‌توان $\hat{\sigma}$ را به‌صورت انحراف معیار استاندارد باقی‌مانده‌ها[۱۲] از برازش روش حد اقل مربعات رگرسیون خطی Y روی Xهای اصلی (تبدیل نیافته) در نظر گرفت. معمولاً مقدار ν را بین ۳ تا ۱۰ انتخاب می‌کنیم و مقدار λ را به‌صورت q -امین چندک پیشین $\hat{\sigma}$، درنظر می‌گیرند به‌طوری‌که $p(\sigma < \hat{\sigma}) = q$.

۴.۲.۲ انتخاب m

تفاوت عمدهٔ BART با روش‌های دیگر در انتخاب تعداد درختان یعنی m است. BART از الگوریتم بازگشتی پس برازش بیزی استفاده می‌کند و چرخهٔ تناوب آن روی m درخت است. اگر BART برای تخمین $f(x)$ یا پیش‌بینی Y استفاده شود، منطقی است که با m مانند یک پارامتر نامعلوم رفتار شود. بهترین مقدار m از طریق روش اعتبارسنجی متقابل انتخاب می‌شود. البته این رویکرد از نظر محاسباتی مقرون به‌صرفه نیست. برای جلوگیری از هزینه‌های محاسباتی، به‌صورت پیش‌فرض $m = ۲۰۰$ در نظر گرفته می‌شود، سپس برای یک یا دو انتخاب متفاوت دیگر m نیز، نتایج بررسی می‌شود تا m مناسب انتخاب شود. به‌طور شهودی ثابت می‌شود که با افزایش مقدار m از ۱، کارایی BART رو به بهبود است و سپس به آهستگی از نقطه‌ای به بعد با افزایش مقدار m کارایی نیز تحت‌شعاع قرار گرفته وکاهش می‌یابد. بنا بر این برای پیش‌بینی، به نظر می‌رسد که انتخاب درست m بسیار اهمیت دارد.

۳.۲ توزیع پسین

همان‌طور که در بخش‌های قبلی مطرح شد، هدف بررسی مدل جمع‌درختان است. مدل جمع‌درختان یک مدل جمع‌پذیر است، که جهت بررسی آن باید پارامترهای مدل مشخص باشد، از طرفی برای پارامترهای این مدل پیشین‌هایی در نظر گرفته شد. بنا بر این برای برآورد پارامترهای این مدل و محاسبهٔ توزیع‌پسین این مدل باید از یک رهیافت بیزی استفاده نمود، که در برازش مدل BART از الگوریتم پس برازش بیزی مونته‌کارلوی زنجیر مارکوفی (BBMCMC) استفاده می‌شود. برای داده‌های مشاهده‌شده y تنظیمات بیزی که برای مدل BART معرفی شد، منجر به توزیع پسینی به شکل

$$p((T_۱, M_۱), \cdots, (T_m, M_m), \sigma|y) \qquad (۳)$$

می‌شود. با توجه به بعد فضای پارامترها ممکن است این‌گونه به‌نظر برسد که محاسبه این پسین ناشدنی است، ولی الگوریتم BBMCMC راه‌حلی برای نمونه‌گیری از این پسین است که می‌توان گفت این الگوریتم درواقع یک نمونه‌گیری گیبز است. برای سهولت، منظور از نماد $T_{(j)}$ همهٔ درختان موجود در مدل جمع‌درختان به‌جز j-امین درخت است و به‌طور مشابه $M_{(j)}$ نیز، تعبیر مشابهی دارد. بنا بر این $T_{(j)}$ مجموعه‌ای از $m-۱$ درخت است که $M_{(j)}$ پارامترهای گره‌های خروجی مربوط است. نمونه‌گیری گیبز در این‌جا معادل با استخراج پی‌درپی $(T_{(j)}, M_{(j)})$ شرطی روی $(T_{(j)}, M_{(j)}, \sigma)$ است،

$$(T_j, M_j)|T_{(j)}, M_{(j)}, \sigma, y, \quad j = ۱, \cdots, m, \qquad (۴)$$

و سپس نمونه‌های σ از توزیع شرطی کامل زیر به دست می‌آید :

$$\sigma|T_۱, \cdots, T_m, M_۱, \cdots, M_m, y. \qquad (۵)$$

[۶] کاربرد نمونه‌گیری گیبز برای مدل‌های جمع‌پذیر و جمع‌پذیر تعمیم‌یافته را برای σ ثابت مورد بررسی قرار دادند. آنها نشان دادند که نمونه‌گیری گیبز برای برخی مدل‌ها تعمیمی از الگوریتم

[۹] Overconcentration

[۱۰] Overdispersion

[۱۱] Sample standard deviation

[۱۲] Residual standard deviation

پس برازش است و به همین دلیل الگوریتم پس برازش بیزی تحت عنوان پس برازش بیزی مونته‌کارلوی زنجیر مارکوفی بیان می‌شود.

۳ تحلیل بقا با مدل BART

تجزیه و تحلیل بقا با مدل BART که نشان‌دهندهٔ کاربرد آمار در زمینهٔ پزشکی است، [۹]. تجزیه و تحلیل بقا با مدل BART توسط تابع surv.bart در نرم‌افزار R ارائه شده است، که رویکردی ساده و مستقیم برای استفاده از BART در تحلیل بقا است که بسیار انعطاف‌پذیر است و می‌توان آن را معادل تحلیل بقای زمان گسسته دانست. در این بخش به بررسی داده‌های مربوط به بیماران سرطان ریه می‌پردازیم.

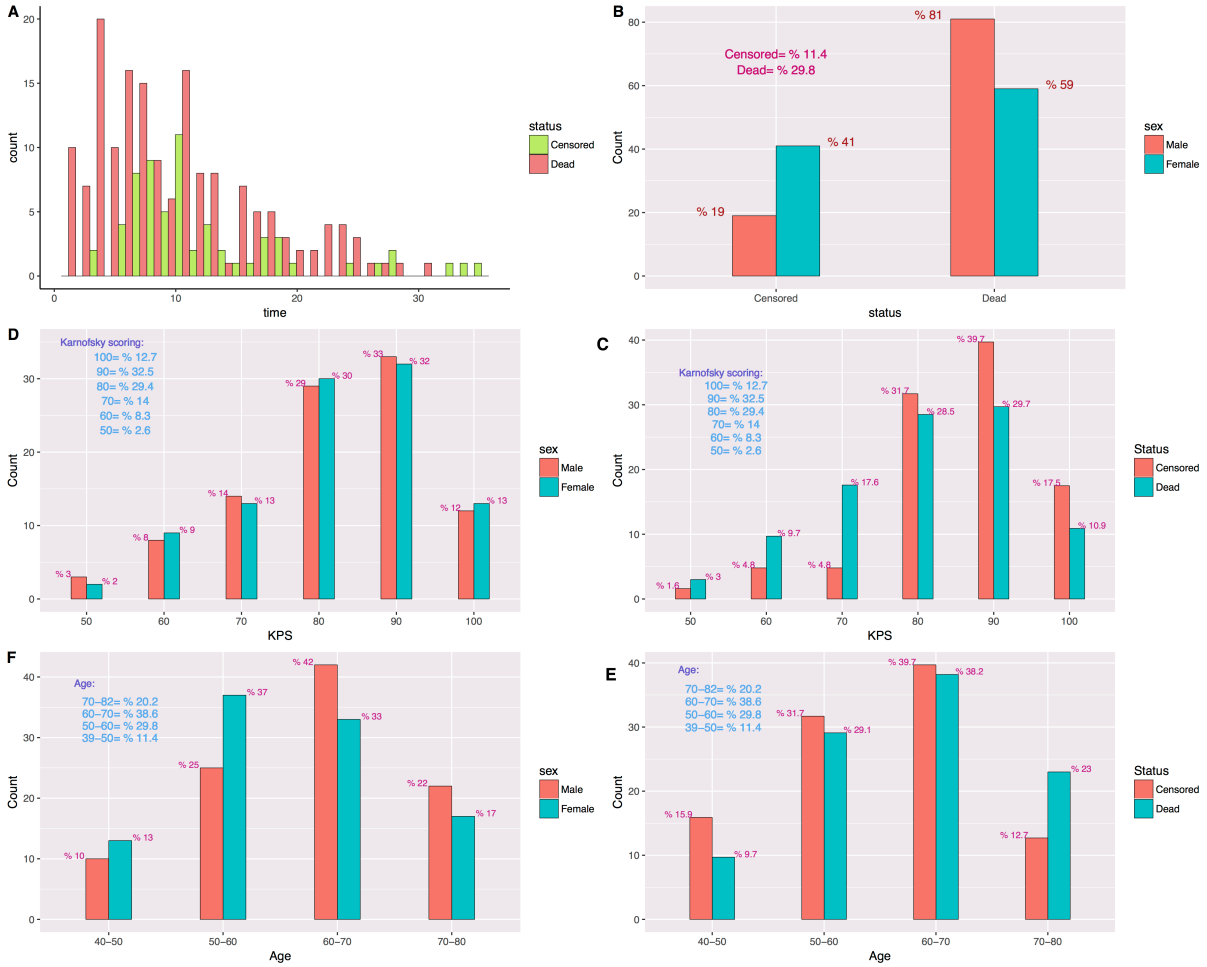
۱.۳ مجموعه‌داده‌ها

داده‌های مورد استفاده در این بخش، مجموعه‌داده lung موجود در نرم‌افزار R است که هدف پیش‌بینی زمان بقا بیماران مبتلا به سرطان ریه است این مجموعه‌داده شامل اطلاعات ۲۲۸ بیمار مبتلا به سرطان ریه است که توسط گروه درمان سرطان مرکزی شمالی جمع‌آوری شده است. در این مجموعه، داده‌ها متغیرهای زمان بقا به روز و وضعیت سانسورشدگی (مرگ=۲ و سانسورشده=۱) موجود است. علاوه بر این متغیرها، متغیرهای کمکی نیز در فایل داده وجود دارد که سه متغیر جنسیت (sex)، سن (age) و متغیر معیار ارزیابی عملکردی در بیماران سرطانی که توسط پزشک مشخص می‌شود به نام کارنوفسکی[۱۳] (KPS) به‌عنوان متغیرهای کمکی انتخاب می‌شوند. معیار (KPS) برای مقایسه کارایی درمان‌های مختلف در بیماران استفاده می‌شود که هرچه مقدار کمتری به بیمار داده شود احتمال بقا برای آن بیمار پایین‌تر است. این معیار توسط دیویدکارنوفسکی و دیگران [۸] برای ارزیابی وضعیت عملکردی بیماران دچار سرطان ریه و انجام شیمی درمانی طراحی شد. معمولاً اعداد از صفر تا ۱۰۰ به آن می‌دهند که در این مطالعه بیماران بر حسب نمره وضعیت عملکردی کارنوفسکی توسط پزشکان و به تجربه به شش گروه ۵۰، ۶۰، ۷۰، ۸۰، ۹۰ و ۱۰۰ تقسیم شده‌اند که معمولاً وضعیت عملکردی مطلوب امتیاز ۱۰۰ - ۸۰ و نامطلوب امتیاز ۷۰ - ۵۰ می‌گیرند که بر اساس فعالیت‌های معمول روزانه که بیمار قادر به انجام آنها است داده می‌شود. سرطان ریه در مراحل ابتدایی هیچ نشانه‌ای از خود بروز نمی‌دهد و اغلب بیماران زمانی به پزشک مراجعه می‌کنند که سرطان در مراحل پیشرفته است و به همین دلیل سرطان ریه کشنده‌ترین سرطان انسان محسوب می‌شود.

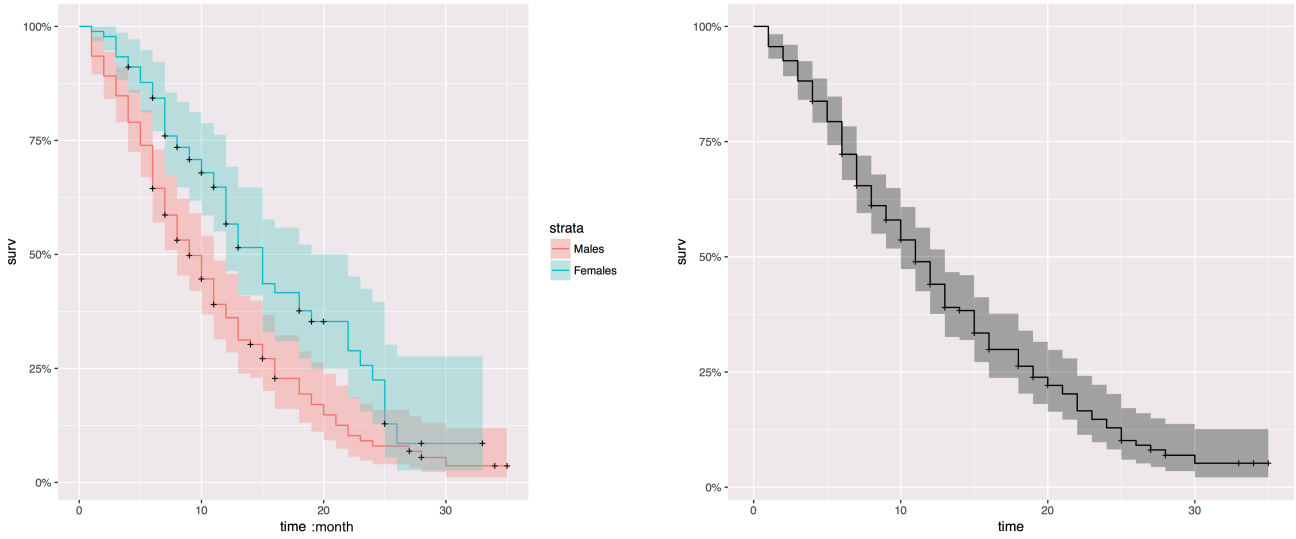
۲.۳ توصیف داده‌ها

پس از بررسی اوّلیه داده‌ها مشخص شد که سن بیماران موجود در این مجموعه‌داده بین ۳۹ تا ۸۲ سال است و میانگین سن ۶۲٫۴۵ و میانه آن ۶۳ سال است. حد اقل زمان بقای بیماران از ابتدای مطالعه ۵ روز و ماکسیمم بقا ۳۵ هفته است. ۲۸ درصد داده‌ها سانسورشده و ۷۲ درصد سانسور نشده است، به این معنا که ۲۸٪ بیماران پس از شیمی درمانی و درمان‌های داده شده بهبود پیدا کرده‌اند و یا قبل از مرگ به هر دلیل دیگری از مطالعه خارج شده‌اند و ۷۲٪ بیماران فوت کرده‌اند. ۶۰ درصد بیماران مرد و ۴۰ درصد زن هستند. درصد بیماران بر حسب KPS به ترتیب ۲٫۶٪، ۸٫۳٪، ۱۴٪، ۲۹٫۴٪، ۳۲٫۵٪ و ۱۲٫۷٪ برای KPS، به‌ترتیب ۵۰، ۶۰، ۷۰، ۸۰، ۹۰ و ۱۰۰ به دست آمد. در شکل ۱ خلاصه‌ای از نمودارهای توصیفی داده‌ها آورده شده است. شکل ۱(A)، فراوانی بیماران سانسورشده و فوت شده را به تفکیک زمان (ماه) نشان می‌دهد، که با توجه به این شکل بیشترین فوت در همان هفته‌های اول است. شکل۱ (B)، درصد سانسورشدگی و مرگ را به تفکیک جنسیت نشان می‌دهد. شکل۱ (C)، درصد سانسورشدگی و مرگ را به تفکیک معیار KPS و شکل۱ (D)، درصد معیار KPS به تفکیک جنسیت است. شکل۱ (E)، درصد سانسورشدگی و مرگ را در رده‌های مختلف سن بیان می‌کند و شکل۱ (F)، درصد رده‌های مختلف سنی به تفکیک جنسیت در داده‌ها را نشان می‌دهد.

[۱۳] Karnofsky performance status



شکل ۱. نمودارهای توصیفی داده‌ها



شکل ۲. برآورد تابع بقا در مقابل زمان بقا (راست) ؛ احتمال بقا به تفکیک جنسیت (چپ).

۳.۳ تحلیل داده‌ها با دو روش KM و BART

داده‌ها با روش ناپارامتری کاپلان-مه‌یر (KM) و روش ناپارامتری رگرسیون جمعی درختی بیزی ($BART$) مورد تحلیل قرارگرفتند. شکل ۲ (راست)، برآورد تابع بقا در مقابل زمان بقا و فاصلهٔ اطمینان ۹۵٪ با روش KM را نشان می‌دهد و بیان‌کنندهٔ این است که احتمال بقا تا ۱۰ ماه حدود ۵۰٪ و ۷۵٪ قبل از ماه بیستم فوت می‌کنند. شکل ۲ (چپ)، برآورد احتمال بقا و فاصلهٔ اطمینان ۹۵٪ به تفکیک جنسیت با روش KM است، که نشان می‌دهد احتمال بقای بیماران زن مبتلا به سرطان ریه بیشتر از بیماران مرد است. شکل ۳ برآورد احتمال بقا و فاصلهٔ اطمینان را با تقسیم‌بندی سن به دو رده بیماران کمتر از ۶۰ سال و بیماران بالاتر از ۶۰ سال را نشان می‌دهد که بیان‌کنندهٔ این مطلب است که بقای بیماران کمتر از ۶۰ سال بیشتر از بیماران بالای ۶۰ سال است. حال داده‌ها با روش BART بررسی می‌شوند. در شکل ۴ برآورد بقای بیماران با روش BART انجام شده است. مزیت این روش این است که می‌توان به‌طور جزئی‌تر به تحلیل داده‌ها پرداخت چون در این روش از روش‌های تکرار شوند و الگوریتم‌های مونته‌کارلویی استفاده می‌شود و پس از ۷۰۰۰۰ تکرار با داغین ۱۵۰۰۰ برای الگوریتم‌های $MCMC$ نمودارها رسم شده‌اند.

در شکل ۴ برآورد بقای بیماران با روش BART به‌صورت جزئی در چهار نمودار مشاهده می‌شود. شکل ۴ (بالا-راست) بقای بیماران ۷۰ ساله به تفکیک جنسیت با معیار $KPS = ۶۰$ که این گروه، بیماران با ریسک خطر بالا نامیده می‌شوند رسم شده است و شکل ۴ (بالا-چپ) نمودار بقای بیماران ۵۰ ساله به تفکیک جنسیت با معیار $KPS = ۹۰$ است، که گروه بیماران با ریسک خطر پایین نامیده می‌شود. همان‌طور که در این دو نمودار مشاهده می‌شود، احتمال بقای بیماران در گروه ریسک خطر پایین بیشتر از گروه بیماران با ریسک خطر بالا است. در شکل ۴ (پایین-راست) و شکل ۴ (پایین-چپ) نمودار برآورد بقای بیماران مبتلا به سرطان ریه به تفکیک جنسیت و دو گروه سنی ۵۰ و ۷۰ رسم شده‌اند. این نمودارها نشان می‌دهند که برآورد بیزی احتمال بقای بیماران ۵۰ ساله بیشتر از بیماران ۷۰ ساله است.

شکل ۵ برآورد بیزی احتمال بقای بیماران در دو گروه ریسک خطر بالا و ریسک خطر پایین و فاصلهٔ اطمینان ۹۵٪ به تفکیک جنسیت را نشان می‌دهد. شکل ۶ نمودار پیشگویی احتمال بقای دو گروه با ریسک خطر بالا و پایین به تفکیک جنسیت و در مقایسه با هم رسم شده است، که نشان می‌دهد احتمال بقای گروه ریسک خطر پایین و همچنین در گروه ریسک خطر بالا احتمال بقای زنان بیشتر از مردان است. این نمودار پیشگویی احتمال بقای دو گروه است، که با استفاده از برآورد مدل BART برازش‌شده در قبل، پیشگویی شده است.

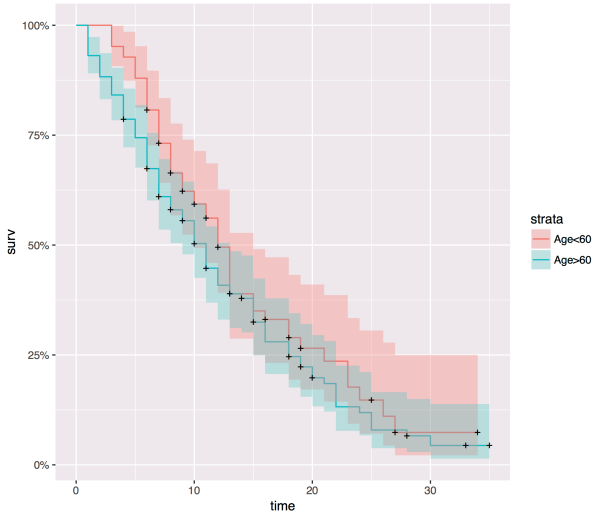
در نرم‌افزار R برای برازش پیشگویی مدل BART از پکیج BART و survbart استفاده می‌شود و از توابع surv.bart و surv.pre.bart و predict به‌ترتیب برای برازش و پیشگویی دو گروه ریسک خطر بالا و پایین استفاده شده است. برای بررسی همگرایی الگوریتم $MCMC$، از آماره $Z - Score$، استفاده می‌شود. در شکل ۷ مشاهده می‌شود که برای ۱۰ نمونه بیمار انتخاب شده اعداد بین ۱٫۹۶ و ۱٫۹۶- هستند.

بحث و نتیجه‌گیری

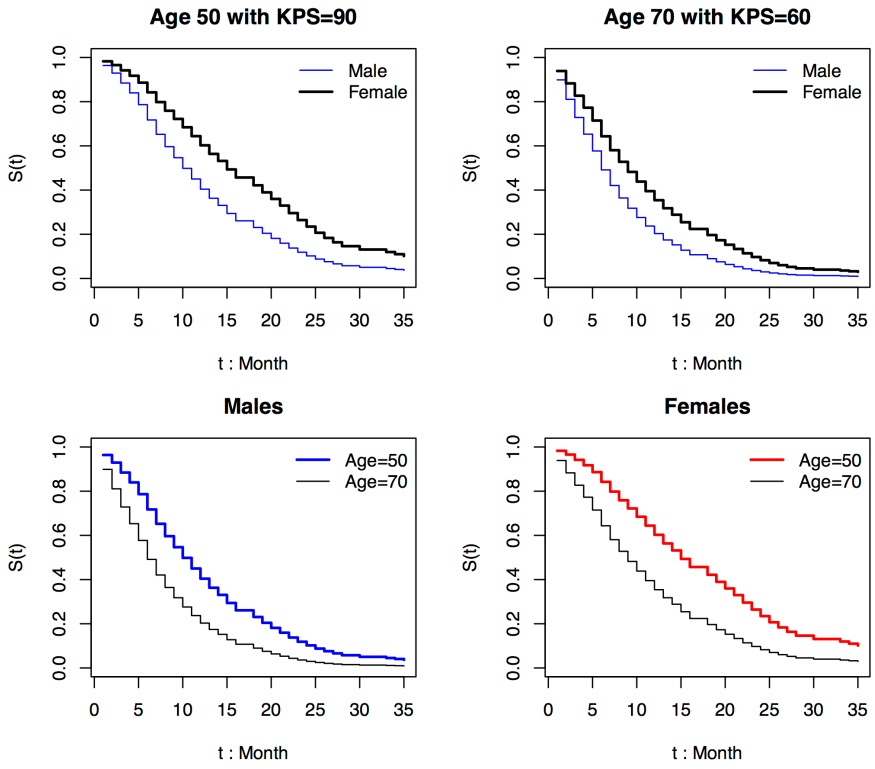
پیشرفت‌های به وجود آمده در گردآوری داده‌ها و قابلیت‌های ذخیره‌سازی در طی دهه‌های اخیر باعث شده بسیاری از علوم با حجم عظیمی از داده‌ها روبرو شوند. محققان در زمینه‌های مختلف مانند مهندسی، اقتصاد، ستاره‌شناسی و زیست‌شناسی هر روز با مشاهدات بیشتر و بیشتری روبرو می‌شوند به‌طوری که گاهی روش‌های آماری سنتی به علت افزایش تعداد مشاهداتناکارامد هستند. مدل‌های درختی یک روش خوب و کارامد برای تحلیل مجموعه‌داده‌های بزرگ است که به‌وسیله تقسیم‌بندی فضای پیش‌بینی کننده‌ها به نواحی ساده‌تر به تحلیل روشن‌تری از داده‌ها می‌پردازد.

در این مقاله رگرسیون درختی جمعی بیزی (BART) مورد بررسی و در تحلیل بقا مورد استفاده قرار گرفت که یک رهیافت ناپارامتری است و برای انواع مختلف داده اعم از پیوسته و گسسته و ترکیب آنها قابل استفاده است. روش BART

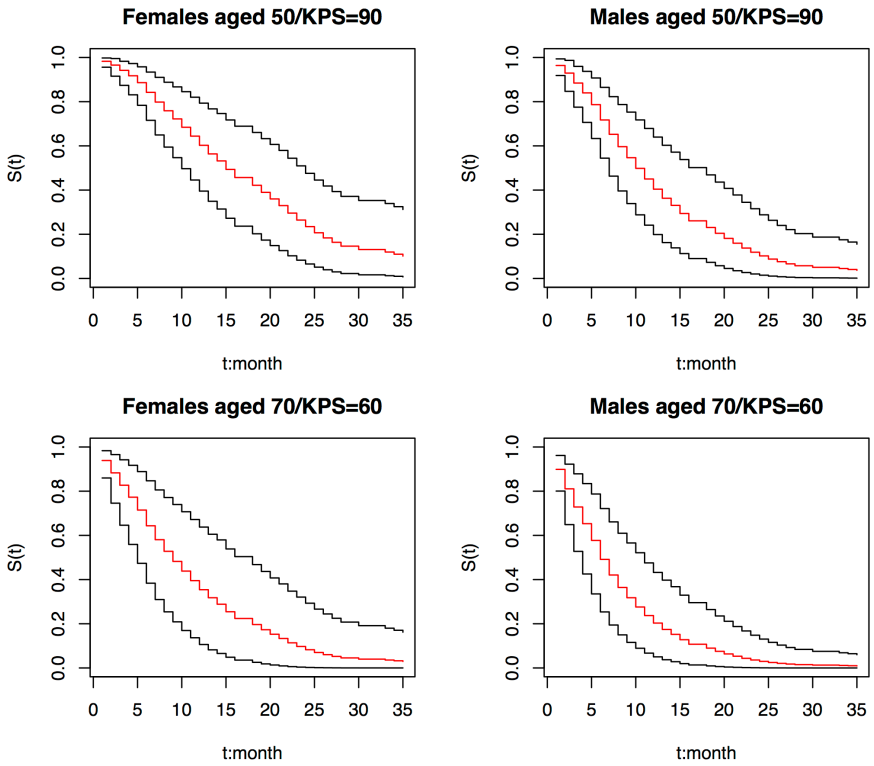
را می‌توان به‌عنوان یک روش کارآمد برای کسب دانش از مجموعه‌داده‌ها تلقی کرد. از مزیت‌های این روش می‌توان به این نکته اشاره کرد که با استفاده از این روش می‌توان به زوایای مختلف مجموعه‌داده‌ها از جمله برآورد پارامترهای مربوط، یافتن مهمترین متغیرها، استنباط و پیش‌بینی آگاه گشت به‌عبارت دیگر چون این روش مبتنی بر درخت است این امکان را به ما می‌دهد تا ویژگی‌ها با تأثیرات بیشتر را در دسته‌بندی اطلاعات تشخیص دهیم.



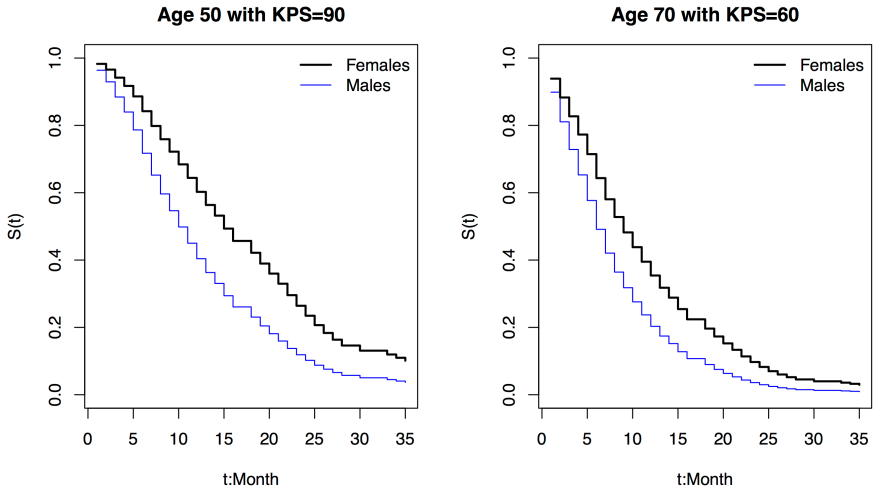
شکل ۳. احتمال بقا با تقسیم سن



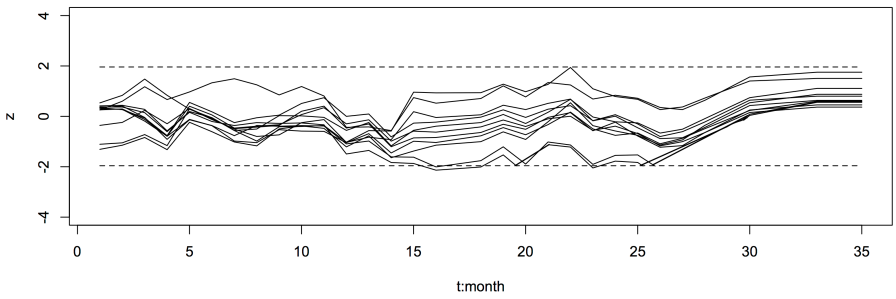
شکل ۴. برآورد بقای بیماران با روش



شکل ۵. برآورد بیزی احتمال بقای بیماران در دو گروه ریسک خطر بالا و پایین



شکل ۶. نمودار پیشگویی احتمال بقای دو گروه با ریسک خطر بالا و پایین به تفکیک جنسیت



شکل ۷. نمودار بررسی همگرایی

مراجع

[1] Chipman, H. A., George, E. I., and Mcculloch, R. E. (2002). Bayesian treed models. *Machine Learning*, **48**, 299–320.

[2] Chipman, H. A., George, E. I., and Mcculloch, R. E. (1998). Bayesian CART model search (with discussion and a rejoinder by the authors). *Journal of the American Statistical Association*, **93**, 935–960.

[3] Chipman, H. A., George, E. I., and McCulloch, R. E. (2007). *B*ayesian ensemble learning, Neural Information Processing Systems, 19.

[4] Crowley, J., and Hoering, A. (2017). *Handbook of Statistics in Clinical Oncology*, Chapman and Hall/CRC.

[5] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Randomized trees, *Journal of Machine Learning Research*, **63**(**1**), 3–42.

[6] Hastie, T., and Tibshirani, R. (2000). Byesian Backfitting. *Statistical Science*, 15, **3**, 196-223.

[7] Holfprd, T. R., (2002). *M*ultivarite methods in epidemiology, New York, OxfordUniversity Press.

[8] Karnofsky, D. A., Abelmann, W. H., Craver, L. F., and Burchenal, J. H. (1948). The use of the nitrogen mustards in the palliative treatment of carcinoma. With particular reference to bronchogenic carcinoma. *Cancer*, **1**(**4**), 634-656.

[9] Saparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W., (2016). Nonparametric Survival analysis using Bayesian Additive Regression Trees (BART), *Statistics in medicine*, **35**(**16**), 2741-2753.

[10] Siciliano, R., and Mola, F. (2000). Multivariate dataanalysis anmodeling through classification and regression tree, *Computational Statistics and Data Analysis*, **32**, 285-301.

[11] Sorokina, D., Caruana, R., and Riedewald, M., (2007). Additivegroves of regression trees. *Machine Learning*, 323–334.