دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی

عنوان:

آنالیز بقای بیماران سرطان دهان

نگارش:

کیوان رضائی

استاد راهنما:

دکتر مهدیه سلیمانی

تیر ۱۴۰۱


بسم الله الرحمن الرحیم

چکیده

پیش‌بینی شدت یک بیماری براساس وضعیت فعلی بیمار در به کارگیری شیوه‌ها و برنامه‌های درمانی مناسب بسیار حائز اهمیت است. توسعه مدل‌هایی که بر مبنای داده‌های کلینیک‌های درمانی بتوانند شدت بیماری در یک بیمار را معین کنند، می‌تواند در این راستا مفید واقع شود. این داده‌ها اکثراً در برگیرنده تعداد زیادی از ویژگی‌های بیماران هستند که غالباً سانسورشده و بخشی از اطلاعات ناقص می‌باشد. این عوامل سبب می‌شوند که برای آنالیز بقا و استخراج اطلاعات از ویژگی‌های بیماران نیاز به شیوه‌هایی داشته باشیم که توانایی حل این چالش‌ها را داشته باشند. اخیراً شیوه‌های مبتنی بر یادگیری ماشین به خاطر توانایی‌شان در یادگیری توابع پیچیده از داده‌های در ابعاد بالا توجهات زیادی در زمینه‌های مختلف از جمله آنالیز بقا به خود جلب کرده‌اند. در این پایان‌نامه ما به بررسی گونه‌های مختلف الگوریتم‌های یادگیری ماشین در زمینه آنالیز بقا می‌پردازیم و آنها را روی داده‌هایمان آزمایش کرده و نتایج مدل‌های مختلف را با یکدیگر مقایسه می‌کنیم. مدل Logistic Hazard و جنگل بقای تصادفی بهترین عملکرد را در بین مدل‌های آزمایش شده دارند. همچنین با کمک این مدل‌ها، ویژگی‌هایی که اهمیت بیشتری در پیش‌بینی شدت بیماری و بقای بیمار دارند را به دست می‌آوریم که به نوعی در تفسیر مدل‌های مربوطه به ما کمک می‌کنند.

کلیدواژه‌ها: یادگیری ماشین، آنالیز بقا، شبکه‌عصبی، انتخاب ویژگی، تفسیرپذیری

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱

مقدمه

در ابتدا در مورد مسأله‌ی آنالیز بقا صحبت خواهیم کرد و آن را شرح خواهیم داد. سپس در مورد بیماری سرطان دهان که در این پایان‌نامه به آن پرداخته‌ایم، صحبت خواهیم کرد و از اهمیت و کاربرد آنالیز بقا در مورد این بیماری خواهیم گفت. سپس کلیتی از آنچه در این پایان‌نامه بررسی شده و اهداف این تحقیق بیان می‌شود. نهایتاً در انتهای این فصل، ساختار پایان‌نامه و کلیتی از محتوای فصل‌های آن ارائه خواهد شد.

۱ـ۱ تعریف مسئله‌ی آنالیز بقا

مسأله‌ی آنالیز بقا در واقع تعیین مدت زمان میان تشخیص یک بیماری و مرگ ناشی از آن است [۱]. هدف این است که بتوانیم براساس داده‌هایی که از یک بیمار داریم تخمینی از مدت زمانی که او زنده می‌ماند و یا شدت بیماری داشته باشیم. این مسأله مدت زیادی است که بررسی می‌شود و روش‌ها و شیوه‌های گوناگونی برای آن وجود دارد. در گذشته روش‌های کلاسیک استفاده می‌شد که مبنای آماری داشتند، اما اخیراً روش‌های متنوع یادگیری ماشین هم برای این هدف به کار گرفته می‌شوند. مسأله آنالیز بقا اکثراً بدین شکل بررسی می‌شود که داده‌های تعدادی از بیماران موجود است و مدل ما، چه به شکل آماری چه با کمک یادگیری ماشین، اطلاعاتی از داده‌ها استخراج کرده و پیش‌بینی را انجام می‌دهد. سپس مدل می‌تواند با گرفتن داده‌های بیماران جدید، تخمینی از مدت زمان بقای آن‌ها ارائه دهد.

نکته‌ای که در خصوص تحلیل مسأله بقا وجود دارد این است که داده‌هایمان به شکل سانسور شده[۱] هستند. به این معنی که ما در داده‌هایمان از بیماران:

- تعدادی بیمار داریم که فوت شده‌اند و مدت زمان بقایشان پس از تشخیص بیماری و انجام آزمایشات پزشکی را می‌دانیم.
- تعدادی بیمار داریم که فوت نشده‌اند یا اطلاعات فوت آنها در دسترس نیست و ما صرفاً تاریخ آخرین مراجعه‌ی آنها را داریم و می‌دانیم که تا زمان آخرین مراجعه زنده بوده‌اند.

این موضوع سبب می‌شود که مسأله آنالیز بقا کمی متفاوت‌تر از سایر مسائل موجود شود و راهکارها و مدل‌های متفاوتی برای حل این مسأله طراحی و به کار گرفته شوند.

۱ـ۲ سرطان دهان و اهمیت تحلیل بقا

سرطان دهان از سرطان‌های به نسبت رایج می‌باشد. در سال ۲۰۲۰ میلادی، تقریباً ۲ درصد سرطان‌هایی که در دنیا تشخیص داده شدند، سرطان دهان بودند. بیش از ۳۷۰ هزار نفر در آن سال مبتلا به سرطان دهان تشخیص داده شدند و ۱۷۰ هزار نفر نیز براساس ابتلا به سرطان دهان فوت کردند [۲]. این سرطان خطرناک و کشنده است و تنها کمی بیشتر از نصف افرادی که به آن مبتلا می‌شوند می‌توانند بیش از ۵ سال زنده بمانند [۳]. براساس این آمار، می‌توان فهمید که نرخ مرگ سرطان دهان بالاست و فرصت طولانی برای درمان و معالجه آن وجود ندارد و ممکن است سرطان به سرعت رو به وخامت برود و سبب فوت بیمار گردد. لذا بهره‌گیری از شیوه‌های درست درمانی برای مقابله با سرطان دهان بسیار مهم است و درمان‌های اشتباه و نامناسب می‌تواند فرصت کوتاه درمان را تلف کند. بیماری که بقای کوتاه‌تری برایش پیش‌بینی شده است، نیاز به اجرای روش‌های درمانی پرریسک‌تر و سنگین‌تر دارد؛ در سمت مقابل برای بیماری که بقای طولانی پیش‌بینی شده است احتمالاً درمان‌های سبک‌تر هم می‌تواند مفید واقع شود [۴]. پس پیش‌بینی بقای بیماران سرطان دهان و نیز تحلیل اهمیت ویژگی‌هایی که از بیماران استخراج شده است و شناخت ویژگی‌هایی که می‌توانند در پیش‌آگاهی به ما کمک بیشتری کنند حائز اهمیت است.

[۱] Censored Data

۱ـ۳ داده‌ها

داده‌هایی که در این پایان‌نامه مورد استفاده قرار می‌گیرند، داده‌هایی هستند که از بیماران سرطان دهان بیمارستان شهید بهشتی به دست آمده‌اند. این داده‌ها در برگیرنده‌ی اطلاعات ۵۲۶ بیمار مبتلا به سرطان دهان هستند. برای هر بیمار ۲۲ ویژگی به دست آمده که ما از این ویژگی‌ها برای تحلیل بقای بیماران استفاده می‌کنیم. در فصل‌های آتی به بیان جزئیات ویژگی‌ها خواهیم پرداخت.

۱ـ۴ اهداف تحقیق

در این پایان‌نامه، قصد داریم گونه‌های مختلف روش‌های تحلیل بقا را روی داده‌های بیماران سرطان دهان که در اختیار داریم، بررسی کنیم و عملکرد آن‌ها را با یکدیگر مقایسه کنیم. تعداد زیادی روش بررسی خواهد شد و هر یک از آنها از نظر معیارهای گوناگونی که برای سنجش کارایی روش‌های تحلیل بقا وجود دارد ارزیابی می‌شوند. همچنین به بررسی ویژگی‌های مختلف در داده‌هایمان می‌پردازیم و سعی می‌کنیم ویژگی‌ها و عواملی که در تشخیص بقای بیماران موثرتر هستند را شناسایی کنیم. به بیان بهتر می‌خواهیم اهمیت ویژگی‌ها را متوجه شویم. مشخص شدن ویژگی‌های مهم‌تر می‌تواند در معاینات و مطالعات پزشکی هم موثر واقع شود.

۱ـ۵ ساختار پایان‌نامه

این پایان‌نامه شامل پنج فصل است. فصل دوم دربرگیرنده‌ی تعاریف و مفاهیم اولیه‌ی مرتبط با پایان‌نامه است. همچنین در خصوص داده‌هایمان و آماده‌سازی آنها برای آزمایشات توضیحاتی ارائه می‌شود. در فصل سوم، به معرفی و بیان روش‌های گوناگون حل مسأله تحلیل بقا خواهیم پرداخت. همچنین چند مورد از معیارهایی که در مسأله تحلیل بقا برای سنجش کارایی و دقت مدل‌ها موجود است را معرفی می‌کنیم. در فصل چهارم، ایده‌هایی که برای تعیین اهمیت ویژگی‌ها در مدل‌های مختلف آزمایش کردیم را شرح می‌دهیم. با کمک این ایده‌ها سعی کردیم که ویژگی‌های مهم‌تر در مسأله تحلیل بقا را شناسایی کنیم. در فصل پنجم نتایج به کارگیری مدل‌های معرفی‌شده روی داده‌هایمان را بیان می‌کنیم و به مقایسه مدل‌های مختلف می‌پردازیم. همچنین نتایج تحلیل اهمیت ویژگی‌ها را شرح می‌دهیم و ویژگی‌هایی که

از نظر مدل‌های مختلف حائز اهمیت بیشتری هستند را مشخص می‌کنیم. فصل ششم به نتیجه‌گیری و پیش‌نهادهایی برای کارهای آتی خواهد پرداخت.

فصل ۲

تعاریف و مفاهیم اولیه

در این فصل به برخی مفاهیمی که در مسأله‌ی تحلیل بقا استفاده می‌شود می‌پردازیم. این مفاهیم در فصل‌های آتی استفاده می‌شود و به طور کلی در این زمینه تعریف‌شده و شناخته‌شده هستند. همچنین برخی از معیارهای بررسی کیفیت ومقایسه مدل‌های آنالیزگر بقا را شرح می‌دهیم . سپس در مورد جزئیاتی داده‌ها و نیز شیوه‌های کلی مدل‌های آنالیز بقا توضیحاتی ارائه خواهد شد.

۲ـ۱ توابع آنالیز بقا

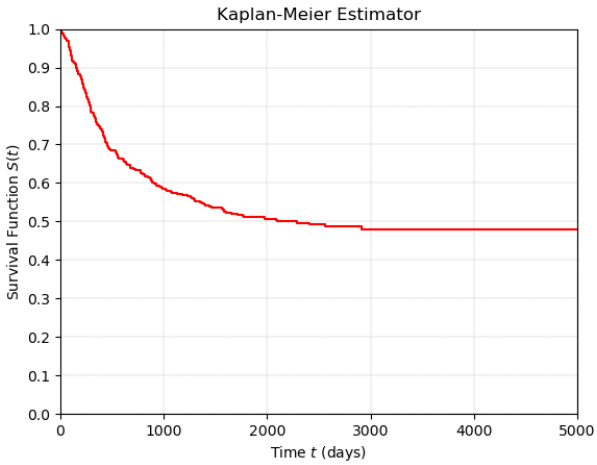
در تعاریف زیر از مقدار t_i استفاده می‌شود. این مقدار برای بیمار شماره‌ی i، در صورتی که در اثر سرطان فوت کرده باشد، فاصله‌ی زمانی تشخیص تا فوت است و در صورتی که فوت نکرده باشد، فاصله‌ی زمانی تشخیص تا آخرین مراجعه‌ی بیمار می‌باشد.

یکی از اهداف مهم در آنالیز بقا، یافتن تابع بقاست [۱].

تعریف ۲ـ۱ (تابع بقا [۵]) *این تابع که آن را با $S : \mathbb{R}^+ \to [0, 1]$ نشان می‌دهیم، به این شکل تعریف می‌شود که اگر زمان فوت بیمار (پس از تشخیص بیماری) را با T نشان دهیم، مقدار $S(t)$ برابر است با احتمال اینکه یک بیمار حداقل به اندازه t پس از تشخیص زنده بماند. به بیان بهتر*

$$S(t) = \mathbb{P}[T \geqslant t].$$

[۱] Survival Function



شکل ۲ـ۱: تابع بقا؛ در اثر اجرای الگوریتم Kaplan-Meier روی کل داده‌ها

این تابع می‌تواند برای یک بیمار و براساس ویژگی‌های او استخراج شود که در این صورت به فرم $S(t|x)$ خواهد بود که x بردار ویژگی‌های بیمار است. همچنین می‌تواند به شکل جمعی و برای یک مجموعه از بیماران محاسبه شود. در این صورت تابع می‌تواند خطرناک بودن بیماری، نرخ مرگ و میر و متوسط عمر افراد فوتی را نمایش دهد. یکی از مدل‌هایی که برای تخمین تابع بقا به شکل جمعی استفاده می‌شود، تخمین‌گر *Kaplan-Meier* می‌باشد. این تخمین‌گر برای هر t، مقدار $S(t)$ را به شکل زیر حساب می‌کند:

$$S(t) = \prod_{i,t_i \leqslant t} (۱ - \frac{d_i}{n_i}).$$

که در رابطه‌ی بالا، d_i برابر تعداد افرادی است که در لحظه‌ی t_i مرده‌اند و n_i تعداد افرادی است که تا پیش از لحظه t_i زنده بوده‌اند.

با اجرای این الگوریتم روی کل داده‌ها، تابع بقا مطابق با شکل ۲ـ۱ خواهد بود. همانطور که مشاهده می‌شود تقریباً نیمی از بیماران براساس داده‌های ما فوت نکرده‌اند، اما برای بیمارانی که فوت کرده‌اند، تابع بقا با شیب زیادی نزولی است که نشان می‌دهد بیماری کشنده بوده و در زمان کمی می‌تواند بیمار را از بین ببرد.

تعریف ۲ـ۲ (تابع خطر [۵]) *تابع خطر که به صورت $h : \mathbb{R}^+ \to [۰, ۱]$ تعریف می‌شود، قرار است میزان خطر در لحظه‌ی t را معین کند. این تابع به صورت زیر محاسبه می‌شود:*

$$h(t) = \mathbb{P}\left[t \leqslant T \leqslant t + \Delta t \mid T \geqslant t\right] = \lim_{dt \to ۰} \frac{\mathbb{P}\left[t \leqslant T \leqslant t + dt\right]}{\mathbb{P}\left[t \leqslant T\right]} = -\frac{S'(t)dt}{S(t)} = -\frac{d}{dt} \ln S(t)$$

تابع خطر همانطور که گفته شد، وقوع مرگ بیمار در لحظه t را بررسی می‌کند. با این فرض که می‌دانیم بیمار تا لحظه‌ی t زنده بوده است، می‌خواهیم ببینیم به چه احتمالی هم‌اکنون فوت می‌کند و در واقع خطر در لحظه‌ی t برای بیمار چقدر است.

به شکل مشابه، تابع خطر تجمعی [۳] نیز تعریف می‌شود.

تعریف ۲ـ۳ (تابع خطر تجمعی [۵]) *تابع تجمعی خطر که آن را به صورت* $H : \mathbb{R}^+ \to [۰, ۱]$ *تعریف می‌کنیم، میزان خطر تا لحظه‌ی* t *را حساب می‌کند. این تابع به شکل زیر محاسبه می‌شود:*

$$H(t) = \int_{s=۰}^{t} h(s)ds = -\ln S(t) + \ln S(۰)) = -\ln S(t)$$

تابع خطر تجمعی، می‌تواند میزان خطر تا لحظه‌ی t را برای یک بیمار معین کند. در واقع اگر تابع بقای یک بیمار را داشته باشیم، تابع خطر جمعی، میزان خطر تجربه شده تا لحظه‌ی t را نمایش می‌دهد. بعضی از مدل‌ها که در ادامه آن‌ها را معرفی خواهیم کرد، تابع $H(t|x)$ که در واقع تابع خطر تجمعی یک بیمار با بردار ویژگی‌های x است را تخمین می‌زنند.

بر اساس تابع بالا، یک امتیاز ریسک [۴] تعریف می‌شود که به کمک آن می‌توان بیماران مختلف یک مجموعه داده را با یکدیگر مقایسه کرد.

تعریف ۲ـ۴ (امتیاز ریسک [۵]) *امتیاز ریسک با داشتن تابع تجمعی خطر در* n *نقطه‌ی مربوط به فوت یا آخرین مراجعه‌ی مجموعه‌ی بیماران محاسبه می‌شود (همان* t_i *ها). این تابع که آن را برای بیمار با ویژگی‌های* x *به شکل* $R(x)$ *نشان می‌دهیم. به شکل زیر محاسبه می‌شود:*

$$R(x) = \sum_{i=۱}^{n} H(t_i \mid x)$$

اکنون برای مقایسه دو بیمار، می‌توانیم امتیاز ریسکشان را محاسبه کنیم و بیماری که امتیاز بالاتری کسب می‌کند، یعنی شدت بیماری برایش شدیدتر بوده و بیشتر در خطر است. برخی از مدل‌ها که در ادامه معرفی می‌شوند از این امتیاز ریسک برای ترتیب‌دهی به بیماران براساس شدت بیماری‌شان استفاده می‌کنند.

[۳] Cumulative Hazard Function
[۴] Risk Score

۲ـ۲ بررسی دقت مدل‌های آنالیز بقا

در این قسمت می‌خواهیم به معرفی چندین معیار [۵] برای سنجش دقت مدل‌های آنالیز بقا بپردازیم. در فصل نتایج، از این معیارها جهت مقایسه و تحلیل مدل‌ها استفاده خواهد شد. محتوای این بخش براساس منبع [۵] می‌باشد.

تعریف ۲ـ۵ (معیار c-index [۶]) *این معیار، بر اساس پیش‌بینی $r_{i=1\dots n}$ مدل که میزان امتیاز ریسک بیماران را نشان می‌دهد، مقدار C_{index} را بدین شکل محاسبه می‌کند:*

$$C_{index} = \frac{\sum_{i,j} I(t_i > t_j, j \textit{ dead}, r_i < r_j)}{\sum_{i,j} I(t_i > t_j, j \textit{ dead})}.$$

در حقیقت معیار C_{index}، قرار است برای هر جفت (i,j) که فرد j فوت کرده باشد و فرد i یا دیرتر از او مرده باشد، یا مطمئن باشیم که مدت طولانی‌تری عمر کرده است (براساس زمان آخرین مراجعه)؛ ببینیم که آیا مدل امتیاز ریسک درستی پیش‌بینی کرده است یا خیر. در حقیقت برای جفت (i,j) که j مرده باشد و i عمر بیشتری کرده باشد، توقع داریم که مدلمان، امتیاز ریسک نفر i را کمتر از امتیاز ریسک نفر j پیش‌بینی کند. معیار C_{index} در واقع نسبت تعداد جفت‌های (i,j) ای است که مدل درست پیش‌بینی کرده به تعداد کل (i,j) های معتبر. طبیعی است که مقدار C_{index} عددی در بازه‌ی $[۰,۱]$ است و هر چه به ۱ نزدیک‌تر باشد، یعنی مدل بهتر توانسته است براساس داده‌ها، امتیاز ریسک بیماران را نسبت به هم پیش‌بینی کند. اگر یک مدل تصادفی داشته باشیم که امتیاز ریسک را به شکل تصادفی استخراج کند، مقدار C_{index} برابر ۰/۵ می‌شود.

تعریف ۲ـ۶ (امتیاز Berier [۷]) *معیار Berier در هر زمان t طبق رابطه‌ی زیر محاسبه می‌شود:*

$$BS(t) = \frac{۱}{n}\sum_{i=۱}^{n} I(t_i \leqslant t, i \textit{ dead})\frac{(۰ - S(t \mid x_i))^۲}{G(t_i)} + I(t_i > t)\frac{(۱ - S(t \mid x_i))^۲}{G(t)}.$$

این معیار در هر لحظه‌ی t میزان خطای مدل را محاسبه می‌کند. برای این منظور،

- برای افرادی که تا لحظه‌ی t فوت شده‌اند $(t_i \leqslant t)$، مقدار $\frac{(۰-S(t \mid x_i))^۲}{G(t_i)}$ را به عنوان خطا در نظر می‌گیرد. علت این است که چون فرد i، پیش از لحظه‌ی t مرده است، پس ما توقع داریم که تابع بقایی که برای او پیش‌بینی می‌کنیم به گونه‌ای باشد که پس از t_i، مقدار $S(t|x_i)$ بسیار کوچک و نزدیک ۰ باشد. به همین جهت مربع فاصله‌ی آن با ۰ را به عنوان خطا در نظر می‌گیریم.

[۵] Criteria

- برای افرادی که تا لحظه‌ی t فوت نشده‌اند ($t_i > t$)، مقدار $\frac{(۱-S(t \mid x_i))^۲}{G(t)}$ به عنوان خطا در نظر گرفته شده است. علت این است که چون فرد i تا لحظه‌ی t زنده بوده است، پس ما توقع داریم که تابع بقای $S(t|x_i)$ تا آن لحظه مقداری نزدیک به ۱ داشته باشد. به همیت جهت مربع فاصله‌ی تابع با ۱ را به عنوان خطا در نظر می‌گیریم. نهایتاً میانگین خطاها برای همه‌ی افراد به عنوان $BS(t)$ گزارش می‌شود.

در رابطه‌ی بالا، $G(t)$ برابر احتمال این است که داده‌ی سانسور شده (که در واقع زمان آخرین مراجعه‌ی بیمار است) که آن را با C نشان می‌دهیم، بیشتر از t باشد. به بیان بهتر $G(t) = \mathbb{P}[C \geqslant t]$ است. همانطور که می‌بینید، تعریفی شبیه تابع بقا دارد و از روش Kaplan-Meier برای تخمین آن استفاده می‌شود. همچنین با تجمیع مقدار امتیاز Berier روی همه‌ی زمان‌های ممکن، می‌توان امتیاز تجمعی Berier [۸] را تعریف کرد.

تعریف ۲ـ۷ (امتیاز تجمعی Berier) *امتیاز تجمعی Berier که آن را با IBS نشان می‌دهیم در بازه‌ی زمانی $[t_s, t_e]$ به صورت زیر محاسبه می‌شود:*

$$\frac{۱}{t_e - t_s} \int_{t_s}^{t_e} BS(t)dt.$$

با کمک این امتیاز می‌توان معیار Berier را به شکل تجمعی روی زمان‌های مختلف حساب کرد و از آن برای سنجش دقت مدل استفاده کرد. در حالت گسسته، انتگرال بالا تنها به ازای n نقطه‌ی مربوط به t_i ها محاسبه می‌شود و به کمک روش ذوزنقه‌ای [۹] انتگرال محاسبه می‌شود. طبیعتاً هر چه امتیاز Berier کوچک‌تر باشد یعنی دقت مدل بهتر بوده است و بهتر توانسته که تابع بقا را براساس داده‌ها پیش‌بینی کند. اگر تابع بقا به شکل تصادفی ساخته شود، این معیار عددی در حدود ۰/۲۵ خواهد داشت.

تعریف ۲ـ۸ (امتیاز AUC [۱۰]) *امتیاز AUC براساس خروجی امتیاز ریسک مدل برای بیماران که با $r_{i=۱...n}$ نشان می‌دهیم، بدین شکل در لحظه‌ی t محاسبه می‌شود:*

$$AUC(t) = \frac{\sum_{i,j} I(y_j > t) I(y_i \leqslant t, i\ dead) w_i I(r_i > r_j)}{\sum_{i,j} I(y_j > t) I(y_i \leqslant t, i\ dead) w_i}.$$

که در رابطه‌ی بالا $w_i = \frac{۱}{G(t_i)}$ است.

[۸] Integrated Berier Score
[۹] Trapezoidal

این معیار در لحظه‌ی t، بیمارانی که تا لحظه‌ی t فوت شده‌اند و بیمارانی که مطمئنیم حداقل تا لحظه‌ی t زنده بوده‌اند را جفت می‌کند و امتیاز بالا می‌رود اگر مدل ما امتیاز ریسک را برای این جفت‌ها به درستی محاسبه کرده باشد. طبیعتاً این امتیاز هر چه به ۱ نزدیک‌تر باشد، یعنی دقت مدل بیشتر است. مشابه بالا می‌توانیم امتیاز AUC تجمیعی را هم در بازه‌ی $[t_s, t_e]$ بدین شکل محاسبه کرد:

$$AUC = \frac{۱}{t_e - t_s} \int_{t=t_s}^{t=t_e} AUC(t)dt.$$

۲ـ۳ مقدمه‌ای بر روش‌های آنالیز بقا

روش‌های آنالیز بقا امروزه عمدتاً با یادگیری ماشین عجین شده‌اند. اما روش‌های کلاسیک هم همچنان استفاده می‌شوند. یکی از این روش‌های معروف شیوه‌ی Cox-PH است. این روش براساس داده‌ها تابع خطر را برای یک بیمار در هر لحظه‌ی t محاسبه می‌کند.

اما روش‌های یادگیری ماشین که به خاطر توانایی‌شان در استخراج دانش از داده‌ها در ابعاد بالا امروزه بیش از پیش مورد توجه قرار گرفته‌اند، در اشکال و انواع گوناگونی در آنالیز بقا هم به کار گرفته می‌شوند. شیوه‌های مبتنی بر یادگیری ماشین در آنالیز بقا، به طور کلی در ۴ دسته قرار می‌گیرند [۵].

مدل‌های دسته‌ی نخست، مبتنی بر جنگل‌های تصادفی [۱۱] می‌باشند که با تجمیع تعداد زیادی مدل ساده‌تر به نام درخت تصمیم [۱۲]، از پیش‌بینی همگی آن مدل‌ها استفاده می‌کنند و ترکیبی از خروجی مدل‌ها را به عنوان نتیجه نهایی اعلام می‌کنند. این فرایند را تجمیع خود راه‌انداز[۱۳] می‌گویند. نکته‌ی قوت این مدل‌ها در سادگی پیاده‌سازی و نیز تفسیرپذیری‌شان است [۶].

دسته دوم از مدل‌ها، مبتنی بر ماشین‌های تقویت گرادیان[۱۴] است. ایده‌ی تقویت در یادگیری ماشین بدین صورت است که یک مدل قوی می‌تواند در اثر تجمیع تعدادی مدل ضعیف [۱۵] ایجاد شود. تفاوت این دسته از مدل‌ها، با مدل‌های دسته‌ی نخست این است که در جنگل‌های تصادفی، هر یک از درخت‌های تصمیم به شکلی مستقل از روی داده چیزی می‌آموزند اما در این دسته از مدل‌ها، مدل‌های ضعیف‌تر به شکل متوالی آموزش داده می‌شوند و هر کدام براساس مدل‌های پیشین خود آموزش می‌بینند. نهایتاً ترکیبی خطی از خروجی این مدل‌های ضعیف، خروجی نهایی را تولید می‌کند [۵، ۶].

[۱۱] Random Forests
[۱۲] Decision Tree
[۱۳] Bagging
[۱۴] Gradient Boosting
[۱۵] Weak Learner

دسته‌ی سوم از مدل‌ها، مبتنی بر ماشین بردار پشتیبانی [۱۶] می‌باشند. این مدل‌ها سعی می‌کنند که به نحوی بقای بیماران را به شکلی خطی تخمین بزنند. برخی از مدل‌های این دسته سعی می‌کنند رتبه‌بندی مناسبی از بیماران ارائه دهند.

نهایتاً در دسته‌ی آخر، شبکه‌های عصبی را داریم که قابلیت تخمین توابع بسیار پیچیده را دارند. تعیین تابع هزینه [۱۷] در این مدل‌ها بسیر حائز اهمیت است. برخی از این مدل‌ها، توابع هدف را به شکل گسسته و برخی به شکل پیوسته پیش‌بینی می‌کنند. مدل‌های گوناگونی امروزه برای آنالیز بقا پیش‌نهاد شده‌اند [۷، ۸، ۹، ۱۰، ۱۱، ۱۲] که ما برخی از آن‌ها را بررسی خواهیم کرد.

در فصل مدل‌ها، ما به بیان دقیق‌تر مدل‌هایی که روی داده‌هایمان استفاده کردیم خواهیم پرداخت و در فصل نتایج عملکرد آن‌ها را بررسی می‌کنیم.

۲ـ۴ داده‌ها

همانطور که اشاره‌شد، ما داده‌های ۵۲۶ بیمار مبتلا به سرطان دهان را جمع‌آوری کردیم. در داده‌هایمان، ما تعدادی فیلد تهی[۱۸] داشتیم که برای پر کردن داده‌های خالی از روش k نزدیک‌ترین همسایه [۱۹] با $k = ۱۱$ بهره گرفته شد. برای هر بیمار ۲۲ ویژگی داریم که توضیحات آنها به شرح زیر است.

- داده‌های دودویی
  - کاهش وزن رخ داده است؟ (Weight loss)
  - آیا غدد لنفاوی درگیر شده‌اند؟ (Lymph node involvement)
  - آیا پرتودرمانی انجام شده است؟ (Radiotherapy)
  - آیا جراحی انجام شده است؟ (Surgery)
  - آیا شیمی‌درمانی انجام شده است؟ (Chemotherapy)
  - آیا برداشت توده انجام شده است؟ (Resection)
  - آیا برداشت توده با برش گردن انجام شده است؟ (Resection with neck dissection)

[۱۶] Support Vector Machine
[۱۷] Loss Function
[۱۸] Null
[۱۹] k-Nearest Neighbors (KNN)

- وضعیت حاشیه (Margin status)
- آیا بیمار مراجعه‌ی مجدد داشته؟ (Regular follow up)
- آیا عود رخ داده؟ (Recurrence)
- آیا عود شیمی‌درمانی شده؟ (Chemotherapy after recurrence)
- آیا عود رادیوتراپی شده؟ (Radiotherapy after recurrence)
- آیا عود جراحی شده؟ (Surgery after recurrence)
- آیا متاستاز رخ داده است؟ (Metastasis)
- آیا بدخیمی ثانویه رخ داده است؟ (Second primary malignancy)
- **بیمار فوت شده است؟** (Death)

- داده‌های دسته‌ای [۲۰]
  - محل اولیه‌ی درگیری (Site)
  - سطح هیستولوژیکال (Histologic grade)
  - زمان عود (Recurrence time)
- داده‌های عددی [۲۱]
  - سایز تومور (Tumor size)
  - تعداد غدد لنفاوی درگیر (Number of involved lymph nodes)
  - سن (Age)
  - **فاصله‌ی زمانی بین تشخیص تا آخرین مراجعه یا فوت**

۲ـ۵ جمع‌بندی

در این فصل، نخست به بیان توابعی که در آنالیز بقا به کار می‌روند، پرداختیم و تابع بقا، تابع خطر، تابع خطر تجمعی و امتیاز ریسک را معرفی کردیم. سپس معیارهایی را که در سنجش عملکرد مدل‌ها

[۲۰] Categorical
[۲۱] Numerical

در آنالیز بقا استفاده می‌شوند شرح دادیم و توضیحاتی در خصوص معیار c-index ، امتیاز Berier و امتیاز AUC ارائه شد. در ادامه دسته‌بندی کلی‌ای بر روش‌های موجود در آنالیز بقا بیان شد و نهایتاً داده‌هایمان در انتهای فصل توصیف شدند.

فصل ۳

مدل‌های آنالیز بقا و تفسیرپذیری

در این فصل به مدل‌هایی که آنالیز بقا را به کمک آن‌ها انجام دادیم می‌پردازیم و آن‌ها را تا حد امکان تشریح کرده و شیوه عملکردشان را شرح می‌دهیم. همچنین در در انتها، در خصوص انتخاب ویژگی‌ها و تعیین اهمیت آن‌ها صحبت می‌کنیم و ایده‌هایی که به این منظور استفاده شدند را شرح می‌دهیم. ضمناً مجموعه‌ی D را مجموعه‌ی افراد فوت‌شده در داده‌هایمان تعریف می‌کنیم. برخی از توابع هزینه از این مجموعه استفاده می‌کنند.

Cox-PH ۱ـ۳

مدل آماری Cox-PH که تابع خطر برای بیمار با بردار ویژگی‌های x را در لحظه‌ی t به شکل

$$h(t|x) = h_0(t) \exp \beta^\intercal x$$

محاسبه می‌کند. این مدل، بردار β را با کمینه کردن تابع جزئی لگاریتم درست‌نمایی [1] به دست می‌آورد [۱۳]. همچنین یک تابع h_0 زمینه‌ای هم موجود است که به شکلی مستقل محاسبه شده و برای همه‌ی بیماران استفاده می‌شود.

تابع جزئی درست‌نمایی، هزینه را برای β بدین شکل محاسبه می‌کند:

$$\mathcal{L}(\beta) = \prod_{i \in D} \frac{h_0(t_i) \exp \beta^\intercal x_i}{\sum_{j, t_j \geqslant t_i} h_0(t_i) \exp \beta^\intercal x_j}$$

[1] partial log likelihood

در حقیقت تابع جزئی درست‌نمایی، برای فرد فوت شده‌ی i، حساب می‌کند که نسبت تابع خطرش در لحظه‌ی مرگ، به حاصل جمع تابع خطر در لحظه‌ی t_i افرادی که حداقل به اندازه‌ی i زنده مانده‌اند، چقدر است. توقع داریم که شانس مرگ فرد i نسبت به بقیه در آن لحظه خیلی بیشتر بوده باشد و این کسر نزدیک به ۱ باشد. با ساده‌کردن عبارت فوق، تابع جزئی درست‌نمایی به صورت زیر می‌شود.

$$\mathcal{L}(\beta) = \prod_{i \in D} \frac{\exp \beta^{\intercal} x_i}{\sum_{j, t_j \geqslant t_i} \exp \beta^{\intercal} x_j}$$

مدل Cox-PH قرینه لگاریتم تابع درست‌نمایی بالا، را کمینه می‌کند. در حقیقت مسأله‌ی کمینه‌سازی زیر را حل می‌کند.

$$-\ln \mathcal{L}(\beta) = -\sum_{i \in D} \left[\beta^{\intercal} x_i - \ln \sum_{j, t_j \geqslant t_i} \exp \beta^{\intercal} x_j \right]$$

۳ـ۲ جنگل بقای تصادفی

در این مدل، تعدادی درخت تصمیم تولید می‌شود. هر درخت تصمیم با بخشی از داده‌ها و زیرمجموعه‌ای ویژگی‌ها آموزش می‌بیند و هر گره‌اش، براساس یکی از ویژگی‌ها به دو بچه تقسیم می‌شود. معیار انتخاب ویژگی و تعیین مقدار آن ویژگی برای جداسازی دو بچه، بیشینه کردن معیار تست log-rank [۲] است [۱۴، ۶]. براین اساس، در هر گره، داده‌ها به گونه‌ای به دو دسته تقسیم می‌شوند که توزیع بقا در دو دسته بیشترین اختلاف را داشته باشد. این مدل ابرپارمترهای گوناگونی از قبیل حداکثر عمق درخت‌ها، تعداد درخت‌ها، تعداد ویژگی‌هایی که هر درخت در نظر می‌گیرد و حداقل تعداد داده‌ها در گره‌های پایانی دارد.

۳ـ۳ ماشین تقویت گرادیان

در ماشین تقویت گرادیان، برای هر بیمار یک تابع $f(x)$ به دست می‌آوریم که قرار است مشابه کارکرد $\beta^{\intercal} x$ در مدل Cox-PH را داشته باشد. این مدل مقدار $f(x)$ را به کمک M مدل ساده‌تر که برای رگرسیون

[2] Logrank Test Statistics

به کار می‌روند مطابق با رابطه‌ی زیر تخمین می‌زند [۱۴، ۶]:

$$f(x) = \sum_{i=1}^{M} \beta_m g(x; \theta_m).$$

در این مدل، تابع هزینه همان تابع جزئی درستنمایی Cox است منتها با تابع f. به بیان بهتر، تابع هزینه به صورت زیر تعریف می‌شود:

$$l(f) = -\sum_{i \in D} \left[f(x_i) - \ln \sum_{j, t_j \geqslant t_i} \exp f(x_j) \right].$$

۳ـ۴ شبکه‌های عصبی

در این فصل مدل‌های یادگیری عمیق استفاده می‌شوند. سه مدل Deep Surv ، Logistic Hazard و PC Hazard [۷، ۱۱، ۱۲] در این پایان‌نامه بررسی شده‌اند.

روش Deep Surv، همان Cox-PH است که در آن قرار است یک شبکه‌ی عصبی با پارامتر θ تابع زمینه‌ای $h_\theta(x)$ را تخمین بزند. در حقیقت خروجی شبکه‌ی عصبی، مقدار زمینه‌ای تابع خطر باشد. بنابر این تعریف، تابع هزینه‌ی این مدل با در نظر گرفتن ضریب منظم‌ساز [۳] به صورت زیر است:

$$l(\theta) = \frac{1}{N_{\text{dead}}} - \sum_{i \in D} \left[h_\theta(x_i) - \ln \sum_{j, t_j \geqslant t_i} \exp h_\theta(x_j) \right] + \lambda \|\theta\|_2^2.$$

روش Logistic Hazard، تابع هزینه‌ای مطابق رابطه‌ی زیر دارد:

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^{n} \left(d_i \ln \left[h(t_i|x_i) \right] + (1 - d_i) \ln \left[1 - h(t_i|x_i) \right] + \sum_{j, t_j < t_i} \ln \left[1 - h(t_j|x_i) \right] \right)$$

که در رابطه‌ی فوق، d_i برای بیمار iاُم مشخص می‌کند که مرده است ($d_i = 1$) یا زنده است ($d_i = 0$). این شبکه‌ی عصبی سعی می‌کند تابع خطر (h) را به گونه‌ای تخمین بزند که هزینه کمینه شود. این روش، تابع خطر را به شکل گسسته تخمین می‌زند و صرفا در نقاط t_i، این تابع برای بیماران مختلف مقدار دارد.

روش PC Hazard مشابه روش Logistic Hazard است، منتها تابع خطر را به شکل پیوسته تخمین می‌زند و به خاطر این موضوع، تابع هزینه‌ی متفاوتی دارد [۱۵].

[3] Regularization

۳ـ۵ اهمیت ویژگی‌ها

یکی از کارهایی که در این در این پایان‌نامه قصد انجام آن را داشتیم، این بود که متوجه شویم کدام ویژگی‌ها اهمیت بیشتری دارند و نقش مهم‌تری در بقا ایفا می‌کنند.

به منظور استخراج این ویژگی‌ها، دو مدل جنگل تصادفی بقا و ماشین تقویت گرادیان، هر دو براساس ویژگی‌هایی درونیشان می‌توانند ضریبی برای اهمیت فیچرها استخراج کنند. در ادامه اما برای مدل Logistic Hazard از یک ایده برای بررسی اهمیت ویژگی‌ها و انتخاب آنها استفاده کردیم. ما دو حالت جستجوی کامل [۴] و جستجوی نیمه‌کامل [۵] را ارائه می‌دهیم که به کمک آنها می‌توان تا حدی در مورد ویژگی‌های مهم که بودنشان در آنالیز بقا اثرگذار است، اطلاعاتی کسب کنیم.

در روش جستجوی کامل، ما از ۲۲ ویژگی موجود، ۲ تا از ویژگی‌ها را کنار می‌گذاریم. سپس برای ۲۰ ویژگی باقی‌مانده از مدلمان استفاده کرده و با این ویژگی‌ها مدل را آموزش می‌دهیم. اگر به ازای همه ۲۰تایی‌ها از ویژگی‌ها، این آزمایش را انجام دهیم، ما تعداد $\binom{۲۲}{۲}$ آزمایش خواهیم داشت. مقدار C_{index} استخراج شده برای همه‌ی این آزمایش‌ها را بررسی می‌کنیم. اگر مقدار C_{index} روی داده‌ی تست برای یک ۲۰تایی از ویژگی‌ها مناسب باشد، یا به عبارت بهتر از حدی بالاتر باشد، می‌توان نتیجه گرفت که احتمالا آن ۲۰ ویژگی می‌توانستند شرایط بیمار را توصیف کنند و به اندازه‌ی کافی دانش در اختیار مدل قرار می‌داده‌اند. نتایج این بررسی و آنچه از آن به دست آمد در فصل بعدی قابل مشاهده است.

در روش جستجوی نیمه‌کامل، تعداد ۱۰۰۰ آزمایش که در هر یک از آن‌ها یک مجموعه‌ی ۱۵تایی تصادفی از ویژگی‌ها انتخاب شده و مدل Logisistic Hazard روی آن ۱۵ ویژگی آموزش می‌بیند بررسی شد. مجدداً آن مجموعه از ویژگی‌ها که بتواند معیار C_{index} قابل قبولی (از حدی بالاتر) داشته باشد، به این معناست که دربرگیرنده‌ی ویژگی‌های مهم هست. در مورد نتایج این بررسی و آن ویژگی‌هایی که با این بررسی مهم تلقی شده‌اند در فصل بعدی بحث شده است.

[4] Exhaustive Search
[5] Semi Exhaustive Search

۳ـ۶ جمع‌بندی

در این فصل به معرفی مدل‌های Cox-PH، جنگل بقای تصادفی، ماشین تقویت گرادیان، Logistic Hazard، Deep Surv و PC Hazard پرداختیم و شیوه‌ی کارکرد و توابع هزینه‌ی آن‌ها را معرفی کردیم. در انتها به اهمیت ویژگی‌ها و شیوه‌هایی که اهمیت ویژگی‌ها را در مدل‌هایمان تخمین می‌زنیم پرداختیم.

فصل ۴

نتایج

۴ـ۱ مقایسه‌ی مدل‌ها

در این فصل، نخست نتیجه مقایسه مدل‌ها را بیان می‌کنیم. برای آموزش و تست مدل‌ها، در ابتدا ٪۸۰ داده‌ها در دسته‌ی داده‌های آموزش و اعتبارسنجی و ٪۲۰ داده‌ها در دسته‌ی داده‌های تست قرار گرفته‌اند. تمامی مدل‌های ارائه شده در فصل قبل با روش اعتبارسنجی متقابل k-fold [1] آموزش دیده‌اند و بهترین مجموعه‌ی ابرپارامترها برای آن‌ها پیدا شده است.

نتیجه‌ی آموزش مدل‌ها روی مجموعه‌ی داده‌هایمان با بهترین ابرپارامترها در جدول ۴ـ۱ قابل مشاهده است. همانطور که مشاهده می‌شود، مدل Logistic Hazard، بهترین نتیجه را در میان تمامی مدل‌ها دارد و بالاترین دقت روی داده‌ی تست را در معیار C_{index} به خود اختصاص می‌دهد. البته مدل جنگل بقای تصادفی هم دقت قابل توجهی دارد. مدل کلاسیک Cox-PH هم همانطور که انتظار می‌رفت، دقت نسبتاً پایین‌تری نسبت به باقی مدل‌ها دارد و سادگی مدل سبب می‌شود که نتواند دقت قابل قبولی روی داده‌ی آموزش یا تست به دست بیاورد. در میان مدل‌های مبتنی بر شبکه‌ی عصبی که عموماً در حجم بالای داده خیلی خوب عمل می‌کنند، مدل Logisitc Hazard بهتر از بقیه عمل می‌کند و دقت به مراتب بهتری دارد. به همین جهت ما نیز از این مدل برای انتخاب ویژگی‌ها و تعیین اهمیت آن‌ها استفاده می‌کنیم.

همانطور که بیان شد، این مدل‌ها می‌توانستند تابع خطر و بعد از روی آن تابع بقا را پیش‌بینی کنند.

[1] k-fold cross validation

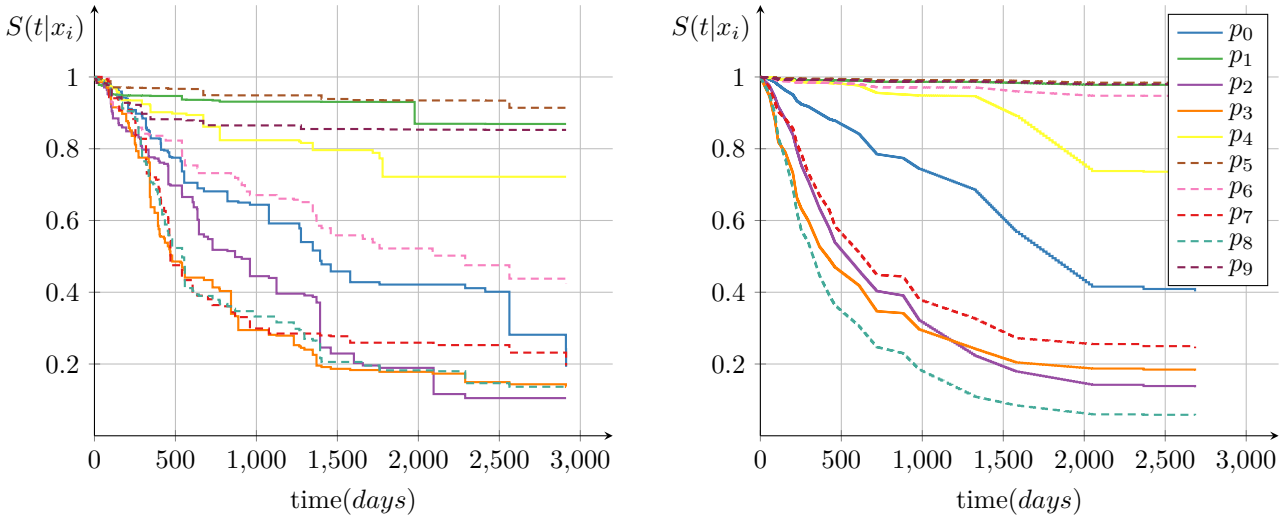
جدول ۴ـ۱: کارایی مدل‌ها روی داده‌های بیماران. امتیاز تجمیعی Berier در یک بازه‌ی ۵ ساله محاسبه شده است. همه‌ی اعداد با دو رقم اعشار گرد شده‌اند. در مدل‌های شبکه‌ی عصبی، در خط دوم، دقت روی داده‌ی اعتبارسنجی مشاهده می‌شود.

Model	Train		Test	
	C-index	IBS	C-index	IBS
RSF	0.89	0.06	**0.84**	0.11
GBS	0.92	0.05	0.82	0.11
Cox PH	0.79	0.12	0.78	0.15
Logistic Hazard	0.89 0.85	0.07 0.10	**0.87**	0.10
DeepSurv	0.93 0.80	0.04 0.11	0.82	0.10
PC Hazard	0.81 0.77	0.10 0.13	0.78	0.12

تابع بقای پیش‌بینی شده برای ۱۰ مورد از بیماران توسط دو مدل جنگل بقای تصادفی و Logistic Hazard در شکل ۴ـ۱ قابل مشاهده است. همانطور که مشاهده می‌شود، در هر دو این توابع، نفرات شماره‌ی ۲، ۳، ۷ و ۸ نفراتی هستند که در خطر بوده و تابع بقا برای آن‌ها به سرعت کاهش می‌یابد. در سمت مقابل، بیماران با شماره‌های ۱ و ۵ وجود دارند که به مراتب مطابق با پیش‌بینی هر دو مدل کمتر در خطر بوده و احتمالاً طولانی‌تر زنده خواهند ماند.

۴ـ۲ اهمیت ویژگی‌ها

دو مدل جنگل بقای تصادفی و ماشین تقویت گرادیان، هر دو به خاطر ساختاری که دارند می‌توانند اهمیت ویژگی‌ها را به دست بیاورند. در مدل جنگل بقا، برای تعیین اهمیت یک ویژگی، هر جا داخل درخت‌های تصمیم، بچه‌های یک گره بر اساس آن ویژگی جدا شده‌اند، مرز جداشدن به شکل تصادفی



(a) Random Survival Forest

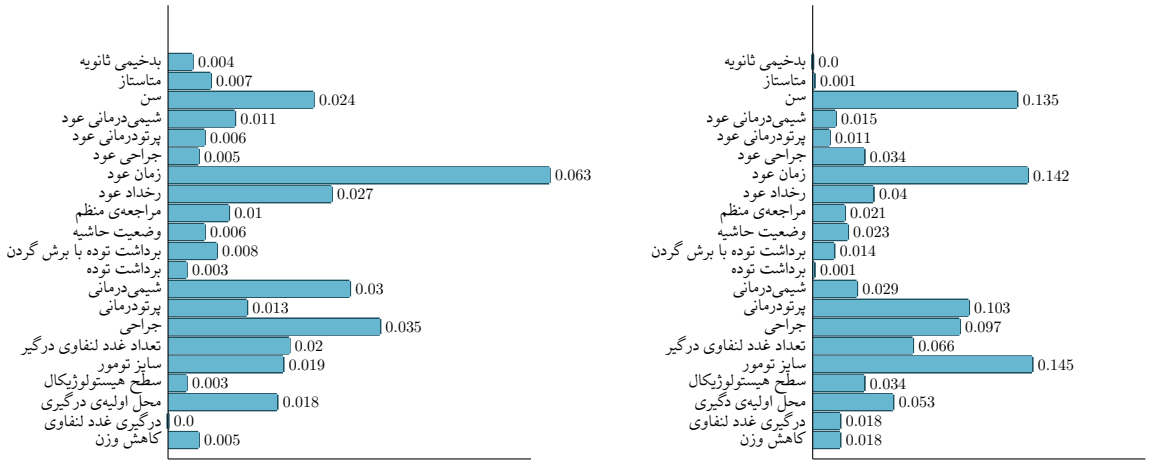
(b) Logistic Hazard

شکل ۴ـ۱: تابع بقا؛ پیش‌بینی شده توسط دو مدل جنگل بقای تصادفی و Logistic Hazard. این دو نمودار، تابع $S(t|x)$ را برای ۱۰ بیمار نشان می‌دهند. بیماران پر خطر و کم‌خطر در هر دو نمودار تقریباً یکسان هستند.

تعیین می‌شود و نهایتاً میزان خطای پیش‌بینی [۲] سنجیده می‌شود. هر چه خطا بیشتر شود، یعنی ویژگی پراهمیت‌تر بوده است. مدل ماشین تقویت گرادیان هم به شکل مشابه چنین قابلیتی را دارد. اهمیت ویژگی‌ها در شکل ۴ـ۲ قابل مشاهده است. این شکل به ما نشان می‌دهد که ویژگی‌هایی از قبیل سن، زمان عود، سایز تومور و جراحی ویژگی‌های پر اهمیت در هر دو مدل‌ها هستند. همچنین پرتودرمانی و شیمی‌درمانی ویژگی‌هایی هستند که به ترتیب در ماشین تقویت گرادیان و جنگل بقا مهم بوده‌اند.

در مورد مدل Logistic Hazard، همانطور که در فصل پیشین مطرح شد، مجموعه آزمایشات جستجوی کامل و جستجوی نیمه‌کامل انجام شد. در مورد جستجوی نیمه‌کامل، هر مجموعه‌ی ۱۵ تایی از ویژگی‌ها که بتواند دقت C_{index} حداقل ۰/۸۳۵ داشته باشد را در نظر می‌گیریم و آن‌ها را مجموعه‌های **امیدبخش** می‌نامیم. شکل آـ۱ توزیع مقدار C_{index} در مجموعه‌های تصادفی مورد آزمایش را نشان می‌دهد. سپس برای هر ویژگی، حساب می‌کنیم که در چه کسری از این مجموعه‌های امیدبخش وجود دارد. طبیعتاً هر چه این کسر برای یک ویژگی بزرگتر باشد، به نوعی نشان می‌دهد که این ویژگی مهم است و اگر در مجموعه‌ی داده‌ها ظاهر شود، می‌تواند سبب تقویت دقت مدل شود و برعکس اگر ظاهر نشود، مدل با احتمال کمی می‌تواند دقت قابل قبولی داشته باشد. پس بدین ترتیب می‌توانیم تخمینی

[2] Prediction Error



(a) Random Survival Forest

(b) Gradient Boosting Machine

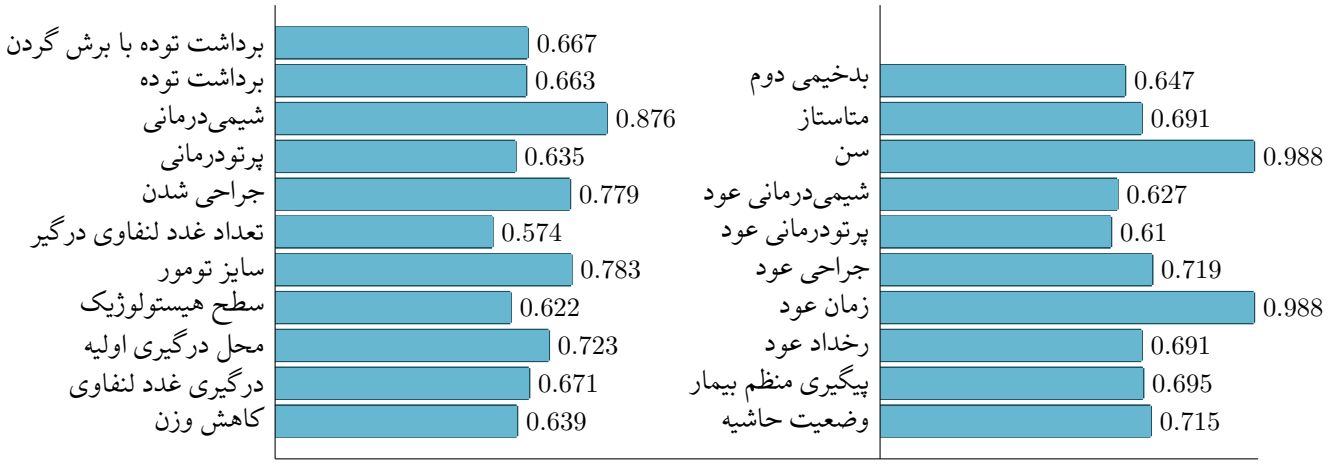
شکل ۴ـ۲: اهمیت ویژگی‌ها در مدل‌های جنگل بقا و ماشین تقویت گرادیان

بر اهمیت ویژگی‌ها داشته باشیم. شکل ۴ـ۳ این کسر را برای ویژگی‌های مختلف نشان می‌دهد. در آزمایشات ما و در مدل Logistic Hazard، به نظر می‌رسد سن، زمان عود، شیمی درمانی، سایز تومور و جراحی ویژگی‌های تاثیرگذار در پیش‌بینی بقا هستند.

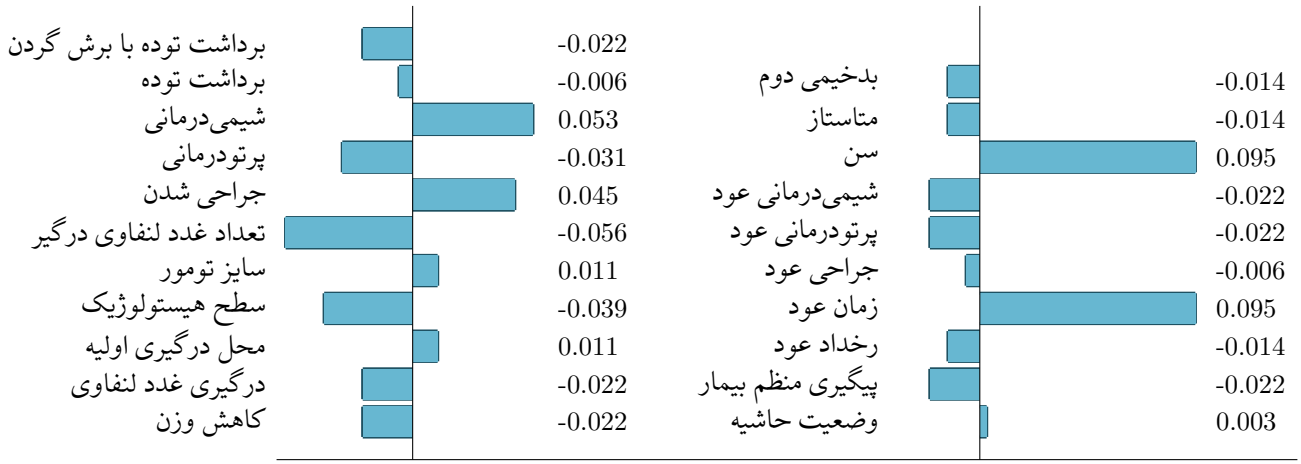
در مورد جستجوی کامل، کاری دقیقاً مشابه بالا انجام شد. توزیع C_{index} در مجموعه‌ی آزمایشات مطابق با شکل آـ۱ است. مجدداً برای هر ویژگی احتمال اینکه در یک مجموعه‌ی **امیدبخش** (در اینجا مجموعه‌ی امیدبخش دقت حداقل ۰/۸۳ در C_{index} دارد) باشد را حساب می‌کنیم. نتیجه محاسبه‌ی این کسر در شکل۴ـ۴ مشاهده می‌شود. نتایج این آزمایش هم تقریباً سازگار با آزمایشات قبلی است و می‌توان متوجه شد که مثلاً ویژگی‌هایی مثل سن، زمان عود، شیمی‌درمانی و جراحی در آنالیز بقا در مورد بیماران ما حائز اهمیت هستند.

۴ـ۳ جمع‌بندی

در این فصل مدل‌هایی که در فصل قبل معرفی کردیم را روی داده‌هایمان بررسی کردیم و به مقایسه آن‌ها پرداختیم. همچنین اهمیت ویژگی‌ها را مورد بررسی قرار دادیم و برای سه تا از مدل‌ها، ویژگی‌هایی که اهمیت بیشتری در پیش‌بینی بقا دارند را شناسایی کردیم.



شکل ۴ـ۳: احتمال اینکه یک ویژگی در یک مجموعه‌ی امیدبخش باشد؛ نتیجه‌ی آزمایش روی ۱۰۰۰ مجموعه‌ی ۱۵ عضوی از ویژگی‌ها. هر چه این احتمال بزرگتر باشد، یعنی ویژگی مهم‌تر است و با نبودنش، احتمالاً مجموعه‌ی ویژگی‌ها امیدبخش نیستند.



شکل ۴ـ۴: احتمال اینکه یک ویژگی در یک مجموعه‌ی امیدبخش باشد؛ آزمایش روی ۲۱۰ مجموعه‌ی ۱۹ عضوی از ویژگی‌ها. مقادیر جدول بالا منهای $\frac{۱۹}{۲۱}$ شده‌اند چون هر ویژگی‌ای در $\frac{۱۹}{۲۱}$ نمونه‌ها وجود دارد. هر چه این احتمال بزرگتر باشد، یعنی ویژگی مهم‌تر است و با نبودنش، احتمالاً مجموعه‌ی ویژگی‌ها امیدبخش نیستند.

فصل ۵

نتیجه‌گیری

در این پایان‌نامه ما به روش‌های حل مسأله‌ی آنالیز بقا پرداختیم و دسته‌بندی کلی‌ای از شیوه‌های کلاسیک و مبتنی بر یادگیری ماشین ارائه دادیم. سپس تعدادی از مهم‌ترین شیوه‌ها را روی مجموعه داده‌هایمان بررسی کردیم و به مقایسه آنها با یکدیگر براساس معیارهای موجود پرداختیم. بنابر نتایج آزمایشات، مدل‌های Logistic Hazard و جنگل بقای تصادفی بیشترین دقت و بازدهی را داشتند و بهتر از بقیه‌ی مدل‌های موجود عمل کردند.

همچنین در ادامه‌ی کار سعی کردیم که ویژگی‌های مهم را شناسایی کنیم. برای این کار از ویژگی‌های جنگل بقا و ماشین تقویت گرادیان استفاده کردیم و نیز شیوه‌ای برای بررسی اهمیت ویژگی‌ها برای مدل Logistic Hazard ارائه دادیم. بر مبنای آزمایش‌هایمان، متوجه شدیم که ویژگی‌های همچون سن، زمان عود، انجام جراحی، انجام شیمی‌درمانی، سایز تومور و پرتودرمانی از مجموعه ویژگی‌هایی هستند که در آنالیز بقا در مورد بیماران سرطان دهان حائز اهمیت‌اند و نبود آن‌ها در مجموعه‌ی ویژگی‌ها می‌تواند باعث شود که مدل‌های بالا به دقت پایینی برسند.

مدل‌های فراوان دیگری نیز در زمینه‌ی آنالیز بقا وجود دارند که می‌توانستند در این پروژه بررسی شوند. همچنین امروزه مدل‌های بسیار متفاوتی مبتنی بر شبکه‌های عصبی ارائه شده‌اند که آن‌ها نیز می‌توانستند بررسی شوند و ما تنها به ۳ مورد از آن‌ها پرداختیم. همچنین در خصوص تعیین اهمیت ویژگی‌ها، می‌توان از ایده‌هایی که مستقیماً به ساختار شبکه‌های عصبی مربوط به این مدل‌ها وابسته هستند (مثل گرفتن مشتق نیست به ویژگی‌ها و ...) نیز استفاده کرد و بیشتر در مورد اهمیت ویژگی‌ها اطلاعات به دست آورد.

پیوست آ

مطالب تکمیلی

آـ۱ ابرپارامترهای مدل‌ها

برای تست و بررسی مدل‌ها از کتابخانه scikit survival و pycox استفاده شد. این کتابخانه‌ها مبتنی بر مقالات موجود، به پیاده‌سازی مدل‌های ارائه شده پرداخته‌اند و تست و بررسی آن‌ها روی داده‌ها به کمک این کتاب‌خانه‌ها ساده‌تر است.

در خصوص ابرپارامترها، برای هر یک از مدل‌ها از مجموعه‌ی ابرپارامترهای زیر استفاده شد. برای یافتن بهترین ابرپارامترها، از جستجوی همگانی و Grid Search استفاده گردید.

- مدل جنگل بقا (`sksurv.ensemble.RandomSurvivalForest`):

```
max_features: 'sqrt', min_samples_leaf: 3, min_samples_split: 3, n_estimators: 50
```

- مدل تقویت گرادیان (`sksurv.ensemble.GradientBoostingSurvivalAnalysis`):

```
learning_rate: 0.05, loss: 'coxph', min_samples_leaf: 8, min_samples_split: 2, n_estimators: 1000
```

- مدل Cox-PH (`sksurv.linear_model.CoxPHSurvivalAnalysis`):

```
alpha: 0, n_iter: 50, ties: 'efron', tol: 1e-09
```

- مدل Deep Surv (pycox.models.CoxPH):

```
network structure: [64, 64, 64]
batch_norm: True
dropout: 0.6
initial_lr: 0.01
```

- مدل Logistic Hazard (pycox.models.LogisticHazard):

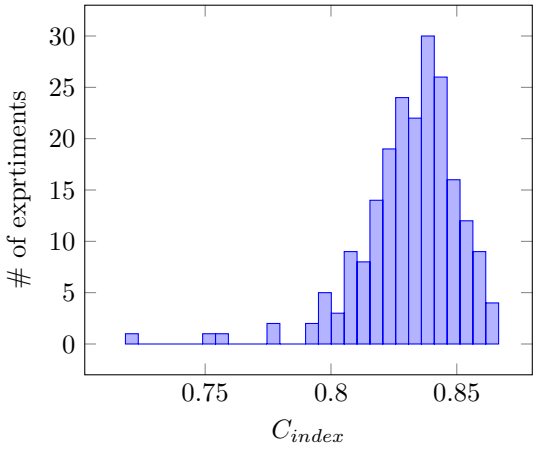
```
num_nodes: [64, 128, 128, 64]
batch_norm: True
dropout: 0.65
initial_lr: 0.05
optimizer: AdamWR
```

- مدل PC Hazard (pycox.models.PCHazard):

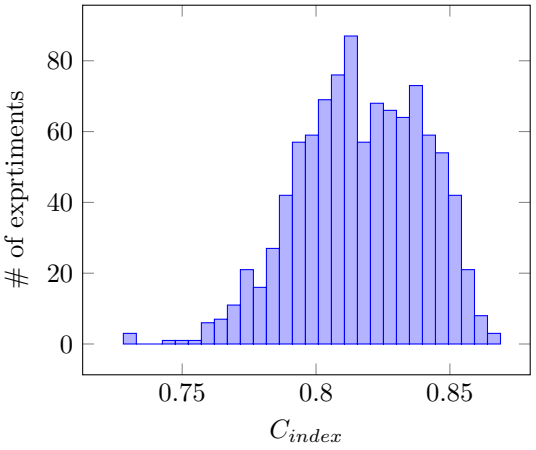
```
num_nodes: [16, 32, 16]
batch_norm: True
dropout: 0.8
initial_lr: 0.05
optimizer: Adam
```

آ‌ـ۲ توزیع C_{index} در آزمایش‌های جستجوی کامل و نیمه‌کامل

در شکل آ‌ـ۱ می‌توان توزیع معیار C_{index} را در آزمایشات مربوط به انتخاب ویژگی‌ها مشاهده کرد. همانطور که مشاهده می‌شود، مقدار C_{index} در محدوده‌ی [۰/۷, ۰/۹] می‌ماند که با توجه به نتایج آزمایش‌هایمان، بازه‌ای قابل قبول است. ضمنا همین توزیع‌ها نشان می‌دهند که می‌توان تعدادی زیادی از ویژگی‌ها را حذف کرد و همچنان دقتی در حد دقت گزارش شده در فصل‌های پیشین که با در نظر گرفتن همه‌ی ویژگی‌ها بود داشت. این موضوع نشان می‌دهد که تعدادی از ویژگی‌های ما عملاً دانش زیادی اضافه نمی‌کنند و نبودنشان اثری ندارد.



(a) Exhaustive Search



(b) Semi-Exhaustive Search

شکل آ‌ـ۱: توزیع C_{index} در مجموعه آزمایشات جستجوی کامل و جستجوی نیمه‌کامل. شکل سمت چپ مربوط به ۲۱۰ آزمایش جستجوی کامل؛ هنگامی که مدل Logistic Hazard روی مجموعه‌ی ۱۹ عضوی از ویژگی‌ها آموزش ببیند. شکل سمت راست مربوط به ۱۰۰۰ آزمایش جستجوی نیمه‌کامل؛ هنگامی که مدل Logistic Hazard روی مجموعه‌ی ۱۵ عضوی از ویژگی‌ها آموزش می‌بیند.

در آزمایش جستجوی کامل، حد پایین ۰/۸۳ برای انتخاب مجموعه‌های امیدبخش در نظر گرفته شده است. اما در آزمایش جستجوی نیمه‌کامل حد ۰/۸۳۵ در نظر گرفته شده است. احتمال موجود در تصاویر برای هر ویژگی، با تقسیم تعداد مجموعه‌هایی امیدبخشی که ویژگی مربوطه را دارند بر تعداد کل مجموعه‌های امیدبخش به دست آمده است.

مراجع

[1] L. Vale-Silva and K. Rohr. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1), 2021.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA*, 71(3):209–249, 2021.

[3] S. Warnakulasuriya. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol*, 45(4-5):309–16, 2009.

[4] D. Kim, S. Lee, S. Kwon, W. Nam, I. Cha, and H. Kim. Deep learning-based survival prediction of oral cancer patients. *Scientific Reports*, 9(1), 2019.

[5] R. Sonabend. A theoretical and methodological framework for machine learning in survival analysis. 2021.

[6] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistical Surveys*, 5:44–71, 2011.

[7] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18, 2018.

[8] J. Wang, N. Chen, J. Guo, X. Xu, L. Liu, and Z. Yi. Survnet: A novel deep neural network for lung cancer survival analysis with missing values. *Front Oncol*, 10, 2021.

[9] A. Spooner and E. Chen. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10, 2020.

[10] E. Giunchiglia, A. Nemchenko, and M. van der Schaar. Rnn-surv: A deep recurrent model for survival analysis. *ICANN*, 2018.

[11] H. Kvamme, B. Ornulf, and I. Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20:1–30, 2019.

[12] Q. Zhong, J. W. Mueller, and J.-L. Wang. Deep extended hazard models for survival analysis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15111–15124. Curran Associates, Inc., 2021.

[13] H. Li, P. Boimel, J. J. Naylor, H. Zhong, Y. Xiao, E. Ben-Josef, and Y. Fan. Deep convolutional neural networks for imaging data-based survival analysis of rectal cancer. *Proc IEEE Int Symp Biomed Imaging*, pages 846–849, 2019.

[14] Wang, Ping, and Y. Li. Machine learning for survival analysis: A survey. *ACM Comput. Surv.*, 51(6), feb 2019.

[15] H. Kvamme and B. Ornulf. Continuous and discrete-time survival prediction with neural networks. 2019.

واژه‌نامه

الف

اهمیت ..... importance

آموزش ..... train

اعتبارسنجی ..... validation

اعتبارسنجی متقابل ..... cross validation

امتیاز ..... score

آنالیز بقا ..... survival analysis

ب

بقا ..... survival

بسته‌بندی ..... bagging

بیشینه درست‌نمایی ..... maximum likelihood

برنامه‌ریزی خطی ..... linear programming

برنامه‌ریزی صحیح ..... integer programming

برداشت توده ..... resection

ت

تبدیل تصویری ..... projective transformation

تابع هزینه ..... loss function

تست ..... test

تجمیعی ..... integrated, cumulative

ج

چنگل تصادفی ..... random forest

جنگل بقای تصادفی ..... random survival forest

جستجوی کامل ..... exhaustive search

جستجوی نیمه کامل ..... semi-exhaustive search

جراحی ..... surgery

خ

خطر ..... hazard

د

درخت تصمیم ..... decision tree

دیسکشن گردن ..... neck dissection

ر

رادیوتراپی radiation therapy

س

سطح level

سن age

ش

شیمی درمانی chemotherapy

شبکه‌ی عصبی neural network

ع

عود recurrence

غ

غده لنفاوی گردنی cervical lymph nodes

ک

کاهش وزن loss weight

م

ماشین تقویت گرادیان gradient boosting machine

مجموعه set

معیار criterion

مراجعه‌ی مجدد follow up

متاستاز metastasis

و

ویژگی feature

Abstract

Predicting the severeness of the disease of a patient could have a significant impact on the methods and treatment plans that should be done to cure a patient. Data gathered from clinical treatments and other studies can be used in order to use for the purpose of survival analysis. These data are usually high-dimensional, censored and contain missing information which brings about the need for methods which can overcome these challenges. Machine learning techniques because of their potential to estimate complex functions and overcome those challenges have recently received remarkable attention. In this paper, we run experiments and use different classes of machine learning models that have been developed for this purpose. We discuss about these classes and compare some instances of them. Also, we analyze the importance of features in our analysis.

Keywords: Machine Learning, Survival Analysis, Neural Networks, Feature Selection

Sharif University of Technology

Department of Computer Engineering

B.Sc. Thesis

Survival Analysis for Oral Cancer Patients

By:

Keivan Rezaei

Supervisor:

Dr. Mahdieh Soleymani

June 2022