پردیس علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

تحلیل بقا براساس گراف بدون‌دور باسو

نگارنده

فاطمه فرتاش اصل

استاد راهنما: زهرا رضایی قهرودی

پایان‌نامه برای دریافت درجه کارشناسی
در رشته آمار

تابستان ۱۴۰۲

چکیده

در نظر گرفتن ساختارهای علّی برای داده‌های مشاهده‌ای بقا، اطلاعات مهمی در مورد روابط بین متغیرهای کمکی و متغیر زمان تا رویداد ارائه می‌دهد. در این پروژه نشان می‌دهیم با بهره‌گیری از نظریه اطلاع کدگذاری (رمزگذاری) منبع، در صورت استفاده از کدگذارهای مناسب و ترکیب دانش گراف غیرمدور جهت‌دار(DAG)، می‌توان به نتایج و تحلیل‌های مناسبی دست یافت. در این پروژه به استنباط های تغییراتی براساس خودرمزگذار تغییراتی شرطی برای پیش‌بینی بقای ساختاریافته علّی می‌پردازیم که به آن DAGSurv گفته می‌شود. همچنین عملکرد DAGSurv را روی مجموعه داده‌های ساختگی بُعدپایین و بُعدبالا و مجموعه داده‌های واقعی مانند METABRIC و GBSG نشان می‌دهیم. روش معرفی شده در این پروژه از سایر روش‌های تحلیل بقا مانند روش خطرهای متناسب کاکس[1] DeepaSurv، و Deephit که نسبت به روابط علی بین موجودیت‌های داده‌ها غافل هستند، بهتر عمل می‌کند.

[1] Cox Proportional Hazards

سپاسگزاری

با تشکر از استاد محترم دکتر زهرا رضایی قهرودی که مرا در این پروژه یاری کردند.

فهرست مطالب

فصل ۱

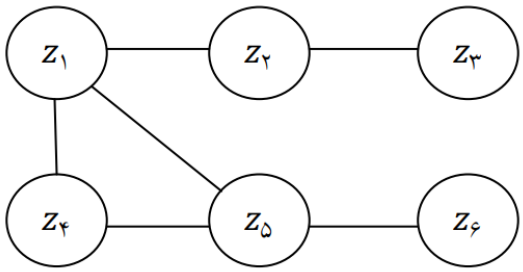
مفاهیم مقدماتی

در دهه‌های اخیر مدل‌های گرافیکی[۱] به عنوان یک ابزار مناسب و کارا در نمایش و مدل‌بندی توزیع‌های چندمتغیره بر اساس استقلال‌شرطی[۲] متغیرها کاربرد گسترده‌ای پیدا کرده است. این مدل‌ها به ما این امکان را می‌دهد که روابط پیچیده بین متغیرهارا با استفاده از گراف بیان کنیم. به همین دلیل ابتدا به معرفی مدل‌های گرافیکی و ویژگی‌های آن [۱] و سپس به بیان مقدماتی درباره تحلیل داده‌های بقا [۲] می‌پردازیم.

۱.۱ گراف

گراف $G(V, E)$ متشکل از مجموعه رئوس(گره‌ها)[۳] V و مجموعه یال‌ها[۴] E می‌باشد. در این مدل‌ها هر راس بیانگر یک متغیرتصادفی و هر یال بیانگر وابستگی و رابطه بین آن دو متغیر می‌باشد. برای مثال در یک گراف با دو راس x_i و x_j یال بین این دو راس را با u_{ij} نمایش می‌دهیم که x_i راس آغازین و x_j راس پایانی می‌باشد. چنانچه x_j راس آغازی و x_i راس پایانی باشد نمایش یال آن به صورت u_{ji} خواهدبود.

[۱] Graphical models
[۲] Conditional Independence
[۳] Vertices
[۴] Arcs



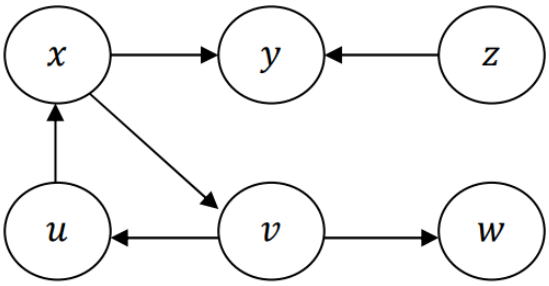
شکل ۱.۱: گرافی با ۶ گره و ۷ یال

۱.۱.۱ تعاریف

گراف جهت‌دار[۵]: گرافی که یال‌های آن جهت‌دار باشند را گراف جهت‌دار می‌گوییم.
گراف‌غیرجهت دار: گرافی که یال‌های آن جهت‌دار نباشند را گراف غیر جهت‌دار می‌گوییم.
زیرگراف: هر زیرمجموعه از V (رئوس) و E (یال‌ها) به عنوان زیرگراف‌های G می‌توانند در نظر گرفته شوند به عبارتی دیگر گراف $\acute{G}(\acute{V},\acute{E})$ را زیرگراف $G\,(V,E)$ می‌گوییم اگر داشته باشیم: $\acute{V} \subseteq V$ و $\acute{E} \subseteq E$
گراف کامل: گراف را کامل می‌گوییم اگر همه راس‌های آن دو به دو به هم متصل باشند.
مسیر[۶]: هر مجموعه از یال‌های پشت سرهم که دو رأس را به هم متصل می‌کنند مسیر نامیده می‌شود؛ در صورتی که یک مسیر از x_i به x_j وجود داشته باشد می‌گوییم راس x_j برای x_i دست‌یافتنی است.
به مجموعه همه راس‌هایی که برای راس x_iدست‌یافتنی است، مجموعه نوادگان [۷] آن راس می‌گوییم و با$des(x_i)$ نشان می‌دهیم؛ همچنین به مجموعه راس‌هایی که راس x_j برای آن‌ها دست‌یافتنی است مجموعه نیاکان [۸] آن راس می‌گوییم و با $an(x_j)$ نشان می‌دهیم.
طول مسیر: تعداد یال‌های موجود در مسیر را طول مسیر می‌گویند.
دور [۹]: یک دور مسیر ساده‌ای است که راس شروع و پایانی آن یکی باشد.
گراف غیرمدور[۱۰]: گراف ساده جهت‌داری را که دارای دور نیست، غیرمدور می‌نامند.
رئوس مجاور یا همسایه: دو راس از یک گراف که به وسیله یک یال به هم متصل شده باشد را دو راس مجاور (یا همسایه) می‌گویند.

---

[۵] Directed Graph
[۶] Path
[۷] Descendant
[۸] Ancestor
[۹] Cycle
[۱۰] Acyclic Graph



شکل ۲.۱: گراف مدور

ماتریس مجاورت: برای گراف G با n راس، ماتریس A با درایه‌های a_{ij} که به صورت زیر تعریف می‌شود را ماتریس مجاورت می‌گوییم. ماتریس مجاورت برای یک گراف با n راس یک ماتریس $n \times n$ می‌باشد.

$$a_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{if } (v_i, v_j) \notin E \end{cases}$$

که در آن v_i و v_j راس‌های گراف و منظور از (v_i, v_j) یالی است که ابتدای آن v_i و انتهای آن v_j می‌باشد.

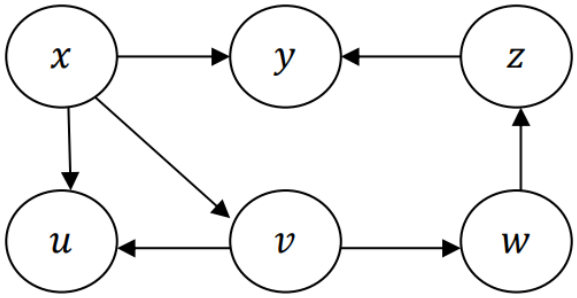
گراف غیرمدور جهت‌دار(DAG)[۱۱]: گراف غیرمدور جهت‌دار گرافی است که همه یال‌های آن جهت‌دار باشد و هیچ مسیر بسته‌ای در آن وجود نداشته باشد. در این گراف با آغاز از یک راس و حرکت در جهت کمان‌ها نمی‌توانیم به راس آغازین برگردیم. در یک DAG برای دو راس x و y اگر $(x, y) \in E$ باشد از نماد $x \to y$ استفاده می‌کنیم و می‌گوییم x والد y و y فرزند x است. یک راس ممکن است به طور همزمان چند والد و فرزند داشته باشد، مجموعه والدهای راس x را با $pa(x)$ نمایش می‌دهیم. اگر مجموعه همه همسایه‌های راس x را با $adj(x)$ نشان دهیم داریم $pa(x) \subseteq adj(x)$

گراف شکل ۲.۱ باتوجه به مسیر $\{x, v, x, u, x\}$ نمی‌تواند یک گراف غیرمدور جهت‌دار یا DAG باشد.

در شکل ۳.۱ DAG نشان داده شده است، که در آن x هیچ والدی ندارد؛ $pa(x) = \emptyset$ پس x یک ریشه است. برای راس u داریم $pa(u) = \{x, v\}$ در این گراف همه راس‌ها برای x دست‌یافتنی می‌باشند و $adj(v) = \{x, u, w\}$

مدل‌های گرافیکی: مدل گرافیکی عبارت است از یک گراف که درآن هر راس نمایانگر یک متغیر تصادفی است و در صورت وجود استقلال کامل یا شرطی بین دو متغیر، هیچ یالی بین دو راس وجود ندارد. استقلال بین دو متغیر x و y را به‌صورت $x \perp y$ نشان می‌دهیم.

[۱۱] Directed Acyclic Graph



شکل ۱.۳: گراف غیرمدور جهت‌دار

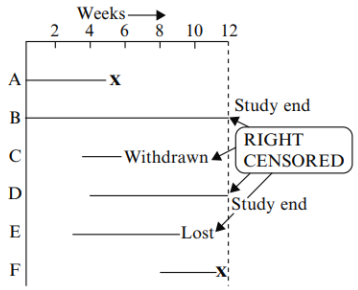
۲.۱ تحلیل بقا

به طور کلی تحلیل بقا مجموعه‌ای از روش‌های آماری برای تحلیل داده‌هایی است که متغیر پاسخ در آن‌ها زمان تا رخداد یک پیشامد خاص است. منظور از زمان در تحلیل بقا می‌تواند تعداد سال‌ها، ماه‌ها، هفته‌ها یا روزها از شروع پیگیری یک فرد تا رخداد پیشامد مورد نظر برای وی باشد.
پیشامد در تحلیل بقا ممکن است مرگ، وقوع بیماری، عود بیماری پس از بهبودی، بازیافتن توانایی و بهبودی (مثلا بازگشت به کار)، یا هر تجربه مورد نظر دیگری که ممکن است برای فرد رخ دهد، باشد.
زمان بقا[۱۲]: در تحلیل بقا معمولا متغیر زمان را زمان بقا می‌نامیم، زیرا این متغیر تعیین‌کننده مدت زمانی است که یک فرد در طول دوره پیگیری بقا یافته است. همچنین چون معمولا در این نوع تحلیل‌ها، پیشامدهای مورد نظر مرگ، وقوع بیماری یا سایر تجربه های فردی است، پیشامد مورد نظر را شکست[۱۳] می‌نامیم. در حالیکه زمان بقا ممکن است زمان تا برگشت به کار پس از یک عمل جراحی باشد، که در این صورت شکست یک پیشامد مثبت خواهد بود.

۱.۲.۱ داده‌های سانسورشده

بسیاری از تحلیل‌های بقا با یک مشکل اساسی به نام سانسور مواجه هستند. سانسور زمانی رخ می‌دهد که ما بخشی از اطلاعات مربوط به زمان بقای فردی را در اختیار داریم اما زمان بقای دقیق او را نمی‌دانیم برای مثال، داده‌های پیگیری بیماران سرطان خون را در نظر بگیرید؛ پیشامد مورد نظر در این مطالعه، خارج شدن از حالت بهبودی است که آن را با X نشان می‌دهیم. اگر برای یک

[۱۲] Survival Time
[۱۳] failure



شکل ۴.۱: مشاهدات B،C،D،F سانسور شده است

بیمار مشخص، مطالعه در حالی به پایان برسد که فرد هنوز در حالت بهبودی باشد(پیشامد برای وی رخ نداده) در این صورت زمان بقای این بیمار به عنوان سانسور در نظر گرفته می‌شود تنها اطلاعی که ما از زمان بقای این بیمار داریم این است که زمان بقای فرد از مدت زمانی که فرد مورد پیگیری قرار گرفته بیشتر می‌باشد. بنابراین اگر وی بعد از پایان مطالعه از حالت بهبودی خارج شود، هیچ اطلاعی از زمان بقای دقیق او نخواهیم داشت.
بطور کلی سه دلیل برای اینکه سانسور رخ دهد وجود دارد:

۱. یک فرد قبل از پایان مطالعه، پیشامد را تجربه نکرده باشد.

۲. یک فرد در طول مدت مطالعه گمشده یا از دست رفته باشد.

۳. یک فرد به دلیل مرگ (اگر پیشامد مورد بررسی مرگ نباشد) یا دلایل دیگر از مطالعه خارج شده باشد.

سه حالت بالا به صورت گرافیکی در شکل ۴.۱ نشان داده شده‌اند. در این نمودار X نشان دهنده فردی است که پیشامد را در طول دوره پیگیری تجربه کرده است. برای مثال فرد A تا هفته پنجم که پیشامد برای آن رخ داده مورد پیگیری قرار گرفته است. این فرد سانسور نشده و زمان بقا وی پنج هفته ثبت می‌شود. فرد B نیز از شروع مطالعه تا پایان مطالعه (که ۱۲ هفته است) مورد پیگیری قرار گرفته، بدون آنکه پیشامدی در این مدت برای وی رخ داده باشد، در این حالت چون تنها می‌توانیم بگوییم زمان بقای وی بیشتر از ۱۲ هفته است، زمان بقای آن سانسور شده می‌باشد. فرد C بین دومین و سومین هفته وارد مطالعه شده و تا هفته ششم که از مطالعه خارج می‌شود، مورد پیگیری قرار گرفته است. زمان بقای این فرد هم بعد از ۳.۵ هفته سانسور شده است. فرد F در هفته هشتم وارد مطالعه شده و تا هفته ۱۱.۵ که پیشامد برای آن رخ داده، مورد پیگیری قرار گرفته است. برای این فرد نیز سانسور نداریم و زمان بقای آن ۳.۵ هفته ثبت می‌شود. به طور خلاصه داریم که از ۲ نفر ۶ نفر پیشامد را تجربه کرده اند (A،F) و ۴ نفر هم سانسور شده‌اند .(D،C،B،E)

در مثال بالا تنها اطلاعاتی که از زمان بقای چهار فرد سانسور شده داریم، این است که زمان بقای واقعی آنها از طول دوره پیگیری بیشتر است و زمان بقا در سمت راست دوره پیگیری ناقص شده است. در این حالت داده‌ها را سانسور شده از راست می‌گوییم. اگرچه داده‌ها می‌توانند از چپ نیز سانسور شوند اما بیشتر داده‌های بقا، سانسور شده از راست هستند. سانسورهای چپ زمانی رخ می‌دهند که زمان بقای واقعی شخص، کمتر یا مساوی زمان بقا مشاهده شده‌ی آن باشد. برای مثال، اگر یک فرد را تا ابتلا به HIV مثبت پیگیری کنیم، این امکان وجود دارد که شکست را زمانی ثبت کنیم، که اولین تست برای ویرروس HIV مثبت شود. در حالی که ممکن است زمان دقیق مواجهه فرد با ویروس HIV را ندانیم. بنابراین به درستی نمی‌توانیم بگوییم که شکست چه زمانی رخ داده است. از آنجا که زمان بقای واقعی (مدت زمان تا رخداد مواجهه) نسبت به پایان زمان پیگیری (مثبت شدن تست فرد) کوتاه‌تر می‌باشد، زمان بقای این فرد را سانسور شده از چپ می‌نامیم.[۴]

۲.۲.۱ اصطلاحات و نمادها

در این بخش به معرفی بعضی اصطلاحات و نمادهای ریاضی که در تحلیل داده های بقا بکار برده میشوند، میپردازیم:
T : نشان دهنده‌ی متغیر تصادفی زمان بقا، برای یک فرد می‌باشد و از آنجا که T نشان دهنده‌ی زمان است تمام اعداد غیرمنفی را شامل می‌شود
d : یک متغیر ۲ حالتی است و از آن برای نشان دادن وضعیت رخ داد پیشامد (شکست) و یا سانسور شدن فرد استفاده می‌کنیم. $d = 1$ برای زمانی است که در طول مطالعه، پیشامد مورد نظر برای فرد رخ داده، $d = 0$ برای حالتی است که به یکی از دلایل گفته شده، زمان بقا برای فرد سانسور شده باشد.
دو عبارت کمی که در تمام تحلیل‌های بقا مورد توجه قرار می‌گیرند، تابع بقا[۱۴] و تابع خطر[۱۵] هستند که به ترتیب با $S(t)$ و $h(t)$ نشان داده می‌شوند.

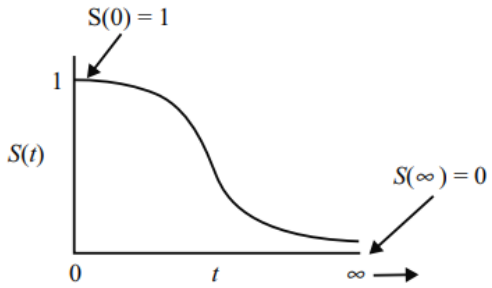
تابع بقا

تابع بقا بیان‌کننده این احتمال است که بقای فرد از مقدار مشخص t بیشتر شود. به عبارت دیگر احتمال اینکه متغیر تصادفی T بزرگتر از زمان مشخص شده t باشد، را نشان می‌دهد.

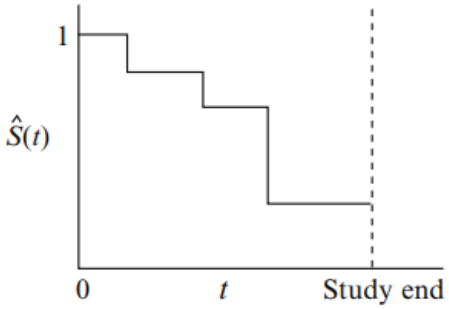
$$S(t) \ = \ p(T \ > t)$$

ویژگی‌های تابع بقا:[۵]

[۱۴] Survival Function
[۱۵] Hazard Function



شکل ۱.۵: تابع بقا



شکل ۱.۶: تابع بقا

- تابع بقا تابعی نزولی است.
- در $t = 0$ داریم $S(t) = 1$ که به معنی شروع مطالعه می‌باشد.
- در تئوری زمانی که $t = \infty$ است $S(\infty) = 0$ که این حالت زمانی رخ می‌دهد که طول دوره مطالعه بدون حد افزایش یابد و در نهایت هیچ فردی که پیشامد را تجربه نکرده، باقی نمانده باشد. در این صورت منحنی بقا سرانجام به صفر می‌رسد.

ویژگی بالا تنها در حالت تئوری برقرار می‌باشند. در عمل، زمانی که از داده‌های واقعی استفاده می‌کنیم اغلب نمودارهای توابع بقا مشابه نمودار توابع پله‌ای (شکل۱.۶) هستند.
علاوه بر این، چون دوره مطالعه هرگز بدون حد افزایش پیدا نمی‌کند و ممکن است خطرهای رقابتی برای شکست وجود داشته باشند، این امکان وجود دارد که همه افراد حاضر در مطالعه پیشامد مورد نظر را تجربه نکنند. بنابراین در عمل، هرگز تابع بقا مساوی صفر نخواهد شد. تابع بقای برآورد شده را با $\hat{S}$ نشان می‌دهیم.

تابع خطر

تابع خطر را با $h(t)$ نشان می‌دهیم و فرمول آن بصورت زیر است

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t | T > t\right)}{\Delta t} \qquad (۱.۱)$$

در رابطه ۱.۱ Δt نشان دهنده فاصله یا بازه کوچکی از زمان می‌باشد. تابع خطر تعیین کننده پتانسیل آنی برای رخداد پیشامد در هر واحد از زمان با فرض بقا یافتن فرد تا زمان t است. برخلاف تابع بقا که بر روی نداشتن شکست تمرکز می‌کند، تابع خطر بر داشتن شکست (رخداد پیشامد) متمرکز می‌شود. زمانیکه به فرمول تابع خطر نگاه می‌کنیم متوجه می‌شویم که صورت کسر در قسمت راست تساوی یک احتمال شرطی است. این احتمال شرطی بیان کننده احتمال این است که زمان بقای یک فرد T در فاصله زمانی t تا $t + \Delta t$ و به شرط اینکه زمان بقا برای آن بزرگتر یا مساوی t باشد. با توجه به صورت کسر تابع خطر، گاهی اوقات این تابع را نرخ شکست شرطی[۱۶] نیز می‌نامند. مشابه تابع بقا، تابع خطر را هم می‌توان به صورت نموداری رسم کرد. این تابع در دامنه t مقادیر مختلفی را می‌گیرد. برخلاف تابع بقا، نمودار تابع خطر می‌تواند از هرنقطه‌ای شروع شده و صعود یا نزول کند. در حالت کلی برای یک مقدار مشخص t تابع خطر $h(t)$ دارای خصوصیات زیر است:

- همیشه غیر منفی است.
- حد بالا ندارد.

هر دو ویژگی بالا با استفاده از فرمول ۱.۱ و با توجه به این نکته که هر دو مقدار صورت و مخرج فرمول $h(t)$، مقادیر غیر منفی هستند و Δt می‌تواند مقادیر صفر تا بینهایت را داشته باشد، قابل اثبات‌اند.

رابطه بین تابع خطر و تابع بقا

اگر $S(t)$ مشخص باشد، می‌توان $h(t)$ متناظر با آن را به دست آورد و برعکس. در حالت کلی می‌توان رابطه‌ی بین $S(t)$ و $h(t)$ را به صورت زیر بیان کرد.

$$S(t) = exp(-\int_0^t h(u)du\ )$$

[۱۶] Conditional Failure Rate

۳.۲.۱ کاپلان مایر

برآوردگر استاندارد برای تابع بقا توسط کاپلان مایر[۱۷] در سال ۱۹۵۸ ارائه شده‌است که به برآوردگر حد حاصل‌ضربی [۱۸] نیز معروف است. اگر d_i تعداد فوت شده‌ها در زمان t_i و y_i تعداد افراد در معرض خطر در زمان t_i بوده و t_iها زمان‌های مجزا باشند، برآوردگر ارائه شده توسط کاپلان مایر برای تابع بقا عبارتست از

$$\hat{S}(t) = \begin{cases} 1 & t < t_i \\ \prod_{t_i \le t} \left[1 - \frac{d_i}{y_i}\right] & t > t_i \end{cases}$$

این برآوردگر برای زمان‌های بزرگتر از بزرگترین زمان مشاهده شده چندان مناسب نیست زیرا اگر در t_{max} تمام موارد با مرگ روبرو شوند مقدار بقا صفر شده و اگر ناتمام در نظر گرفته شوند $S(t)$ قابل محاسبه نخواهدبود. افرون[۱۹] در سال ۱۹۶۷ پیشنهاد کرده‌است که در زمان t_{max} تمام افراد فوت شده در نظر گرفته شوند (لذا اریبی منفی ایجاد خواهد شد) گیل[۲۰] در سال ۱۹۸۰ در نظر گرفتن فوت تمامي افراد را در بینهایت پیشنهاد می‌کند (با اریبی مثبت). کلاین[۲۱] در سال ۱۹۹۱ نشان داد اگر قرار باشد بین این دو پیشنهاد یکی را انتخاب نماییم پیشنهاد گیل مناسب‌تر است. گرین[۲۲] و وود[۲۳] برآوردگری براي واریانس $S(t)$ بر مبنای برآورد کاپلان مایر به صورت زیر ارائه کرده‌اند که به نام خود آنها معروف شده است.

$$\hat{V}[\hat{S}(t)] = [S(t)]^2 \sum_{t_i \le t} \frac{d_i}{y_i(y_i - d_i)}$$

تابع مخاطره تجمعی براساس تعریف بر مبنای روش نلسون آلن عبارتست از:

$$\hat{H}(t) = \begin{cases} 0 & t < t_i \\ \sum_{t_i \le t} \frac{d_i}{y_i} & t > t_i \end{cases}$$

در سال ۱۹۷۸ این محققین برآوردی براي واریانس تابع مخاطره تجمعی فوق ارائه کردند.

$$\hat{V}[\hat{H}(t)] = \sum_{t_i \le t} \frac{d_i}{{y_i}^2}$$

---

[۱۷] Kaplan-Meier
[۱۸] product limit
[۱۹] Efron
[۲۰] Gill
[۲۱] Klein
[۲۲] Green
[۲۳] Wood

در این روش بر عکس روش کاپلان مایر تابع بقاء با کمك تابع مخاطره تجمعی بدست می‌آید:

$$\hat{S}(t) = exp[-\hat{H}(t)]$$

۴.۲.۱ مدل خطرهای متناسب کاکس

مدل خطرهای متناسب کاکس یکی از پرکاربردترین مدل‌ها برای برازش داده‌های بقا است که براساس فرض‌های همگنی جامعه، استقلال و هم‌توزیع بودن داده‌های بقا بنا شده‌است. اما در بسیاری از مواقع خطرهای واحدهای آماری متفاوت بوده و فرض همگنی جامعه برقرار نیست. یکی از دلایل این تفاوت وجود عوامل خطر ناشناخته یا مشاهده نشده است که لحاظ نکردن آنها و استفاده از مدل‌هایی همچون مدل خطرهای متناسب کاکس می‌تواند نتایج گمراه‌کننده‌ای را به همراه داشته باشد.
فرمول مدل خطرهای متناسب کاکس به صورت زیر است:

$$h(t, X) \ = \ h_0(t)e^{\sum_{i\ =\ 1}^{p} \beta_i X_i}$$

$$X = (X_1, X_2, \cdots, X_p)$$

که در آن X مجموعه‌ی متغیرهای پیشگو است. با این مدل می‌توان خطر را برای یک فرد مشخص، که دارای مجموعه مقادیر پیشگوی X است، در زمان t محاسبه کرد. فرمول مدل کاکس که خطر را در زمان t محاسبه می‌کند حاصل ضرب دو کمیت $h_0(t)$ ، که تابع خطر اولیه نامیده می‌شود و $e^{\sum_{i\ =\ 1}^{p} \beta_i X_i}$ می‌باشد و p تعداد متغیرهای پیشگو است. ویژگی مهم فرمول مدل خطرهای متناسب کاکس این است که تابع خطر اولیه تابعی از t و مستقل از Xها است. این ویژگی را فرضیه خطر متناسب می‌نامیم. در مقابل $e^{\sum_{i\ =\ 1}^{p} \beta_i X_i}$ تنها تابعی از Xها است. در این صورت Xها را متغیرهای مستقل از زمان می‌نامیم.
این امکان وجود دارد که Xها وابسته به t باشند که در این صورت آنها را متغیرهای وابسته به زمان می‌نامیم. اگر Xها وابسته به زمان باشند باز هم می‌توان از مدل کاکس استفاده کرد اما دیگر فرضیه خطر متناسب برای آنها مناسب نخواهد بود و باید از مدل کاکسی استفاده کنیم که مدل کاکس تعمیم‌یافته نامیده می‌شود. متغیر مستقل از زمان به متغیری گفته می‌شود که مقدار آن برای یک فرد با گذشت زمان تغییر نکند، مانند جنسیت. یکی دیگر از ویژگی های مدل کاکس این است که اگر همه Xها برابر صفر باشند و از مدل حذف شوند، فرمول به تابع خطر اولیه کاهش می‌یابد. همین ویژگی فرمول کاکس باعث شده تا $h_0(t)$ را تابع خطر اولیه بنامیم.

$$X_1 = X_2 = \cdots = X_p = 0 \rightarrow h(t, X) = h_0(t)e^{\sum_{i\ =\ 1}^{p} \beta_i X_i} = h_0(t)e^0 = h_0(t)$$

بنابراین $h_0(t)$ را می‌توان به عنوان یک شروع یا پایه تابع خطر قبل از در نظر گرفتن متغیرهای کمکی دانست. $h_0(t)$ یک تابع نامشخص است وهمین ویژگی باعث می‌شود تا مدل کاکس یک

مدل نیمه پارامتری باشد. می‌دانیم مدل پارامتری مدلی است که بجز مقادیر مجهول پارامترها اجزاء دیگر آن معلوم و مشخص باشد.
همانطور که قبلا اشاره کردیم خطر اولیه در مدل کاکس نامشخص است. با این حال می‌توان به وسیله آن برآوردهای خوبی برای ضرایب رگرسیون، نسبت خطر مورد نظر و منحنی‌های بقای تعدیل‌شده تولید کرد. همین ویژگی یک علت مهم برای محبوبیت مدل کاکس است. به عبارت دیگر مدل کاکس یک مدل نیرومند[۲۴] است. یعنی نتایج حاصل از برازش مدل کاکس به طور تقریبی بسیار نزدیک به نتایج حاصل از بکارگیری مدل‌های پارامتری است. برای مثال اگر مدل پارامتری مناسب برای داده‌ها، مدل وایبل باشد و ما از مدل کاکس برای به دست آوردن برآوردها استفاده کنیم، نتایج به دست آمده مشابه نتایجی خواهد بود که در صورت برازش مدل وایبل به دست می‌آوریم. اگر مطمئن باشیم که مدل پارامتری ما صحیح است ترجیح می‌دهیم برای تحلیل داده‌های بقا از مدل پارامتری استفاده کنیم. اگر چه راه‌های مختلفی برای ارزیابی نیکویی برازش مدل پارامتری وجود دارد اما هیچ‌گاه نمی‌توان مطمئن بود که مدل پارامتری انتخاب شده مناسب است. بنابراین زمانی که در انتخاب مدل پارامتری مناسب شک داریم، برازش مدل کاکس روی داده‌ها می‌تواند نتایج مطمئنی را تولید کند و کاربر لازم نیست نگران انتخاب اشتباه مدل پارامتری باشد.
در مدل کاکس عبارت نمایی این اطمینان را می‌دهد که مدل برازش شده همیشه برآورد خطر غیرمنفی را تولید می‌کند $(0 < h(t, X) < \infty)$.
ویژگی دیگر مدل کاکس این است که با وجود نامشخص بودن $h_0(t)$ امکان برآورد βها وجود دارد. پس از برآورد βها از آنها برای تعیین اثر متغیرهای پیشگوی مورد نظر استفاده می‌کنیم. همچنین می‌توان اندازه اثر که نسبت خطر (HR)[۲۵] نامیده می‌شود را بدون برآورد کردن تابع خطر اولیه محاسبه کرد. نکته قابل توجه این است که می‌توان $h(t, X)$ و منحنی بقای متناظر آن یعنی $S(t, X)$ را برای مدل کاکس برآورد کرد، بدون آنکه احتیاجی به برآورد کردن $h_0(t)$ داشته باشیم. بنابراین با استفاده از مدل کاکس می‌توانیم با کمترین فرضیات، اطلاات اولیه برای تحلیل بقا یعنی نسبت خطر و منحنی بقا را به دست آوریم. آخرین نکته درباره خصوصیات مدل کاکس این است که هرگاه پیامد مورد نظر زمان بقا است و در داده‌ها سانسور وجود دارد، مدل کاکس بر مدل لوژستیک برتری دارد. مدل کاکس نسبت به مدل لوژستیک که متغیر خروجی تنها مقادیر ۰ و ۱ را در برمی‌گیرد و زمان بقا و سانسورها را در نظر نمی‌گیرد، از اطلاعات بیشتری استفاده می‌کند.

[۲۴] Robust
[۲۵] Hazard Ratio

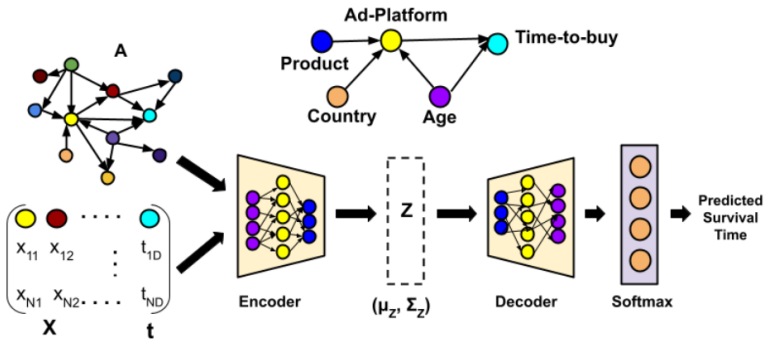
فصل ۲

DAGSurv

۱.۲ مقدمه

تحلیل و پردازش داده‌های نوین شامل حجم وسیعی از داده‌ها است که ساختار آن حاوی اطلاعات مهمی در مورد روابط متقابل بین آن‌ها است. این ساختار و روابط اغلب از گراف‌ها قابل استخراج است، به گونه‌ای که یک یال وزن دار/بدون وزن روش انعطاف‌پذیری برای نمایش رابطه بین گره‌ها ارائه می‌دهد. در دهه‌های گذشته، الگوریتم‌های پردازش سیگنال و یادگیری ماشین به تحلیل داده‌های گرافیکی می‌پردازد [۶]. در حوزه یادگیری ماشین، نادیده گرفتن این روابط بین متغیرهای کمکی ممکن است منجر به پیش‌بینی‌های اریب و نادرست شود. از این رو، ترکیب کردن دانش توپولوژی گراف در الگوریتم‌های یادگیری بسیار مهم است.

گراف‌های غیرمدور جهت‌دار (DAG) امکان مدل‌سازی آماری متغیرهای کمکی با اعمال نظم توپولوژیکی برآنها را فراهم می‌کند. DAG ها در پاسخ به سؤالات «چه اتفاقی می‌افتد اگر» مانند «اگر متغیری به جای B روی یک مقدار A تنظیم شود، سیستم چه رفتاری خواهد داشت؟» با تمرکز بر انجام اقداماتی که باعث ایجاد تغییر کنترل شده در سیستم‌ها می‌شود، مفید هستند. به عنوان مثال، هنگام قرار دادن یک تبلیغ در پلتفرم‌های آنلاین، سؤال «چه اتفاقی می افتد اگر» با پلتفرم مورد استفاده برای قرار دادن تبلیغات مرتبط است و روی خروجی آن که زمان تا خرید است، تاثیرگذار است. در صورتی که تنها رابطه علت و معلولی بین پلتفرم و نتیجه و خروجی در نظر گرفته شود، پیش‌بینی‌های اشتباهی رخ خواهد داد؛ زیرا متغیرهای کمکی مانند سن، جغرافیا، رفتار خرید آنلاین، رده‌های اقتصادی و غیره نیز هرچند غیرمستقیم، همانطور که در شکل ۱.۲ نشان داده شده است می‌تواند بر خرید تأثیرگذار باشد [۷]. مدل‌سازی چنین داده‌هایی توسط یک مدل گرافی، همانطور که در شکل ۱.۲ نشان داده شده است، به ما اجازه می‌دهد تا ساختار گراف را با استفاده از استقلال شرطی بین متغیرهای تصادفی که با رئوس نشان داده شده‌اند، کدبندی کنیم. در این کار، فرض می‌کنیم توزیع توام متغیرهای کمکی با استفاده از ماتریس مجاورت یک DAG که



شکل ۱.۲: چارچوب .DAGSurv ورودی VAE شرطی شامل مجموعه داده D (که در ادامه تعریف می‌شود) و ماتریس مجاورت A است. متغیر پنهانی که A و D را کدگذاری می‌کند با Z نشان داده شده. توزیع شرطی $P(t|X,Z)$ و یک لایه پنهان softmax برای به دست‌آوردن زمان بقای پیش‌بینی شده اعمال می‌کنیم. همچنین گراف مثال تبلیغات را نیز نشان می‌دهیم.

رئوس آن ویژگی‌های مجموعه داده هستند، تجزیه شود.
تحلیل بقا (SA) یک روش آماری شناخته شده برای مطالعه رویدادهای زمانی است که در آن داده‌های زمان تا یک رویداد با استفاده از یک تابع احتمالی از متغیرهای کمکی کاملاً یا جزئی مشاهده شده مدل‌سازی می‌شوند. یک مانع در مدل‌سازی داده‌های زمان تا رویداد، وجود مشاهدات سانسور شده است؛ به عنوان مثال، مواردی که پیشامد یا رویداد موردنظر مشاهده نشود (و از این رو، اطلاعات زمان تا رویداد گم‌شده است). نادیده گرفتن داده‌های سانسور شده، باعث اریبی در فرآیند استنتاج می‌شود و از این رو، تحلیل چنین داده‌هایی نیازمند روش‌های آماری و یادگیری ماشینی متفاوت است [۸]. علاوه بر این، روش‌های ماکسیمم درستنمایی برای تحلیل بقا هیچ رابطه‌ای بین ویژگی‌ها اعمال نمی‌کند و مدل اثرات متقابل بین ویژگی‌ها و نتایج زمان تا رویداد را یاد می‌گیرد. در پژوهش معرفی شده در این پروژه، DAG ها به عنوان ورودی در نظر گرفته می‌شوند که در آن ویژگی‌ها به عنوان گره‌های DAG ها و اثرات متقابل بین ویژگی‌ها توسط یال‌های DAG ها نشان داده می‌شوند.
در این پژوهش، رابطهٔ علت و معلولی بین متغیرها و پاسخ زمان تا رویداد را از طریق کدگذاری ساختار علّی DAG با تحلیل داده‌های زمانی ادغام می‌کنیم. سهم و مشارکتی که این روش در بهبود تحلیل‌ها دارد به شرح زیر است:

- با استفاده از نظریه اطلاع کدگذاری منبع، نشان می‌دهیم که استفاده از دانش ماتریس مجاورت همراه با متغیرهای کمکی ورودی، در مقایسه با حالتی که متغیرهای کمکی از نظر آماری مستقل فرض می‌شوند، منجر به کدگذاری بهینه توزیع منبع می‌شود.

- با استفاده از نظریه کدگذاری منبع، یک خودرمزگذار تغییراتی شرطی(CVAE[1]) براساس یادگیری عمیق جدید پیشنهاد می‌شود تا از دانش DAG علّی برای پیش‌بینی داده‌های بقای ساختاریافته، که ما به آن DAGSurv می‌نامیم، استفاده کند.

- عملکرد چارچوب پیشنهادی DAGSurv با استفاده از شاخص هماهنگی(CI[2]) زمان‌ـ وابسته به عنوان یک سنجه، روی مجموعه داده‌های ساختگی و واقعی مانند Metabric و GBSG بررسی می‌شود.

با استفاده از نتایج تجربی، نشان می‌دهیم که ترکیب DAG علی در پیش‌بینی بقا نه تنها برای بهبود نتایج، بلکه برای اعتبار سنجی پویایی علّی مفروض یک سیستم نیز مفید است. در مجموعه داده‌های واقعی، DAG، به راحتی در دسترس نیستند و از این رو، از مرحله پیش‌پردازش برای براورد گراف براساس مجموعه داده مشخص استفاده می‌شود و از این گراف به عنوان ورودی مدل پیشنهادی استفاده می‌شود. نتایج شبیه‌سازی فرضیه‌های مطرح شده مبنی بر اینکه ترکیب DAG در مدل یادگیری ماشین به نمایش بهتر داده‌ها منجر می‌شود را تأیید می‌کند که در نهایت این کار منجر به بهبود مقادیر CI زمان‌ـوابسته، در مقایسه با روش‌های تحلیل بقای مرسوم می‌شود.
درنهایت، مقدمات ریاضی تحلیل بقا را به دنبال استدلال کدگذاری منبع برای فشرده‌سازی بهینه منبع در صورتی که ماتریس مجاورت مشخص باشد، توصیف می‌کنیم. در ادامه، چارچوب پیشنهادی DAGSurv تعریف و براساس نتایج تجربی نتیجه‌گیری می‌شود.

۲.۲ کارهای مرتبط

بارها ثابت شده است که ترکیب دانش ساختار گراف با مدل‌های یادگیری ماشین مزایای بسیار زیادی به همراه دارد. شبکه‌های پیچشی گراف (GCN[3]) ابزارهای قدرتمندی هستند که با گراف‌های غیر جهت‌دار برای رده‌بندی نیمه‌نظارت شده در هر نمونه از مجموعه داده‌ها استفاده می‌شوند [۹].
در این پژوهش، بر استخراج رابطه بین متغیرهای کمکی در یک مجموعه داده تمرکز شده است، و بنابراین، GCN به طور مستقیم کاربردی ندارند. گراف دانش، توانایی برقراری روابط بین متغیرها را به شیوه‌ای کارآمد که قابل توضیح و استفاده مجدد باشد به ارمغان می‌آورد. با این حال، این روابط اغلب معنایی هستند [۱۰]، و ممکن است ارتباط آماری نداشته باشد.
در حالتی که گراف‌ها اطلاعات آماری را ارائه دهند، چارچوب مدل‌های گرافیکی احتمالاتی نقش مهمی ایفا می‌کنند [۱۱]. در مدل‌های گرافیکی احتمالی، گره‌های یک گراف به‌عنوان متغیرهای تصادفی در نظر گرفته می‌شوند و اطلاعات متغیرهای کمکی و هدف برای فهم این متغیرهای تصادفی در نظر گرفته می‌شوند. بدیهی است که یال‌های بین متغیرهای تصادفی روابط آماری بین

[1] conditional variational autoencoder
[2] concordance index
[3] Graph convolutional networks

متغیرهای تصادفی را نشان می‌دهند و از این رو، گراف، یک توزیع توام روی مجموعه داده تشکیل می‌دهد. در سناریوهایی که گراف اصلی معلوم است، از شبکه‌های عصبی عمیق همراه با مدل‌های گرافیکی برای پیش‌بینی استفاده می‌شود [۱۲]. در این پژوهش، از چارچوب مبتنی بر مدل‌های گرافیکی احتمالی برای پیش‌بینی بقا مبتنی بر گراف استفاده شده است.
در زمینه تحلیل بقا، روش کاپلان‌ـ‌مایر (KM) یک روش ناپارامتری محبوب اما ساده برای به دست آوردن براورد تجربی تابع بقا است [۱۳]. پیشرفتی که در روش KM بوجود آمده است، مدل خطرات متناسب کاکس (CPH) [۱۴] است که متغیرهای کمکی را برای استنباط ترکیب می‌کند. چندین روش پارامتری با استفاده از توزیع‌های وایبل یا لگ‌نرمال [۱۵] و روش‌های ناپارامتری با استفاده از فرآیندهای گاوسی برای تحلیل بقا پیشنهاد شده است. روش‌های مدرن مبتنی بر شبکه‌های عصبی عمیق (DNN[۴]) برای تحلیل زمان تا رویداد [۱۶] استفاده شده است، به طوری که در آن‌ها برای مدل‌سازی رابطه بین متغیرهای کمکی و مخاطره، نمایش غیرخطی جایگزین مدل‌های خطی می‌شود. اما محدودیت این روش‌ها، فرض ثابت بودن نرخ خطر و خطی بودن نرخ لگاریتم خطر است. در مقاله (لی و همکاران ۲۰۱۸)، نویسندگان یک رویکرد منحنی شاخص تجمعی (CIC[۵]) را پیشنهاد کرده‌اند که از احتمالات حاشیه‌ای یک پیشامد یا رویداد، در حضور چندین پیشامد رقیب استفاده می‌کند. این روش فرض ثابت بودن نرخ خطر یا هر فرض دیگری در مورد مدل را درنظر نمی‌گیرد.
از مدل‌های گرافیکی احتمالی در تحلیل بقا استفاده شده است [۱۷] که در آن الگوریتم‌هایی استنباطی گراف‌مبنا برای پیش‌بینی بقا با فرض نرخ خطر ثابت پیشنهاد شده است. در مقابل، در این پژوهش روش VAE شرطی (CVAE) برمبنای رویکرد مدل‌های گرافیکی برای پیش‌بینی بقای ساختاریافته پیشنهاد شده است که در آن نرخ خطر ثابت فرض نشده است. این پژوهش ارتباط نزدیکی با DAG-GNN (یو و همکاران، ۲۰۱۹) دارد. لازم به ذکر است که CVAE پیشنهادی از جنبهٔ طراحی از DAG-GNN الهام گرفته شده است، اما از نظر عملکرد در مقایسه با DAG-GNN تفاوت اساسی دارد (یو و همکاران، ۲۰۱۹). در VAE, DAG-GNN (و نه CVAE) به گونه‌ای طراحی شده است که به یادگیری ماتریس مجاورت وزنی DAG بپردازد و تنها مربوط به حوزه یادگیری ماشین نیست. در این پژوهش، از ماتریس مجاورت به عنوان یک پارامتر معلوم استفاده می‌شود و سپس یک مدل یادگیری ماشینی عاری از فرض برای پیش بینی بقا به دست می‌آید. اگرچه، موضوع این پژوهش تحلیل بقا است، اما از این مدل می‌توان برای کارهای رده‌بندی و رگرسیون نیز استفاده کرد.
روش‌های متعددی که نمایش گرافی روابط ویژگی‌ها را با رویکردهای پیش‌بینی با استفاده از GCN ترکیب می‌کنند، در مرور مطالب و مقالات پیشنهاد شده است. اما، این روش‌ها حوزه‌های جداگانه‌ای برای تعبیه‌سازی یا جاسازی گراف و رگرسیون، رده‌بندی یا تحلیل بقا ترکیب می‌کنند. به عنوان مثال، در [۱۸]، یک گراف بین تکه‌های تصاویر آسیب‌شناسی در نظر گرفته می‌شود و نمایش ویژگی که از طریق GCN تولید شده است، برای تحلیل بقا در نظر گرفته شده است. از طرف دیگر، ما

[۴] deep neural networks
[۵] cumulative index curve

دانش گراف را در شبکه جاسازی کرده‌ایم و به طور خاص به مسئلهٔ تحلیل بقا پرداخته‌ایم. یک کار مشابه و مرتبط مربوط به مقاله (چن و همکاران ۲۰۱۹) است که در آن نویسندگان با استفاده از یک مدل گرافیکی احتمالی با توزیع خانواده نمایی، یک تحلیل بقای مبتنی بر گراف بدون جهت را برای توصیف رابطه بین متغیرهای کمکی، پیشنهاد کرده‌اند. در مقایسه با آن کار، این پژوهش به طور خاص از DAG ها برای مدل‌سازی روابط علی استفاده کرده است و از مدل‌های احتمالی خاص برای بررسی روابط بین متغیرهای کمکی استفاده نکرده است.

۳.۲ تحلیل بقا مبتنی بر گراف غیرمدور جهت‌دار

در این بخش، مسئلهٔ تحلیل بقای مبتنی بر DAG شرح داده می‌شود. ابتدا، مقدمات ریاضی پیش‌بینی بقا ارائه می‌شود و در ادامه مسئله بر اساس استدلال کدگذاری منبع فرمول‌بندی می‌شود. در این پژوهش، چارچوب CVAE به عنوان یک رمزگذار (کدگذار) منبع ممکن که از دانش DAG برای پیش‌بینی بقا استفاده می‌کند، پیشنهاد شده است. در این پژوهش، تابع زیان تغییراتی به گونه‌ای بسط داده شده است که دو منظوره است به این معنا که DAG علّی را همراه با پارامترهای سیستم یادگیری برای پیش‌بینی بقا ترکیب می‌کند.

۱.۳.۲ مقدمات ریاضی

مجموعه داده‌های زمان تا رویداد معمولا به صورت $D = (X^{(n)}, t^{(n)}, \delta^{(n)})_{n=1}^{N}$ مشخص‌سازی می‌شوند که شامل سه متغیر برای n امین نمونه است. که در آن، $x^{(n)} \in \mathbb{R}^L$ مربوط به نمونه n ام، $X \in \mathbb{R}^{N \times L}$ است. L معرف تعداد متغیرهای کمکی است. زمان بقای $t^{(n)}$ گسسته و افق زمانی متناهی در نظر گرفته شده‌است به گونه‌ای که $t \in T$ که در آن $T = \{0, \ldots, M\}$ حداکثر افق زمانی از پیش تعریف شده M. علاوه بر این، $t \in \mathbb{R}^{N \times 1}$ نشان دهنده زمانی است که در آن رویداد رخ داده است و $\delta^{(n)} \in \{0, 1\}$ یک متغیر نشانگر است که مشخص می‌کندnامین نمونه سانسور شده است یا نه.
مدل‌های زمان تا رویداد با تابع بقایی که در زیر ارائه شده است، مشخص می‌شود:

$$S(t|x) = P(T > t\,|x) = 1 - F(t|x)$$

که به عنوان کسری از جمعیتی که تا زمان t زنده مانده است، تعریف می‌شود؛ که در آن $F(t|x)$ تابع توزیع تجمعی زمان تا رویداد با معلوم بودن متغیر x را نشان می‌دهد. یک آماره مهم دیگر تابع نرخ خطر شرطی $h(t|x)$ است که به عنوان نرخ لحظه‌ای وقوع یک رویداد در زمان t با معلوم بودن متغیرهای کمکی x تعریف می‌شود. رابطه بین $h(t|x)$ و $S(t|x)$ براساس تعاریف استاندارد به صورت زیر ارائه می‌شود:

$$h(t|x) = \lim_{dt \to 0} \frac{P(t < T < t + dt|x)}{P(T > t|x)} = \frac{f(t|x)}{S(t|x)} \qquad (۱.۲)$$

که در آن $f(t|x)$ تابع چگالی شرطی بقا است و $S(t|x)$ نیز قبلا توضیح داده شده است. مدل خطرهای متناسب کاکس (کاکس، ۱۹۷۲) یک مدل خطی نیم‌پارامتری است که در آن تابع خطر شرطی $h(t|x)$ از طریق تابع خطر پایه $h_0(t)$ به زمان و متغیرهای کمکی مستقل x بستگی دارد به طوریکه

$$h(t|x) = h_0(t)exp(x^T\gamma) \qquad (۲.۲)$$

برای یک مجموعه داده مشخص با N مشاهده، همانطور که قبلا توضیح داده شده مدل خطرهای متناسب کاکس، ضرایب رگرسیون $\gamma \in \mathbb{R}^L$ را به گونه‌ای براورد می‌کند که درستنمایی جزئی را ماکسیمم می‌کند (کاکس، ۱۹۷۲). در ،DeepSurv نویسندگان یک مدل خطرهای متناسب کاکس را براساس DNN پیشنهاد می‌کنند.علاوه بر این، DeepHit به طور مستقیم توزیع توام زمان‌ها و رویدادهای بقا را یاد می‌گیرد و به طور موثر از فرضیات خطرهای متناسب اجتناب می‌کند. در این روش‌ها، متغیرهای کمکی مستقل فرض می‌شوند و هیچ مکانیسم رسمی وجود ندارد که با استفاده از آن بتوان هر گونه وابستگی بین متغیرهای کمکی را در نظر گرفت. در چن (۲۰۱۹)، یک گراف بدون جهت بین متغیرهای کمکی در نظر گرفته شده و یک مدل گرافیکی احتمالی مبتنی بر توزیع نمایی در تحلیل درنظر گرفته شده است.اما، در مقابل، در این پژوهش، یک چارچوب مبتنی بر CVAE با درنظرگرفتن یک DAG بین متغیرهای کمکی برای پیش‌بینی بقا طراحی شده است. توجه داشته باشید که روش پیشنهادی به هیچ یک از فرضیات مدل‌بندی مانند آنچه در چن (۲۰۱۹) اشاره شده است، نیاز ندارد، و از این رو، برای همه مجموعه داده‌ها مناسب است.

۲.۳.۲ فرمول‌بندی مسئله

در این پژوهش، از DAG که با $G(V, E)$ نمایش داده می‌شود، برای توصیف رابطه علی بین ویژگی‌ها در مجموعه داده D استفاده شده است. هر رأس در $G(V, E)$ معرف یک متغیر تصادفی با $V = \{1, ... , L + 1\}$ است که نشان‌دهندهٔ اندیس‌های این متغیرهای تصادفی است. به عنوان مثال، X_l یک راس است که $l \in V$. بعلاوه، فرض کنید $V \times V$ شامل تمام جفت‌های اندیس‌ها در V است. یک جفت متغیر تصادفی $\{X_l, X_m\}$ یال گراف G نامیده می‌شود $(l, m) \in E \subset V \times V$. $L + 1$ راس شامل L متغیرهای کمکی در X است و راس $L + 1$ ام متغیر هدف داده شده توسط زمان بقا t است. فرض کنید $A \in \mathbb{R}^{(L+1)\times(L+1)}$ نشان دهنده ماتریس مجاورت وزنی این DAG است.

در این پژوهش، ماتریس کمکی X و ماتریس مجاورت A برای ارائهٔ یک نمایش کارآمد برای پیش‌بینی بقای ساختاریافته، کدگذاری شده است. مسئلهٔ کدگذاری توام X و A به عنوان یک مسئلهٔ کدگذاری منبع است. در این پژوهش، از اصول اولیه نظریه اطلاع که محدودیت اساسی برای فشرده‌سازی اطلاعات ایجاد می‌کند، استفاده شده است. برای فشرده‌سازی بهینه منبع، طول مورد انتظار کد منبع باید بزرگتر یا مساوی با آنتروپی منبع باشد [۱۹]. لازم به ذکر است که ماتریس

مجاورت رابطه احتمالی بین ویژگی‌ها را همانطور که در قضیه زیر نشان داده شده است، کنترل می‌کند.

گزاره ۱.۲. ماتریس مجاورت A از گراف غیرمدور جهت‌دار $G(V,E)$، توزیع توام $p(t,X)$ را مشخص‌سازی می‌کند.

گزاره ۲.۲. ماتریس مجاورت A یک ماتریس غیرصفر است اگر و تنها اگر جمله i ام در تجزیهٔ $p(X|K_A)$ که با $p(X_i\ |X_{pa(i)})$ نمایش داده می‌شود، برای هر i ، برابر با $p(X_i)$ نباشد.

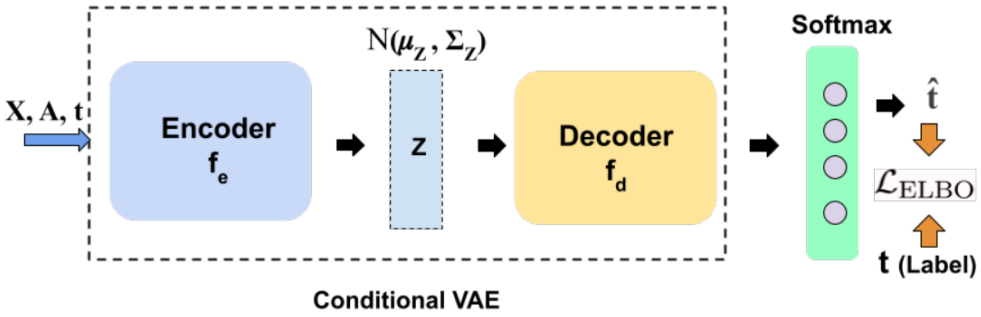
گزاره ۳.۲. اگر به ازای هر i، تجزیه iام $p(t,X|K_A)$ که بصورت $p(X_i|X_{pa(i)})$ نشان داده می‌شود برابر با $p(X_i)$ نباشد، داریم $H(X) < \sum_{i=1}^{L} H(X_i)$ که در آن $H(.)$ تابع آنتروپی است.

در گزاره‌های ذکر شده در بالا، مشاهده می‌کنیم که اگر برای همه i,j، $A(i,j) \neq 0$ باشد، آنتروپی منبع به شدت کوچک‌تر از آنتروپی منبعی است که از نظر آماری مستقل‌اند. علاوه بر این، اگر دانش حاصل از ماتریس مجاورت A برای نمایش داده‌ها تهیه نشده باشد، کدگذار بهینه ممکن است برای همه i,j، $A(i,j) = 0$ را در نظر بگیرد و در نتیجه تعداد بیت‌های مورداستفاده برای نشان دادن منبع به اندازه $\sum_{i=1}^{L} H(X_i)$ است. بنابراین، بدیهی است که دانش A باید به طور مناسب برای نمایش داده‌های منبع استفاده شود، به طوری که تعداد بیت‌های مورد نیاز برای رمزگذاری چنین منبعی به شدت کمتر از $\sum_{i=1}^{L} H(X_i)$ باشد. اکنون، این استدلال رمزگذاری منبع نظری اطلاعات بنیادی را به دلیل اینکه انگیزه قوی برای طراحی رمزگذارهای کارآمد را به ما می‌دهد، بیان و اثبات می‌کنیم. در این راستا، از دانش ماتریس مجاورت A در مفهوم پیش‌بینی بقای ساختاریافته استفاده می‌کنیم.

CVAE و تابع هزینه

یک رویکرد ممکن برای استفاده از دانش ماتریس مجاورت برای رمزگذاری منبع، استفاده از خودرمزگذار تغییراتی (VAE) است [۲۰]. نویسندگان متعددی از VAE ها برای کدگذاری توام منبع به روش کدگذاری کانال [۶] استفاده کرده‌اند (چو و همکاران ۲۰۱۹). با این انگیزه، یک چارچوب خودکدگذار تغییراتی شرطی (CVAE) برای پیش‌بینی بقای مبتنی بر DAG استخراج شده است به گونه‌ای که دانش A را در برگیرد.
از CVAE استاندارد (چو و همکاران ۲۰۱۵) برای ترکیب DAG در پیش‌بینی بقا استفاده شده است. عبارت شرطی به تابع چگالی احتمال شرطی (pdf) که در CVAE استفاده می‌شود، به جای تابع چگالی احتمال توام که در VAE استفاده می‌شود، اشاره دارد. اگرچه VAE و CVAE

[۶] اصلاح خطای روبه‌جلو(FEC) یا کدگذاری کانال (Channel coding) در مخابرات، نظریه اطلاع و نظریه کدگذاری به افزودن تعدادی بیت زائد (Redundant bits) به بیت‌های اطلاعات برای مقابله با بروز خطا در خطوط نویزدار گفته می‌شود. این کار برای انتقال بدون خطای (یا کم‌خطا) اطلاعات لازم است.



شکل ۲.۲: X،A و t به عنوان ورودی به CVAE در طول یادگیری داده می‌شوند. رمزگشا، توسط لایه softmax دنبال می‌شود، به طوری که خروجی $\hat{t}$ نشان دهنده احتمال اینکه فردی رویدادی را در یک زمان معین تجربه کند، است. در طول زمان آزمون، تنها از رمزگشای (f_d) استفاده می‌شود که در آن X و Z (نمونه‌های ورودی به رمزگشا از $N(0, I)$ هستند) ترفند پارامترسازی مجدد تضمین می‌کند که Z از $N(\mu_Z, \Sigma_Z)$ نمونه‌گیری شده و این توزیع در طول یادگیری در رمزگشا تعبیه شده است) و A به عنوان ورودی و $\hat{t}$ به عنوان خروجی تعیین شده است.

از فرمول مبتنی بر متغیر پنهان استفاده می‌کند، شرطی کردن x منحصر به CVAE است. کار جدید روش پیشنهادی، ترکیب دانش DAG و ویژگی‌های فردی برای تحلیل بقا از طریق کدگذاری ساختار DAG در گراف به عنوان اطلاعات اضافی است. جنبه مولد و تولیدی CVAE این است که به چارچوب ELBO امکان کدگذاری گراف در شبکه عصبی را می‌دهد و قابلیت پیش‌بینی DAGSurv نتیجه قابلیت پیش‌بینی CVAE است. برای طراحی ،DAGSurv از فرآیند تولید نمونه بر اساس SEM تعمیم‌یافته که به صورت

$$t = f_d\left(\left(I - A^T\right)^{-1} g\left(\left[X^T, Z^T\right]\right)\right) \qquad (۳.۲)$$

مشخص می‌شود، استفاده شده است که در آن A^T ترانهاده ماتریس A است.
$f_d : \mathbb{R}^{(L+1)\times N} \to \mathbb{R}^{M\times 1}$ و $g : \mathbb{R}^{(2L+1)\times N} \to \mathbb{R}^{(L+1)\times N}$. همچنین، ورودی رمزگشا (کدگشا)، ماتریس مجاورت A و یک ماتریس پیوستگی شامل X و Z است. در اینجا $Z \in \mathbb{R}^{N\times(L+1)}$ یک متغیر پنهان با توزیع پیشین گاوسی میانگین صفر $N(0, I)$ است و I ماتریس همانی است. اغلب، رابطهٔ (۳.۲) به مدل رمزگشا اشاره دارد، و مدل رمزگذار متناظر به صورت زیر

$$Z^T = (I - A^T) f_e(\widetilde{X}^T) \qquad (۴.۲)$$

مشخص می‌شود که در آن $f_e : \mathbb{R}^{(L+1)\times N} \to \mathbb{R}^{(L+1)\times N}$ یک تابع پارامتری از رمزگذار (کدگذار) است و $\widetilde{X} \in \mathbb{R}^{N\times(L+1)}$ نشان‌دهنده ماتریس افزوده متشکل از ویژگی‌های X و بردار

زمان تا رویداد t ، یعنی $\widetilde{X} = [X, t]$ است. توجه داشته باشید که اگر در رابطهٔ بالا $A = 0$ باشد، رمزگذار به صورت $Z^T = f_e(\widetilde{X}^T)$ و رمزگشا با $t = f_d\ (g[X^T, Z^T])$ مشخص می‌شود، که مشابه رمزگذار و رمزگشا مطابق با CVAE مرسوم است، که در آن متغیرهای کمکی X به عنوان ورودی و از نظر آماری مستقل در نظر گرفته می‌شوند.
برای اهداف پیش‌بینی زمان بقای داده‌رهنمون، پارامترهایی که رمزگذار و رمزگشا را تشکیل می‌دهند، با ماکسیمم‌سازی لگاریتم شواهد $\frac{1}{N}\sum_{n=1}^{N} ln(p(t^n|x_x))$ یاد می‌گیرند (آموزش می‌بینند)، که در آن x_n نشان‌دهنده متغیرهای کمکی نمونه n ام در X است. ماکسیمم‌سازی لگاریتم شواهد اغلب غیرقابل حل است، و از این رو، در این پژوهش به نظریه استنباط تغییراتی روی آورده شده است که امکان ماکسیمم‌سازی حد پایین شواهد را می‌دهد که به عنوان ELBO شناخته شده است (بیشاپ، ۲۰۰۶). رابطه بین لگاریتم شواهد و ELBO به صورت
(۵.۲)

$$\frac{1}{N}\sum_{n=1}^{N} ln(p(t^n|x_x)) \ \geq\ \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q(z_n|x_n,t^{(n)})}\left[ln\left(\frac{p(t^{(n)}, z_n|x_n)}{q(z_n|x_n, t^{(n)})}\right)\right] \equiv\ \iota_{ELBO}$$

مشخص شده است. در اینجا، $q(z_n|x_n, t^{(n)})\ , 1 \leq\ n \leq N$ توزیع پسین متغیر را نشان می‌دهد، که نمونه‌ها را در متغیر پنهان z_n رمزگذاری می‌کند. همچنین z_n نشان‌دهنده n امین سطر Z است. برخلاف VAE مرسوم، رمزگشا در CVAE برای پیش‌بینی متغیر هدف که در این پژوهش، زمان تا رویداد t برای نمونه‌هایی که قبلا دیده نشده‌اند است، آموزش داده شده است. به طور خاص، میانگین و کوواریانس توزیع شرطی $p(t|X, Z)$ و پیش‌بینی ها از طریق نمونه‌گیری از توزیع شرطی به دست آمده‌اند. علاوه بر این، در این پژوهش، ι_{ELBO} به صورت زیر ساده شده است (بیشاپ، ۲۰۰۶)[۲۱].
(۶.۲)

$$\iota_{ELBO} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q(z_n|x_n,t^{(n)})}\left[ln(p(t^{(n)}, z_n|x_n))\right]\ -\ D_{KL}(q(z_n|x_n, t^{(n)})||p(z_n))$$

که در آن $D_{KL}(.||.)$ تابع واگرایی KL و $p(z_n)$ توزیع پیشین z_n است. از این رو، ELBO منجر به یک تابع هدف مبتنی بر درستنمایی مورد انتظار می‌شود که توسط واگرایی KL محدود می‌شود. از آنجا که داده‌های زمان تا رویداد سانسور می‌شوند، بنابراین

$$lnp(t^{(j)}|x_j, z_j)\ =\ \delta_j ln\ f\ (t^{(j)}|x_j, z_j) + (1 - \delta_j) ln\ S(t^{(j)}|x_j, z_j) \qquad (۷.۲)$$

به طوریکه δ_j همانطور که قبلاً تعریف شد، یک متغیر شاخص است. $f(t|x, z)$، چگالی شکست و $S(t|x, z)$، تابع بقا است. $\hat{t}$ یک توزیع احتمال است $\hat{t} = [\hat{t}_1, \cdots, \hat{t}_M]$ یعنی با توجه به متغیرهای X، همانطور که در شکل ۲.۲ نشان داده شده است، t_k احتمال این است که فرد رویداد را در دوره زمانی k ام تجربه کند. مشابه (لی و همکاران، ۲۰۱۸)، تابع هزینه در رابطهٔ ۷.۲

شبکه را به یادگیری روابط غیرخطی و غیرمتناسب بین متغیرهای کمکی و ریسک‌ها برای یک رویداد معین هدایت می‌کند. از این رو، تابع هزینه کلی CVAE مبتنی بر بقا، تابع هزینه فوق را در LELBO ادغام می‌کند. به منظور انجام طرح پیشنهادی، از مدل رمزگذار استفاده شده است که از یک پرسپترون چند لایه (MLP) با وزن‌های W_e که با f_e نشان داده می‌شود، استفاده شده است. بر این اساس، در رمزگشا، f_dیک MLP با وزن‌های W_d و به دنبال آن یک لایه softmax است. رمزگشای CVAE نمونه‌ها را از توزیع شرطی $p(t|Z,X)$ تولید می‌کند که به صورت زیر نشان داده می‌شود.

$$\hat{t} \leftarrow Softmax((I - A^T)^{-1}Z, W_d, X) \qquad (۸.۲)$$

به گونه‌ای که Z در رمزگذار تولید می‌شود. وزن‌های W_e و W_d و توابع f_e و f_d با به حداکثر رساندن ι_{ELBO}، همانطور که در ۶.۲ ارائه شده است، آموزش داده می‌شوند. از آنجا که تابع هزینه مبتنی بر SA ارائه شده در ۷.۲ را در ι_{ELBO} ادغام می‌کنیم، می‌توان CVAE را برای پیش‌بینی کارآمد، مبتنی برگراف و زمان تا رویداد آموزش داد. همانطور که در شکل ۲.۲ نشان داده شده است، برای پیشبینی نمونه‌هایی که قبلا دیده نشده بودند، فقط از رمزگشا استفاده می‌شود.

فصل ۳

نتایج شبیه‌سازی

در این بخش، کارایی DAGSurv روی چند مجموعه داده‌های بالینی ساختگی و واقعی که در دسترس عموم است مانند ،METABRIC GBSG وRotterdam [۲۲] نشان داده شده است.
در این بخش به معرفی مجموعه داده‌ها به همراه مراحل پردازش، و معیارهای ارزیابی، رویکردهای پایه و ویژگی‌های پیاده‌سازی DAGSurv پرداخته شده است و کد این پژوهش در لینک : $https$ $//github.com/rahulk207/DAGSurv$
در دسترس عموم است.

۱.۳ معرفی مجموعه داده و پردازش داده‌ها

۱.۱.۳ مجموعه داده ساختگی

ما یک DAG تصادفی، $G(V, E)$ را با استفاده از مدل اردوش و رینی (اردوش و رینی ۱۹۵۹). که در آن $|V| = L + 1$ که L تعداد متغیرهای کمکی و ۱ متغیر هدف که همان زمان تا رویداد است، اشاره دارد. برای شبیه سازی، درجه گره مورد انتظار، سه، درنظر گرفته شده است. وزن یال‌ها

جدول ۱.۳: توصیف مجموعه‌داده ساختگی و واقعی(C_{max} ماکسیمم زمان سانسور است)

مجموعه‌داده	سانسورشده	ویژگی‌ها	T_{max}	C_{max}
Synthetic-small	%50.06	9	377	91
Synthetic-large	%51.58	49	395	235
METABRIC	%42.06	9	355	337
GBSGS	%43.23	7	83	87

جدول ۲.۳: فراپارامترهایی که در مجموعه‌داده‌های مختلف نشان داده شده است: n_l و n_h تعداد لایه و تعداد گره‌های پنهان هر لایه را نشان می‌دهد و lr نرخ یادگیری است.

مجموعه‌داده	n_l, n_h (رمزگذار)	n_l, n_h (رمزگشا)	فعال‌سازی	lr
Synthetic-small	$5, 128$	$3, 64$	ReLU	$1e-4$
Synthetic-large	$5, 64$	$4, 32$	ReLU	$1e-5$
METABRIC	$3, 256$	$3, 64$	SELU	$1e-5$
GBSGS	$3, 128$	$3, 32$	ReLU	$1e-5$

به طور یکنواخت اما به طور تصادفی مقداردهی اولیه شده است، یعنی $\forall e \in E$ وزن یال DAG به صورت $W(e) \sim U(0.5,\ 2)$ درنظر گرفته شده است. رابطه مبتنی بر DAG در میان متغیرهای کمکی با استفاده از معادلات زیر تعبیه شده است (یو و همکاران، ۲۰۱۹).

$$X^T = A^T(cos(\widetilde{X}\ +\ 1))\ +\ Z_X^T\ ,\quad t = max(0, c\ expA^T(cos(\widetilde{X}^T\ \ +1))\ +\ Z_t^T \tag{۱.۳}$$

که در آن ورودی های Z_X و Z_t به ترتیب به طور مستقل از $N(0,1)$ و $N(30,70)$ نمونه‌گیری شده است. علاوه بر این، 1 یک ماتریس است که تمام مولفه‌های آن یک است، 0 یک ماتریس تمام صفر است، و c ثابت انتخاب شده است به طوری که ما t را در یک محدوده مشخص به دست آوریم. $c = 90$ در نظر گرفته شده است. با استفاده از این فرآیند تولید داده، ۱۰۰۰۰ نقطه داده به دست می‌آوریم. اگرچه محافظه کارانه است، اما ۵۰ درصد از داده ها سانسور شده است و زمان سانسور به صورت یکنواخت اما تصادفی به صورت $U(0, max(t))$، نمونه‌گیری شده است. با استفاده از موراد درنظر گرفته شده در بالا، دو مجموعه داده زیر تولید شده است.

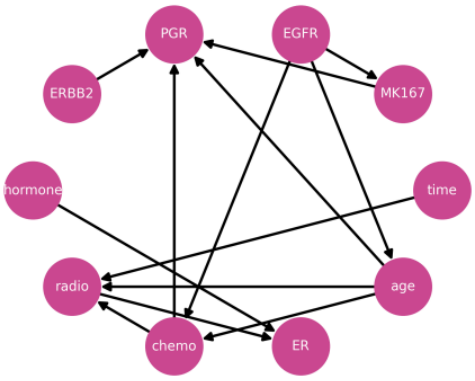
۱. Synthetic-small: $L = 9$ متغیر کمکی درنظر گرفته شده است (در نتیجه $|V| = 10$).

۲. Synthetic-large: برای آزمون مقیاس‌پذیری و عملکرد مدل روی یک مجموعه داده بزرگ، $L = 49$ در نظر گرفته شده‌است.

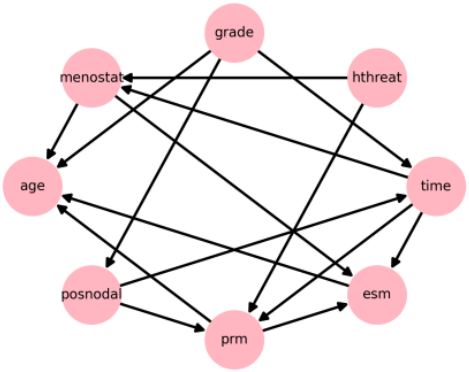
۲.۱.۳ مجموعه دادهٔ واقعی

در مجموعه داده‌های واقعی، DAG ورودی معلوم نیست. با توجه به متغیرهای کمکی در مجموعه داده، ساخت دستی یک DAG به دلیل نیازمندی به تخصص خاص، ممکن است غیرممکن باشد و می‌تواند یک فرآیند پرهزینه باشد. از این رو، از دو الگوریتم معروف برای پیش محاسبه ماتریس مجاورت A استفاده شده است که در ادامه معرفی شده است:

۱. bnlearn R package: از الگوریتم Hill Climbing (HC) برای یادگیری ساختار شبکه بیزی استفاده کردیم که منجر به یک گراف جهت‌دار بدون‌وزن می‌شود.



شکل ۱.۳: DAG: METABRIC



شکل ۲.۳: DAG: GBSGS

۲. DAG-GNN :DAG-GNN یک مدل یادگیری عمیق است که برای تولید یک DAG وزن‌دار است که ساختاری را در میان ویژگی‌های یک مجموعه داده مشخص ایجاد می‌کند.

از این الگوریتم‌ها در داده‌های واقعی به صورت زیر استفاده شده است:

- METABRIC: کنسرسیوم بین‌المللی تاکسونومی مولکولی سرطان سینه (METABRIC) یک مجموعه داده بالینی است که شامل بیان ژنی است که برای تعیین زیر گروه‌های مختلف سرطان سینه استفاده می‌شود. ما داده‌ها را برای ۱۹۰۴ بیمار در نظر می‌گیریم که هر بیمار دارای ۹ متغیر کمکی شامل چهار شاخص ویژگی ژن (MKI67, EGFR, PGR, ERBB2) و ۵ ویژگی بالینی (شاخص درمان هورمونی، شاخص رادیوتراپی، شاخص شیمی‌درمانی، شاخص ER مثبت، سن تشخیص) است. همچنین از مجموع ۱۹۰۴ بیمار، ۸۰۱ نفر(42.06%) راست سانسور شده و مابقی فوت کرده‌اند (رویداد) . ما DAG را همانطور که در شکل ۱.۳ نشان داده شده است با استفاده از یک الگوریتم اصلاح شده DAG-GNN به دست آوردیم.

- GBSGS: گروه مطالعاتی سرطان سینه روتردام و آلمان (GBSG) حاوی داده‌های سرطان سینه از بانک تومور روتردام است. مجموعه داده شامل ۲۲۳۲ بیمار است که از این تعداد ۹۶۵ نفر (43.23%) سانسور شده‌اند، بقیه فوت کرده‌اند (رویداد)، و هیچ مقدار گم‌شده ای وجود ندارد. در مجموع، ۷ ویژگی برای هر بیمار وجود دارد که شامل هورمون درمانی ،(hthreat) سن بیمار، وضعیت یائسگی، درجه تومور، تعداد گره‌های مثبت، گیرنده پروژسترون (در fmol) و گیرنده استروژن (در fmol) است. گراف این مجموعه داده با استفاده از bnlearn به دست آمده است و در شکل ۲.۳ نشان داده شده است.

۲.۳ پیاده‌سازی و ارزیابی

در این بخش، جزئیات ارزیابی تجربی ارائه شده است که شامل معیار ارزیابی، مدل‌های پایه، ویژگی‌های پیاده‌سازی و نتایج تجربی است. داده‌ها به تصادف به مجموعه آموزش (80%) و مجموعه آزمون (20%) تقسیم شده است و(20%) از مجموعه آموزش برای اعتبار سنجی درنظر گرفته شده است.
همانطور که در شکل ۲.۲ نشان داده شده است، DAGSurv یک CVAE است که از MLP ها به عنوان رمزگذار و رمزگشا تشکیل شده است. مدل دارای معماری DNN است و از جستجوی شبکه‌ای برای انجام یک جستجوی فراپارامتری گسترده در مورد تعداد لایه‌ها، تعداد واحدهای پنهان، تابع فعال‌سازی و نرخ یادگیری استفاده شده است. مقادیر فراپارامترهایی که برای به دست آوردن نتایج گزارش شده در این پژوهش استفاده شده است، در جدول ۲ آمده است. برآورد گشتاوری سازوار (آدام) به عنوان الگوریتم بهینه‌سازی گرادیان نزول انتخاب شد و کل ماژول با استفاده از pyTorch

کدگذاری شد. پس از پیاده‌سازی در DAG-GNN (یو و همکاران، ۲۰۱۹)، واریانس متغیر پنهان Σ_Zرا به عنوان I_{L+1} تنظیم شد که ماتریس همانی در ابعاد $L+1$ به دست آید. فقط μ_Z قابل آموزش در نظر گرفته شد، زیرا مشاهده شد که مقدار Σ_Z به دلیل عبارت توان، به ویژه در مجموعه داده‌هایی با مقادیر زمان تا رویداد بزرگتر، بسیار بزرگ می‌شود. توجه داشته باشید که با وجود این اصلاح، نتایج بی‌تأثیر باقی می‌مانند.

۱.۲.۳ سنجه ارزیابی

از شاخص هماهنگی زمان‌ـوابسته (CI) به عنوان معیار ارزیابی به دلیل اینکه در برابر تغییرات در خطر بقا در طول زمان پایدار است، استفاده شده است. این شاخص از نظر ریاضی به صورت زیر تعریف می‌شود:

(۲.۳)

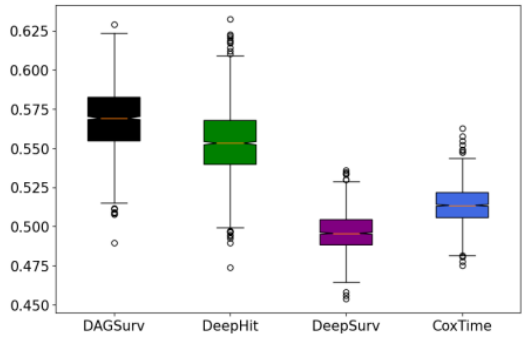
$$C_{td} = P(F(t^{(i)}|x^{(i)}) > F(t^{(i)}|x^{(j)})|t^{(i)} < t^{(j)}) \approx \frac{\sum_{i \neq j} R_{i,h} 1(F(t^{(i)}|x^{(i)}) > F(t^{(j)}|x^{(j)}))}{\sum_{i \neq j} R_{i.j}}$$

که در آن $1(.)$ تابع نشانگر است و $R_{i.j} \triangleq 1(t^{(i)} < t^{(j)})$. به عبارت دیگر، از یک براورد تجربی وابسته به زمان (CI) به عنوان متر استفاده شده است (لی و همکاران، ۲۰۱۸). برای آزمون استواری مدل‌های آموزش‌دیده روی داده‌های دیده نشده، بوت استرپ روی مجموعه آزمون اجرا شده است. با استفاده از مقادیر C_{td} بوت استرپ به‌دست‌آمده در مجموعه آزمون، نمودارهای جعبه‌ای رسم شده است و روش‌های پیشنهادی و پایه مقایسه شده‌اند. (۹۵%) فاصله اطمینان (CI) در اطراف میانه نیز ارائه شده است که می‌تواند به عنوان $median \pm 1.57IQRb$ محاسبه شود، که در آن IQR فاصله میان چارکی است و شامل ۵۰% از داده‌ها است و b نشان‌دهنده تعداد نمونه‌های بوت استرپ است.

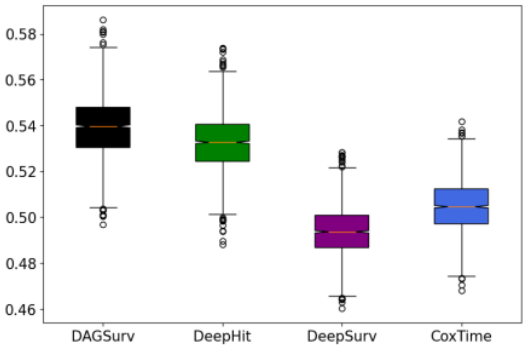
۲.۲.۳ مدل پایه

در این بخش، رویکردهای پایه زیر برای پیش‌بینی بقا مورد بحث قرار می‌گیرد که DAGSurv پیشنهادی با آن مقایسه می‌شود:

- CoxTime: یک Cox-PH کلاسیک و یکی از اساسی‌ترین رویکردهای پایه برای مقایسه است. در حالیکه فرض PH برای این مدل‌ها ضروری است، آنها امکان تفسیر و رتبه‌بندی آسان عوامل خطر را فراهم می‌کنند. در این پژوهش از CoxTime (کوام و همکاران، ۲۰۱۹) استفاده شده است که یک مدل شبکه عصبی با ریسک نسبی است که رگرسیون کاکس را فراتر از خطرات نسبی خطی گسترش می‌دهد.
- DeepSurv: یک توسعه DNN از مدل کلاسیک Cox-PH است. به طور کلی بهتر از مدل Cox-PH عمل می‌کند زیرا ساختاری غیرخطی را در نظر می‌گیرد که می‌تواند در زمینه مجموعه داده‌های واقعی مهم باشد.



شکل ۳.۳: نمودار جعبه‌ای: C_{td} مجموعه‌داده‌های Synthetic-small



شکل ۴.۳: نمودار جعبه‌ای: C_{td} مجموعه‌داده‌های Synthetic-large

- DeepHit: برخلاف الگوریتم‌های پیش‌بینی خطر بقا مانند DeepSurv/Cox، مستقیماً زمان تا رویداد را پیش‌بینی می‌کند. علاوه بر این، Deephit ذاتاً مبتنی بر فرض PH نیست، و از این رو، رویکرد پایه مهم برای مقایسه با آن است.

۳.۳ نتایج تجربی

در این بخش، مقادیر CI زمان‌ـوابسته (C_{td})، همراه با فواصل اطمینان (95%) را با استفاده از جداول و نمودارهای جعبه‌ای نشان می‌دهیم.

جدول ۳.۳: C_{td} مجموعه‌داده‌های Synthetic-small و Synthetic-large

Synthetic-large		Synthetic-small	
الگوریتم‌ها	$C_{td}(\%95CI)$	الگوریتم‌ها	$C_{td}(\%95CI)$
DAGSurv	0.5692 ± 0.0009	DAGSurv	0.5396 ± 0.0006
DeepHit	0.5532 ± 0.0009	DeepHit	0.5326 ± 0.0005
DeepSurv	0.4956 ± 0.0005	DeepSurv	0.4936 ± 0.0004
CoxTime	0.5134 ± 0.0005	CoxTime	0.5045 ± 0.0005

جدول ۴.۳: C_{td} مجموعه‌داده‌های METABRIC و GBSGS

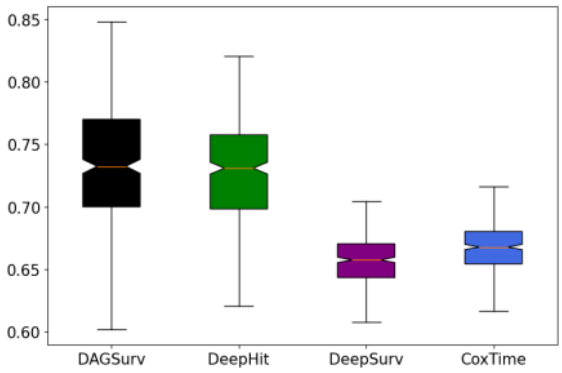
GBSGS		METABRIC	
الگوریتم‌ها	$C_{td}(\%95CI)$	الگوریتم‌ها	$C_{td}(\%95CI)$
DAGSurv	0.7323 ± 0.0056	DAGSurv	0.6892 ± 0.0023
DeepHit	0.7309 ± 0.0047	DeepHit	0.6602 ± 0.0026
DeepSurv	0.6575 ± 0.0021	DeepSurv	0.6651 ± 0.0020
CoxTime	0.6679 ± 0.0020	CoxTime	0.6687 ± 0.0019

۱.۳.۳ داده‌های ساختگی

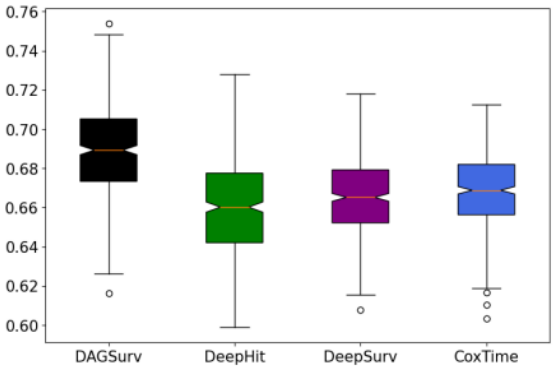
در این بخش، نتایج به‌دست‌آمده با استفاده از روش‌های پیشنهادی و پایه را بر روی مجموعه داده‌های ساختگی کوچک و بزرگی که در بخش قبل تعریف کردیم، ارائه می‌کنیم. مشاهده می‌شود که یادگیری مدل پایه برای اکثر مدل‌ها دشوار است و مقادیرC_{td} که در مجموعه آزمون اندازه‌گیری می‌شوند پایین هستند. از جدول ۳.۳ می‌توان مشاهده کرد که Deepsurv و CoxTime موفق به یادگیری یک مدل معنی دار نمی‌شوند و مقادیر C_{td} آنها نزدیک به ۰.۵ است. با فرضیات مبتنی بر مدل کمتر، DeepHit و DAGSurv قادر به یادگیری مدل با C_{td} قابل قبول هستند. توجه داشته باشید که ما از قیدی در شاخص تطابق مانند Deephit استفاده نمی‌کنیم. به طور کلی محاسبه این قید برای مجموعه داده‌های بزرگ دشوار است، زیرا به محاسبات زوجی نیاز دارد. دانش DAG ورودی به DAGSurv کمک می‌کند تا در غیاب قید بهتر از DeepHit عمل کند. همانطور که انتظار می‌رفت، نمودار جعبه تغییرات کمتری را در مقادیر C_{td} نسبت به مجموعه آزمون نشان می‌دهد، زیرا در مورد داده‌های مصنوعی، نمونه‌های آزمون و آموزش از یک فرآیند تولید داده می‌آیند.

۲.۳.۳ مجموعه داده واقعی

در این بخش، عملکرد رویکرد پیشنهادی و طرح‌های پایه در مجموعه داده‌های واقعی که قبلاً توضیح دادیم، نشان داده می‌شود. مشاهده می کنیم که DAGSurv به طور مداوم در مقایسه با



شکل ۵.۳: نمودار جعبه‌ای: C_{td} مجموعه‌داده‌های METABRIC



شکل ۶.۳: نمودار جعبه‌ای: C_{td} مجموعه‌داده‌های GBSGS

طرح‌های پایه بهتر عمل می‌کند یا به همان اندازه رقابتی است.

علاوه بر بهبود عملکرد، DAGSurv تفسیرپذیری بهتری نیز دارد. اول از همه، امتیاز هماهنگی به عنوان اعتبارسنجی برای گراف ورودی در نظر گرفته می‌شود. به عنوان مثال، اگر C_{td} با در نظر گرفتن $A = 0$ در DAGSurv بهبود یابد، به این معنی است که گراف به مدل های ML بهتر برای تحلیل بقا کمک نمی‌کند. علاوه بر این، به ایجاد رابطه بین متغیرهای کمکی و نتیجه کمک می‌کند. به عنوان مثال، ما از گراف مربوط به مجموعه داده GBSG در شکل ۲.۳ مشاهده می‌کنیم که درجه تومور بر هر دو، تعداد گره‌های لنفاوی مثبت و همچنین زمان تا رویداد (مرگ) تأثیر می گذارد. از این رو، رابطه بین تعداد غدد لنفاوی مثبت و زمان بقا، باید درجه تومور را در نظر بگیرد. چنین نتایج تفسیرپذیر ابزارهای قدرتمندی برای متخصصان و پزشکان هستند، و ما قصد داریم جنبه‌های هوش مصنوعی تفسیرپذیر را در تمام کارهای آینده خود بررسی کنیم.

۴.۳ نتیجه‌گیری

در این پژوهش، روش DAGSurv پیشنهاد شده‌است که در آن از دانش DAG علّی استفاده شده‌است و یک چارچوب جدید CVAE برای تحلیل بقا طراحی شده‌است. با استفاده از کدگذاری منبع، ثابت شد که دانش DAG منجر به کاهش آنتروپی در مقایسه با منبعی می‌شود که در آن متغیرهای آماری مستقل در نظر گرفته شده است. از CVAE به عنوان یک رمزگذار منبع برای دستیابی به نمایش کارآمد داده استفاده شده است. اما، CVAE انتخاب بهینه‌ای نیست و طراحی رمزگذار بهینه منبع برای کارهای آینده پیشنهاد می‌شود.

با استفاده از مجموعه داده‌های ساختگی و واقعی، نشان دادیم که DAGSurv از نظر شاخص هماهنگی عملکرد بهبود یافته‌ای دارد و همچنین تفسیرپذیرتر نیز است. استفاده از این روش نیاز به دانش DAG دارد که عموماً مشخص نیست. در غیاب دانش کارشناسان، نشان داده شد که می‌توان از یکی از چندین الگوریتم موجود برای به دست آوردن یک DAG از یک مجموعه داده مشخص استفاده شود. برخلاف CoxTime ، DeepSurv و DAGSurv را می‌توان در حضور خطرات متغیر با زمان استفاده کرد. علاوه بر این، DAGSurv می تواند برای اعتبار سنجی روابط علی در هر مدل گرافیکی استفاده شود.

بسط و توسعهٔ تحلیل ارائه شده در این پژوهش به ریسک چندگانه می‌تواند جز کارهای آتی باشد. برخی توسعه‌های دیگر شامل تحلیل رویدادهای تکراری یا بازگشت‌پذیر (گوپتا و همکاران، ۲۰۱۹) است.

واژه‌نامه فارسی به انگلیسی

آ

استقلال شرطی ..... conditional independence

ب

برآورد ..... estimation
برآورد تجربی ..... empirical estimat
بدون وزن ..... unweighted
بدون جهت ..... undirected

پ

پیش‌بینی ..... prediction

ت

تابع بقا ..... survival function
تابع واگرا ..... divergence function
تابع خطر ..... hazard function
تابع هزینه ..... cost function
تحلیل بقا ..... survival analysis
توزیع پیشین ..... prior distribution

توزیع پسین ..... posterior distribution

ج

جهت‌دار ..... directed

خ

خطرهای متناسب کاکس ..... cox proportional hazards
خودرمزگذار تغییراتی شرطی ..... conditional vriational autoencoder
خودرمزگذار ..... autoencoder

د

دور ..... acycle

ر

رأس آغازین ..... initial vertex
رأس پایانی ..... terminal vertex
رأس‌ها ..... vertices
رمزگذار ..... encoder
رمزگشا ..... decoder

ز

زمان بقا ..... time survival
زمان‌ـوابسته ..... time-dependant

ش

شاخص هماهنگی ..... concordance index
شبکه عصبی عمیق ..... deep neural networks

شبکه‌های پیچشی گراف ..... Graph convolutional networks
شکست ..... failure

ک

کاپلان مایر ..... kaplan meier
کمان ..... arc

گ

گراف ..... graph
گراف استقلال شرطی ..... conditional independence graph
گراف غیرمدور جهت‌دار ..... directed acyclic graph

م

ماتریس افزوده ..... augment matrix
مدل‌های احتمالاتی ..... probabilistic model
متغیر پنهان ..... latent variable
متغیر کمکی ..... covariate

ن

نیاکان ..... ancestor
نرخ شکست شرطی ..... conditional failure rate
نسبت خطر ..... hazard ratio

و

وزن‌دار ..... weighted

ه

واژه‌نامه انگلیسی به فارسی

A

adjacency ..... مجاورت، همسایگی
ancestor ..... نیاکان
arc ..... کمان
augment matrix ..... ماتریس افزوده
autoencoder ..... خودرمزگذار

C

censored ..... سانسورشده
concordance ..... هماهنگی
conditional independence ..... استقلال شرطی
conditional independence graph ..... گراف استقلال شرطی
conditional failure rate ..... نرخ شکست شرطی
covaraite ..... متغیر کمکی
cox proportional hazards ..... خطرهای متناسب کاکس
conditional variational autoencoder(CAVE) ..... خودرمزگذار تغییراتی شرطی
concordance index ..... شاخص هماهنگی
cycle ..... دور
cost function ..... تابع زیان

D

decoder ..... کدگشا، رمزگشا

descendent ..... نوادگان

deep neural networks(DNN) ..... شبکه‌های عصبی عمیق

directed ..... جهت‌دار، باسو

directed acyclic graph ..... گراف غیرمدور جهت‌دار

divergence function ..... تابع واگرا

E

encoder ..... کدگذار، رمزگذار

edge ..... یال

empirical estimate ..... برآورد تجربی

estimate ..... برآورد

F

failure ..... شکست

failure rate ..... نرخ شکست

G

graph ..... گراف

Graphical model ..... مدل گرافیکی

Graphical convolution networks (GCN) ..... شبکه‌های پیچشی گرافی

H

hazard function ..... تابع خطر

hazard ratio ..... نسبت خطر

I

initial vertex ..... راس آغازین

K

kaplan meier ..... کاپلان مایر

L

latent variable ..... متغیر پنهان

M

maximum likelihood ..... ماکسیمم درستنمایی

N

node ..... گره

P

prediction ..... پیش‌بینی

prior distribution ..... توزیع پیشین

posterior distribution ..... توزیع پسین

probabilistic model ..... مدل‌های احتمالاتی

pre-processing ..... پیش پردازش

S

Semi supervised ..... نیمه نظارت‌شده

Survival analysis ..... تحلیل بقا

Survival function ..... تابع بقا

Survival time ..... زمان بقا

T

terminal vertex ..... راس پایانی

time-dependant ..... زمان‌ـوابسته

U

undirected ..... بدون جهت

unweighted ..... بدون وزن

V

vertices ..... رأس‌ها

W

weighted ..... وزن‌دار

کتاب‌نامه

[1] Anant Vickram Jaitha. (2017, April). An Introduction to the Theory and Applications of Bayesian Networks.

[2] David G. Kleinbaum Mitchel Klein. Survival Analysis A Self-Learning Text.Third Edition.

[3] Sharma, A. K., Kukreja, R., Prasad, R., and Rao, S. (2021, November). Dagsurv: Directed ayclic graph based survival analysis using deep neural networks. In Asian Conference on Machine Learning (pp. 1065-1080). PMLR.

[4] Gebski V. 1997. Analysis of Censored and Correlated Data (ACCORD). Data Analysis and Research Technologies, Eastwood, NSW, Australia.

[5] Berkson J. and Gage R.P. 1952. Survival curve for cancer patients following treatment. J. Amer. Statist. Assoc.

[6] Antonio G Marques, Negar Kiyavash, Jos´e MF Moura, Dimitri Van De Ville, and Rebecca Willett. Graph signal processing: Foundations and emerging directions [from the guest editors]. IEEE Signal Processing Magazine, 37(6):11–13, 2020.

[7] Sachin Kumar, Garima Gupta, Ranjitha Prasad, Arnab Chatterjee, Lovekesh Vig, and Gautam Shroff. Camta: Casual attention model for multi-touch attribution. DMS Workshop, ICDM, 2020.

[8] Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In Proc. AAAI, 2018.

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. ICLR, 2017.

[10] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1):11– 33, 2015.

[11] Daphne Koller and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

[12] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proc. ICML, volume 97, pages 7154–7163. PMLR, 2019.

[13] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282):457–481, 1958.

[14] David Roxbee Cox. Analysis of survival data. Routledge, 2018.

[15] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR), 51(6):1–36, 2019.

[16] David Faraggi and Richard Simon. A neural network model for survival data. Statistics in medicine, 14(1):73–82, 1995.

[17] Sunayan et. al Bandyopadhyay. Data mining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data. Data Mining and Knowledge Discovery, 29(4):1033–1069, 2015.

[18] Donglin Di, Shengrui Li, Jun Zhang, and Yue Gao. Ranking-based survival prediction on histopathological whole-slide images. In MICCAI, pages 428–438. Springer, 2020.

[19] Thomas M Cover. Elements of information theory. John Wiley and Sons, 1999.

[20] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. Foundations and Trends in Machine Learning, 12(4):307–392, 2019.

[21] Christopher M Bishop. Pattern recognition and machine learning. Springer, 2006.

[22] M Schumacher, G Bastert, H Bojar, K Huebner, M Olschewski, W Sauerbrei, C Schmoor, C Beyerle, RL Neumann, and HF Rauschecker. Randomized 2x2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. Journal of Clinical Oncology, 12(10):2086–2093, 1994.

Abstract

Causal structures for observational survival data provide crucial information regarding the relationships between covariates and time-to-event. We derive motivation from the information theoretic source coding argument, and show that incorporating the knowledge of the directed acyclic graph (DAG) can be beneficial if suitable source encoders are employed. As a possible source encoder in this context, we derive a variational inference based conditional variational autoencoder for causal structured survival prediction, which we refer to as DAGSurv. We illustrate the performance of DAGSurv on low and high-dimensional synthetic datasets, and real-world datasets such as METABRIC and GBSG. We demonstrate that the proposed method outperforms other survival analysis baselines such as Cox Proportional Hazards, DeepSurv and Deephit, which are oblivious to the underlying causal relationship between data entities.

College of Science
School of Mathematics, Statistics, and Computer Science

Survival analysis based on DAG

Fateme Fartash Asl

Supervisor: Zahra Rezaei Ghahroodi

A thesis submitted in partial fulfillment of the requirements for
the degree of B.Sc. in Statistics

Aug 2023