پانزدهمین سمینار احتمال
و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان

15th Seminar On Probability
and Stochastic Processes
August 30-31, 2025
University of Kurdistan

عنوان مقاله

فاطمه محمدی [۱] ، محمد آرشی[۱] آرزو حبیبی‌راد [۱] ابوالفضل محمدزاده مقدم [۲]

[۱]گروه آمار دانشگاه فردوسی مشهد

[۲]گروه عمران دانشگاه فردوسی مشهد

چکیده: حوادث ترافیکی یکی از چالش‌های اصلی در مدیریت ترافیک شهری هستند که می‌توانند تاثیرات منفی بر ایمنی و کارایی سیستم‌های حمل و نقل داشته باشند. یکی از آثار منفی حوادث ترافیکی، ایجاد ترافیک و تاخیر در عبور و مرور است. اقدامات نادرست آژانس‌های مدیریت ترافیک در برخورد با تصادفات می‌تواند منجر به تاخیر در ترافیک و خطر تصادف شود. آژانس‌های حمل و نقل برای مدیریت حوادث ترافیکی و به کارگیری استراتژی‌های مناسب برای کاهش زمان حادثه و به دنبال آن کاهش ازدحام ترافیکی نیاز به پیش‌بینی مدت زمان حادثه دارند. داده‌های زمان حوادث ترافیکی داده‌های زمان تا رویداد هستند. یکی از ویژگی‌های این داده‌ها وابستگی به مدت زمان است. یکی از مدل‌های رایج برای در نظر گرفتن این ویژگی، مدل‌های نرخ خطر مبنا است. مدل‌های نرخ خطر مبنا شامل مدل‌های زمان شکست شتابیده و مدل خطرات متناسب پارامتری و نیمه پارامتری کاکس می باشد. عملکرد مدل‌های نرخ خطر مبنا در پیش‌بینی زمان حوادث ترافیکی، با توجه به فرضیاتی که دارند ممکن است دارای محدودیت‌هایی باشد.از جمله این فرضیات می‌توان به فرض خطرات متناسب، در نظر گرفتن توزیع خاص برای داده‌ها، اثر ثابت با زمان عوامل تاثیرگذار و رابطه خطی بین متغیرهای مستقل و وابسته، اشاره کرد و ممکن است برخی از این فرضیات در داده‌های حوادث ترافیکی وجود نداشته باشد. یکی دیگر از روش‌های رایج برای مدل‌بندی زمان حوادث ترافیکی مدل‌های یادگیری ماشین می‌باشد. نقطه قوت این مدل‌ها این است که برخی از فرضیات مشکل‌ساز مدل‌های نرخ خطر مبنا را ندارند. یکی از روش‌های یادگیری ماشین، شبکه‌های عصبی عمیق می‌باشد. اگر هدف این مدل‌ها پیش‌بینی مدت زمان پاکسازی حادثه باشد، آن ها نمی‌توانند وابستگی به مدت زمان را در داده‌های حوادث ترافیکی در نظر بگیرند. لذا از مدل‌های شبکه‌های عصبی عمیق با هدف برآورد تابع بقا استفاده می‌کنیم. اما این مدل‌ها هم نمی‌توانند به صراحت مدت زمان یک حادثه ترافیکی را پیش‌بینی کنند. بنابراین در این مطالعه علاقه‌مندیم که مدل یادگیری عمیق مبتنی بر تحلیل بقا طراحی کنیم که هم تابع بقا را برآورد کند و هم زمان حوادث ترافیکی را پیش‌بینی کند. مدل پیشنهادی می‌تواند وابستگی مدت زمان را هنگام پیش‌بینی مدت زمان حادثه با برآورد همزمان تابع بقا بر اساس مفهوم یادگیری چند وظیفه‌ای در نظر بگیرد.

واژه‌های کلیدی:تحلیل بقای عمیق، مدل‌های نرخ خطر مبنا، زمان پاکسازی حادثه، تابع بقا، یادگیری چند وظیفه‌ای

---

[۱] فاطمه محمدی، fm357016@gmail.com

کد موضوع‌بندی ریاضی (۲۰۲۰): xxAxx، xxBxx، xxCxx.

۱ مقدمه

حوادث ترافیکی تاثیرات منفی بر ایمنی و کارایی سیستم‌های حمل و نقل دارند. از جمله این تاثیرات می‌توان به ایجاد ترافیک و تأخیر در عبور و مرور، آسیب‌ها و تلفات جانی و بروز حوادث ثانویه اشاره کرد. بنابراین به عنوان یکی از چالش‌های اصلی در مدیریت ترافیک محسوب می‌شوند. سازمان‌های مدیریت ترافیک برای به‌کارگیری استراتژی‌های مناسب و کاهش این اثرات منفی نیاز به پیش‌بینی مدت زمان حادثه دارند. مدت زمان حادثه را می‌توان به چهار مرحله تقسیم کرد: زمان تشخیص، زمان پاسخ، زمان پاکسازی و زمان بازیابی. زمان پاکسازی که زمان ورود بین یک تیم پاسخ و پاکسازی کامل حادثه است طولانی‌ترین مرحله است. یکی از اهداف مدل‌بندی داده‌های حوادث ترافیکی، پیش‌بینی زمان پاکسازی حادثه است. داده‌های زمان پاکسازی حادثه داده‌های مدت زمان هستند. یکی از ویژگی‌های این داده‌ها، وابستگی به مدت زمان است. برای در نظر گرفتن این ویژگی عمدتاً از مدل‌های آماری نرخ خطر مبنا استفاده شده است . نام و منرینگ (۲۰۰۰)؛ جونز و همکاران (۱۹۹۱)؛ چانگ (۲۰۱۰) .نام و منرینگ (۲۰۰۰) مدت زمان حادثه را با استفاده از یک مدل خطر متناسب پارامتری که متغیر وابسته آن نرخ خطر است، تحلیل کردند. آنها گزارش داده‌اند که پارامترهای براورد شده و عملکرد پیش‌بینی مدل زمان شکست شتابیده (AFT) در طول سال‌های مختلف پایدار است. مدل خطرات متناسب کاکس دارای فرض خطرات متناسب هستند که این فرض ممکن است برای همه متغیرهای مستقل برقرار نباشد. در مطالعات قبلی برای پیش‌بینی مدت زمان حادثه از مدل‌های زمان شکست شتابیده (AFT) استفاده شده است. جونز و همکاران (۱۹۹۱) مدل زمان شکست شتابیده (AFT) که متغیر وابسته آن میانگین مدت زمان حادثه بود برای مدل‌بندی داده‌های حوادث ترافیکی به کار بردند. چانگ (۲۰۱۰)مدل زمان شکست شتابیده را برای مجموعه داده‌های حادثه در مقیاس بزرگ در بزرگراه اعمال کرد توسلی حجتی و همکاران (۲۰۱۳) ناهمگنی مشاهده نشده را با اعمال پارامترهای تصادفی در مدل زمان شکست شتابیده در نظر گرفت. مدل زمان شکست شتابیده (AFT) یک مدل پرکاربرد است که اثرات ثابت با زمان عوامل تاثیرگذار را فرض می‌کند. همچنین یکی دیگر از فرضیات این مدل‌ها رابطه خطی متغیرهای مستقل با لگاریتم میانگین زمان پاکسازی است. این دو فرض ممکن است برای داده‌های حوادث ترافیکی و در یک آزادراه برقرار نباشد. بنابراین استفاده از این مدل‌ها دارای محدودیت‌هایی می‌باشد. یکی دیگر از روش‌های رایج برای مدل‌سازی زمان پاکسازی حادثه مدل‌های یادگیری ماشین از جمله مدل جنگل تصادفی، ماشین گرادیان تقویتی و شبکه‌های عصبی مصنوعی است. با این حال مدل‌های یادگیری ماشین میانگین زمان پاکسازی را به عنوان متغیر وابسته در نظر می‌گیرند و ویژگی وابستگی به مدت زمان را نادیده می‌گیرند. مطالعه‌ای اخیر گزارش داده است که عملکرد مدل‌های یادگیری ماشین قابل مقایسه یا بدتر از مدل‌های زمان شکست شتابیده است . وی و همکاران (۲۰۰۷) برای پیش‌بینی مدت زمان حادثه از پیش‌بینی متوالی استفاده کردند که عملکرد مدل‌های شبکه های عصبی مصنوعی را بهبود بخشید. حمد و همکاران (۲۰۲۰) مدل جنگل تصادفی (RF) را برای پیش‌بینی زمان پاکسازی حادثه توسعه دادند. همچنین ما و همکاران (۲۰۱۷) یک مدل GBM را برای پیش‌بینی زمان پاکسازی حادثه و همچنین تحلیل عوامل تاثیرگذار توسعه دادند. نتایج آن‌ها عملکرد پیش‌بینی بهتر GBM را نسبت به RF وANN نشان داد. لی و همکاران (۲۰۲۰) پیش‌بینی مدت زمان حادثه را با استفاده از شبکه عصبی عمیق انجام داد. نتایج آن‌ها نشان داد که DNN بهتر از مدل‌های یادگیری ماشین معمولی با

هدف پیش‌بینی میانگین زمان حادثه، عمل می‌کند. مطالعه اخیر تانگ و همکاران (۲۰۲۰) گزارش داد که عملکرد پیش‌بینی مدل‌های یادگیری ماشین قابل مقایسه یا بدتر از مدل‌های زمان شکست شتابیده (AFT) با توزیع های فرضی مختلف بوده است. مطالعات قبلی از مدل‌های یادگیری ماشین و شبکه های عصبی عمیق به عنوان رگرسیون برای زمان پاکسازی حادثه استفاده می‌کردند. اگر هدف این مدل‌ها پیش بینی زمان پاکسازی حادثه باشد نمی‌توانند وابستگی به مدت زمان را که یکی از ویژگی‌های مهم داده‌های زمان تا رویداد است به طور مستقیم در نظر بگیرند. این مدل‌ها تنها یک وظیفه انجام می‌دهند و آن هم پیش‌بینی میانگین زمان پاکسازی حادثه است. اگر مدل‌های یادگیری ماشین و شبکه‌های عصبی عمیق[1] با هدف برآورد تابع بقا طراحی شوند می‌توانند مدل‌های مناسبی برای تحلیل داده‌های زمان تا رویداد باشند زیرا می‌توانند ویژگی وابستگی به مدت زمان در این داده‌ها را در نظر بگیرند. ایشواران و همکاران (۲۰۰۸) با توسعه جنگل بقای تصادفی که مدل جنگل تصادفی را برای داده‌های زمان تا رویداد گسترش می‌دهد، تابع بقا را برآورد کند. هوثورن و همکاران (۲۰۰۶) مدل ماشین تقویتی گرادیان اصلاح شده[2] را پیشنهاد کردند. کاتزمن و همکاران (۲۰۱۸) Deepsurv را ابداع کردند که با استفاده از شبکه‌های عصبی عمیق روابط غیرخطی و تعاملی را در نظر بگیرد و تابع بقا را برآورد کند. این مدل از مدل‌های خطرات متناسب کاکس و جنگل بقای تصادفی بهتر عمل کرد. لی و همکاران (۲۰۱۸) DeepHitرا پیشنهاد کردند که به طور مستقیم تابع بقا را با استفاده از شبکه‌های عصبی عمیق برآورد می‌کند. با اینکه مدل‌های یادگیری ماشین با هدف براورد تابع بقا می‌توانند وابستگی به مدت زمان را در نظر بگیرند اما به صراحت نمی‌توانند زمان پاکسازی حادثه را پیش‌بینی کنند. در ادامه مدلی ارائه می‌شود که می‌تواند به طور همزمان هم تابع بقا را برآورد کند و هم مدت زمان حادثه را با استفاده از ساختار یادگیری چند وظیفه‌ای پیش‌بینی کند.

۲ تعاریف اولیه

داده‌های زمان پاکسازی حادثه داده‌های مدت زمان هستند که زمانی است تا وقوع رویداد. منظور از رویداد پاکسازی کامل حادثه است. برخی از مفاهیم مورد نیاز در تحلیل داده‌های زمان تا رویداد شامل تابع بقا، تابع نرخ خطر، داده سانسورشده، ویژگی وابستگی به مدت زمان و میانگین طول عمر می‌باشد. در ابتدا لازم است این مفاهیم توضیح داده شود.

تعریف ۱.۲. *تابع بقا*[3] *احتمال این است که زمان بقا بیش از یک زمان مشخص باشد. این تابع، که به نام تابع قابلیت اطمینان نیز مشهور است احتمال زنده ماندن در طول زمان t را محاسبه می‌کند. عموماً این تابع را با استفاده از روش‌های ناپارامتری مانند کاپلان مه یر یا نلسون آلن و یا روش های پارامتریک و از روی تابع توزیع به دست می‌آورند.*

$$S(t) = Pr(T < t) = ۱ - F(x)$$

تعریف ۲.۲. *تابع نرخ خطر*[4] *میزان مرگ و میر لحظه‌ای را نشان می‌دهد. این تابع شانس رخداد پیشامد را در لحظه t با شرط آن‌که واحد آزمایشی تا این لحظه زنده باشد، بیان می‌کند.*

- *اگر T متغیر تصادفی گسسته باشد:*

$$h(t) = Pr(T = t | t \leq T)$$

[1] Deep Neural Network
[2] modified Gradient Boosting Machine
[3] Survival Function
[4] Hazard Function

- *اگر T متغیر تصادفی پیوسته باشد:*

$$\lim_{\Delta t \to ۰} \frac{Pr(t < T < t + \Delta t)}{\Delta t}$$

با استفاده از تابع نرخ خطر می‌توان تابع بقا را برآورد کرد.

$$S(t) = exp(-\int_{۰}^{t} h(z)\,\mathrm{dz}) = exp(-H(t))$$

تعریف ۳.۲. *درکنار داده‌های زمان تا رویداد، وضعیت آن داده هم ثبت می‌شود. منظور از وضعیت داده این است که آیا داده سانسور شده است یا داده کامل است. منظور از داده کامل این است که اطلاعات داده از اول مطالعه تا زمانی که رویداد برایش اتفاق می‌افتد موجود باشد. منظور از داده سانسورشده هم این است که اطلاعات کاملی از زمان دقیق وقوع رویداد برای برخی نمونه‌ها وجود ندارد. سانسورها می توانند انوا ع مختلفی داشته باشند:*

- *سانسور راست*

  زمانی‌که رویداد بعد از پایان مطالعه رخ دهد یا به عبارتی دیگر رویداد تا انتهای مطالعه رخ نمی‌دهد.

- *سانسور چپ*

  زمانی‌که رویداد قبل از شروع مطالعه رخ داده باشد.

- *سانسور فاصله ای*

  زمانی‌که رویداد بین دو بازه زمانی مشخص رخ داده باشد.

تعریف ۴.۲. *ویژگی وابستگی به مدت زمان به این معنی است که احتمال پایان پاکسازی حادثه به طول مدت زمانی که پاکسازی به طول انجامیده است بستگی دارد. بدین منظورکه در این نوع داده‌ها متغیر وابسته، زمان تا رویداد است. این زمان همواره مثبت است و در مطالعات تحلیل بقا می‌تواند به صورت پیوسته یا گسسته در نظرگرفته شود.*

تعریف ۵.۲. *مساحت زیر منحنی تابع بقا، میانگین طول عمرگفته می‌شود.*

$$E(T) = \int_{۰}^{\infty} S(u)\mathrm{du}$$

۳ مدل‌های تحلیل بقا

در آمار برای مدل‌بندی داده‌های زمان تا رویداد از مدل‌های تحلیل بقا استفاده می‌شود این مدل‌ها شامل مدل‌های نرخ خطر مبنا هستند. مدل‌های نرخ خطر مبنا قادر هستند که ویژگی وابستگی به مدت زمان را در نظر بگیرند. این مدل‌ها می‌توانند دو نوع متغیر وابسته ازجمله میانگین زمان رویداد و تابع نرخ خطر را در نظر بگیرند. مدل‌های نرخ خطر مبنا که متغیر وابسته آن‌ها تابع نرخ خطر است شامل مدل‌های خطرات متناسب کاکس[۵] است. همچنین مدل‌هایی که متغیر وابسته آن‌ها میانگین زمان بقا است شامل مدل‌های زمان

[5] Cox proportional hazard Model

شکست شتابیده (AFT) می‌باشد. در حالی‌که هدف مدل خطرات متناسب کاکس برآورد اثرات متغیرهای مستقل بر توابع نرخ خطر یا بقا است، هدف مدل زمان شکست شتابیده برآورد مستقیم اثرات بر روی مدت زمان (یا لگاریتم زمان)است.

- مدل خطرات متناسب کاکس

$$h(t,x) = h_{\circ}(t) \times exp(\beta_{\text{۱}} x_{\text{۱}} + \cdots + \beta_p x_p)$$

این مدل‌ها یک خطر متناسب را فرض می‌کنند که نشان می‌دهد اثر متغیرهای مستقل ثابت از زمان است. منظور از فرض خطرات متناسب این است که نسبت مخاطره برای واحد ام‌i به واحد j ام مستقل از زمان باشد. نرخ خطر پایه مقدار خطر را زمانی‌که همه متغیرهای مستقل صفر هستند، نشان می‌دهد. خطر به عنوان یک تابع ضربی از یک خطر پایه و یک تابعی از متغیرهای مستقل است. مدل کاکس یک مدل نیمه پارامتری است ولی می‌تواند پارامتری هم باشد.

- مدل زمان شکست شتابیده

این مدل یک رابطه خطی بین لگاریتم زمان بقا و متغیرهای مستقل را فرض می‌کند.

$$\ln t = \beta_{\circ} + \beta_{\text{۱}} x_{\text{۱}} + \cdots + \beta_p x_p + \epsilon$$

متغیر وابسته این مدل را می‌توان تابع نرخ خطر نیز در نظر گرفت:

$$h(t,x) = exp(-\underline{\beta} x) \cdot h_{\circ}(exp(-\underline{\beta} x)t)$$

می توان توزیع‌های فرضی وایبول، لگ نرمال و لگ لجستیک را در نظر گرفت که در نرخ خظر پایه تاثیرگذار هستند.

- مدل‌های شکنندگی

در برخی مطالعات تحلیل بقا، ممکن است زیر گروه‌های جامعه دارای یک متغیر تصادفی مشترک مشاهده نشده باشند که قادر به وارد کردن آنها در مطالعه نیستیم. اگر اثر این متغیرها در نظر گرفته نشود بخشی از تغییرات تابع نرخ خطر که می‌توانست با وجود این عامل توضیح داده شود کاهش می‌یابد و با جمله خطا جمع شده و در نتیجه باعث افزایش تغییرات تابع نرخ خطر نسبت به حالتی که عامل مورد نظر در مدل وجود داشته باشد، می‌گردد که باعث به دست آمدن براوردگرهای اریب برای پارامترهای مدل رگرسیونی می‌گردد. رویکرد پارامتر تصادفی (عامل شکنندگی) روشی است که به طور گسترده برای پرداختن به ناهمگنی مشاهده نشده پذیرفته شده است به طوری‌که به برخی از پارامترها اجازه می‌دهد تا در مشاهدات با توزیع فرضی، تغییر کنند.

$$\beta_n = \beta + \delta_n$$

این عامل شکنندگی را می‌توان هم در مدل‌های خطرات متناسب به کار برد و هم در مدل‌های زمان شکست شتابیده از آن استفاده کرد. اگر این عامل تصادفی در مدل‌های زمان شکست شتابیده به کار رود به نام مدل‌های (Random parameter AFT) شناخته شده و پیش‌بینی و تحلیل دقیق‌تری از عوامل تاثیرگذار مرتبط با ناهمگنی مشاهده نشده ارائه می‌کنند.

۴ مدل‌های یادگیری ماشین

یکی دیگر از روش‌های رایج برای مدل‌سازی داده‌های زمان پاکسازی حادثه، مدل‌های یادگیری ماشین می‌باشد. از جمله می‌توان به مدل‌های شبکه های عصبی مصنوعی (ANN)، جنگل تصادفی(RF) و ماشین گرادیان تقویتی(GBM) اشاره کرد. مدل‌های شبکه های عصبی عمیق(DNN) یک روند اجتناب ناپذیر در زمینه‌های مختلف تحقیقاتی می‌باشد.

۱.۴ جنگل بقای تصادفی

جنگل تصادفی[۶] یکی از الگوریتم‌های یادگیری ماشین مبتنی بر درخت تصمیم است که برای مسائل رگرسیون و طبقه‌بندی استفاده می‌شود. همچنین اگر مدل جنگل تصادفی با هدف برآورد تابع بقا طراحی شود[۷] می‌تواند برای داده‌های زمان تا رویداد مناسب باشد زیرا ویژگی وابستگی به مدت زمان را در این داده‌ها در نظر می‌گیرد. در این نوع الگوریتم‌ها از چندین مدل ضعیف استفاده می‌شود تا عملکرد مدل کلی را بهبود ببخشد. در مدل جنگل تصادفی از چندین درخت تصمیم استفاده می‌شود. همان‌طور که در جنگل تصادفی با هدف رگرسیون مدل‌های ضعیف درخت تصمیم به صورت تصادفی و مستقل آموزش می‌بینند در مدل جنگل بقای تصادفی (RSF) هم درخت‌ها به طور جداگانه با هدف برآورد تابع بقا آموزش می‌بینند. در مدل جنگل بقای تصادفی هر درخت تصمیم به جای پیش‌بینی یک مقدار عددی در گره اخر، تابع بقا را به روش‌های ناپارامتری مثل کاپلان مه یر برآورد می‌کند. پس از آموزش درخت‌ها، برای هر نمونه جدید پیش‌بینی تابع بقا به صورت میانگین از پیش‌بینی تمام درخت‌ها محاسبه می‌شود.

الگوریتم ۱. *فرض کنید داده‌ها به صورت زیر باشند:*

$$\mathcal{D} = \{(X_i, T_i, \delta_i)\}_{i=1}^{n}$$

که در آن: X_i: بردار ویژگی‌ها، T_i: زمان مشاهده‌شده (کمترین مقدار بین زمان وقوع و سانسور)، و $\delta_i \in \{0, 1\}$: نشانگر وقوع رویداد است. (۱=رخ داده، ۰=سانسور شده)

- *ساخت درخت‌ها*

۱. *یک نمونه بوت‌استرپ از داده‌ها بگیرید.*

۲. *در هر گره، m ویژگی به‌طور تصادفی انتخاب کنید.*

۳. *برای هر تقسیم، آماره آزمون لاگ-رتبه را محاسبه کنید.*

۴. *تقسیم با بیشترین مقدار آماره لاگ-رتبه را انتخاب کنید.*

۵. *تقسیم گره‌ها تا رسیدن به شرط توقف ادامه می‌یابد.*

- *برآورد در گره نهایی*

در هر گره نهایی، تابع بقا با استفاده از برآوردگر کاپلان-مایر محاسبه می‌شود:

[6] Random Forest

[7] Random Survival Forest

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(۱ - \frac{d_j}{r_j}\right)$$

که: d_j تعداد رخدادها در زمان t_j و r_j تعداد افراد در معرض خطر در زمان t_j ، می‌باشد.

- *پیش‌بینی نهایی جنگل*

میانگین برآوردهای بقا از تمام درخت‌ها به صورت زیر محاسبه می‌شود:

$$\hat{S}(t \mid X) = \frac{۱}{B} \sum_{b=۱}^{B} \hat{S}^{(b)}(t \mid X)$$

۲.۴ ماشین گرادیان تقویت‌شده

ماشین گرادیان تقویتی[۸] یکی از الگوریتم‌های یادگیری جمعی است . این مدل هم بر مبنا چندین درخت تصمیم است. تفاوت بین مدل‌های Adaboost و ماشین گرادیان تقویتی در نحوه به روزرسانی مدل و نحوه تعامل مدل‌های ضعیف است. Adaboost مدل‌های ضعیف درخت تصمیم را به صورت متوالی تقویت می‌کند و در هر مرحله به مواردی که اشتباه پیش‌بینی شده‌اند وزن‌هایی اختصاص می‌دهد تا در مدل بعدی بیش‌تر به آن‌ها توجه شود. در مدل ماشین گرادیان تقویتی به همه نمونه‌ها وزن‌های یکسانی اختصاص داده می‌شود و با استفاده از روش عددی گرادیان نزولی پیش‌بینی‌ها بهبود می‌بخشد. به عبارت دیگر در هر مرحله در مدل ماشین گرادیان تقویتی هدف بهبود عملکرد کلی مدل در پیش‌بینی است.

۳.۴ شبکه عصبی مصنوعی

شبکه‌های عصبی[۹] شامل سه نوع لایه از جمله لایه ورودی، لایه خروجی و لایه پنهان هستند. تعداد نورون‌ها در لایه ورودی برابر با تعداد ویژگی‌های داده‌های ورودی است و تعداد نورون‌ها در لایه خروجی بستگی به هدف پیش‌بینی دارد. تعداد نورون‌های لایه پنهان بین تعداد نورون‌های لایه ورودی و تعداد نورون‌های لایه خروجی قرار می‌گیرد. به‌عبارتی، تعداد نورون‌ها در لایه پنهان ابرپارامتری است که با آزمایش مقادیر مختلف تعیین می‌شود تا بهترین عملکرد آموزشی شبکه عصبی به دست آید.

آموزش شبکه عصبی شامل دو مرحله از پیش‌انتشار و پس‌انتشار است. پیش‌انتشار فرآیندی است که در آن داده‌های ورودی به لایه ورودی ارسال می‌شود، توسط لایه پنهان پردازش می‌شود و در نهایت به لایه خروجی منتقل می‌گردد. جزئیات پیش‌انتشار به‌صورت زیر است:

اگر شبکه عصبی با یک لایه پنهان و یک لایه خروجی داشته باشیم، رابطه بین لایه ورودی و لایه پنهان به‌صورت ریاضی به‌صورت زیر بیان می‌شود:

$$\text{neth}_۱ = \sum w_i x_i + b_۱$$

[8] Gradient Boosting Machine

[9] Neural Network

که در آن x_i بردار ورودی است، w_j بردارهای وزن بین لایه ورودی و لایه پنهان هستند، و b_1 بایاس بین لایه‌های محاسبه‌شده است که به‌طور انعطاف‌پذیری برای تنظیم عملکرد شبکه عصبی طراحی شده است. سپس، neth_1 از توابع فعالسازی از جمله ReLu عبور می‌کند.

$$h_1 = g(\text{neth}_1) = \max(0, \text{neth}_1)$$

Z به‌صورت زیر محاسبه می‌شود:

$$\text{Net}_z = \sum w_j h_i + b_2$$

و

$$Z = g(\text{Net}_z) = \max(0, \text{Net}_z)$$

که در آن w_j بردارهای وزن بین لایه پنهان و لایه خروجی هستند و b_2 بایاس بین لایه پنهان و لایه خروجی است.

اگر Z با خروجی مورد انتظار مطابقت نداشته باشد، خطای پیش‌بینی با استفاده از تابع ضرر انتخاب‌شده محاسبه خواهد شد. سپس، پس‌انتشار خطای پیش‌بینی را به نورون‌های شبکه باز می‌گرداند و از الگوریتم نزول گرادیان برای اصلاح وزن‌های نورون‌ها به‌منظور کاهش خطاهای پیش‌بینی استفاده می‌کند. هنگامی که خطای پیش‌بینی با انتظارات مطابقت داشت، پیش‌بینی شبکه عصبی انجام می‌شود.

۵ مدل تحلیل بقای عمیق چندوظیفه‌ای

در بحث شبکه‌های عصبی اگر چند وظیفه باهم مرتبط بودند می‌توانیم یک مدل شبکه عصبی طراحی کنیم که همزمان قادر به انجام دو وظیفه باشد. مدل مورد استفاده در این مطالعه یک شبکه عصبی عمیق (منظور تعداد زیاد لایه‌های پنهان) دو وظیفه ای است که می‌تواند ویژگی وابستگی به مدت زمان را با استفاده از اطلاعات مشترکی که از برآورد تابع بقا به دست می‌آورد در پیش‌بینی زمان پاکسازی حادثه در نظر بگیرد. در یادگیری چند وظیفه ای، چندین کار مرتبط به طور همزمان یاد می‌گیرند و از اطلاعات مشترک بین آن‌ها برای بهبود عملکرد کلی استفاده می‌شود. یادگیری چند وظیفه ای به جای آموزش یک مدل جداگانه برای هر کار، یک مدل واحد را برای انجام چندین کار آموزش می‌دهد. با ورود یک بردار متغیرهای مستقل به مدل، چندین خروجی (هدف یا متغیرهای وابسته) به دست می‌آوریم. در یادگیری چند وظیفه‌ای، برخی از لایه‌ها یا پارامترها در بین وظایف به اشتراک گذاشته می‌شوند و به مدل اجازه می‌دهند ویژگی‌های مشترکی را که برای همه وظایف مفید هستند، یاد بگیرد. این مدل به طور همزمان روی وظایف مختلف آموزش داده می‌شود و لایه‌های مشترک بر اساس زیان ترکیبی از همه وظایف به روز می‌شوند. علاوه بر لایه‌های مشترک، مدل‌های یادگیری چند وظیفه ای معمولا دارای لایه‌های اختصاصی مربوط به وظایف هستند که جنبه‌های منحصر به فرد هر وظیفه را مدیریت می‌کنند. از جمله مزایای یادگیری چند وظیفه‌ای این است که در صورت مرتبط بودن وظایف، عملکرد آنها با اطلاعات مشترک همدیگر بهبود می‌بخشد و مدل یادگیری چند وظیفه‌ای می‌تواند به عنوان یک تنظیم‌کننده عمل کند و از توجه بیش از حد مدل در یک وظیفه جلوگیری کند. همچنین از معایب ساختار یادگیری چند وظیفه‌ای می‌توان گفت با افزایش تعداد وظایف پیچیدگی و هزینه محاسباتی یادگیری چند وظیفه ای می‌تواند به طور قابل توجهی افزایش

یابد. نتایج نشان می‌دهند که انجام وظایف چندگانه به طور مکمل عملکرد وظایف مختلف را افزایش می دهد اگر اطلاعات نهفته ای که از مولفه‌های مشترک استخراج می‌شود به دو وظیفه مشابه کمک کند ساختار یادگیری چند وظیفه ای می‌تواند به طور موثر اطلاعات پنهان مشترک را جذب کند. وظیفه پیش‌بینی میانه زمان پاسخگویی بخشی از برآورد تابع بقا است بنابراین دو وظیفه مرتبط هستند. در این مطالعه به دنبال انجام دو وظیفه یعنی برآورد تابع بقا و پیش‌بینی میانگین زمان پاکسازی حادثه به طور همزمان هستیم. وظیفه برآورد تابع بقا مربوط به مدل‌های خطرات متناسب کاکس است. همچنین پیش‌بینی میانگین زمان پاکسازی حادثه وظیفه مدل‌های زمان شکست شتابیده است.

برای لایه اختصاصی مربوط به وظیفه براورد تابع بقا، این مطالعه ۱۰ دقیقه را به عنوان فاصله زمانی برای برآورد تابع بقا با توجه به کاربرد آن در مدیریت حادثه در نظر گرفته است. تابع بقا یک احتمال است، احتمال بقا در هر کدام از بازه‌های ۱۰ دقیقه ای محاسبه می‌شود. برای این منظور تابع فعالسازی لایه آخر سافت مکس[۱۰] در نظر گرفته می‌شود. برای وظیفه برآورد تابع بقا دو تابع زیان در نظر گرفته شده است:

- تابع زیان لگاریتم درستنمایی

در داده‌های زمان تا رویداد داده‌های سانسور شده وجود دارد. این داده‌ها اطلاعات کاملی درباره چگالی به ما نمی‌دهند و بنابراین در تابع درستنمایی باید به همان میزانی که به ما اطلاعات می‌دهند اثر داشته باشند بنابراین باید تابع ماسکی تعریف شود تا داده‌های سانسور شده را از بدون سانسورتشخیص دهد.

$$L_1 = -\sum_{t=1}^{n_o ut} y_t \cdot \log \widehat{y_t}$$

- تابع زیان رتبه بندی[۱۱]

این تابع زیان با استفاده از نرخ خطر تجمعی ترتیب وقوع رویدادها را بررسی می‌کند به طوری‌که حادثه‌هایی که رویداد برای آن‌ها زودتر اتفاق می‌افتد احتمال بقا کمتر و نرخ خطر بیش‌تری نسبت به حادثه‌هایی دارند که همان رویداد را دیرتر تجربه کردند. اگر مدلی بتواند این ترتیب‌ها را درست پیش‌بینی کند مثل این است که توانسته تابع بقا را دقیق برآورد کند. این تابع زیان برای بررسی این است که آیا مدل ترتیب درست وقوع رویدادها را پیش‌بینی کرده است یا خیر. هر چه این زیان کمتر باشد یعنی براورد تابع بقا دقیق‌تر است. با توجه به عملکرد این تابع زیان باید وقتی که می‌خواهیم این زیان را محاسبه کنیم تابع ماسکی تعریف کنیم تا برای داده‌هایی که در طی زمان اندازه‌گیری شدند هم بتواند ترتیب درست رویدادها را به درستی اعمال کند. این تابع زیان مانند شاخص تطابق وابسته به زمان (C-Index) عمل می کند.

$$L_2 = \sum_{i \neq j} A_i, j\eta(\hat{H}(t_i|x_{ti}), \hat{H}(t_i|x_{tj}))$$

$$\eta(x, y) = exp(\frac{y - x}{\sigma})$$

[10] Softmax

[11] Rank Loss

$$A_i, j = ۱ \; t_i < t_j$$

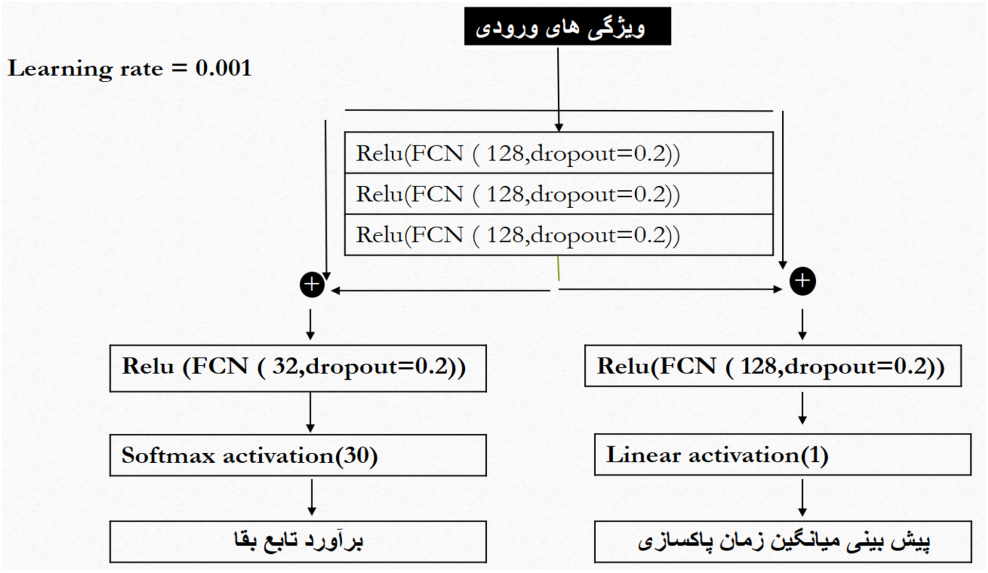
در لایه آخر شبکه عصبی مربوط به وظیفه پیش‌بینی میانگین زمان پاکسازی حادثه، از تابع فعالسازی خطی استفاده شده است. تابع زیان مربوط به آموزش این وظیفه مجموع میانگین مربعات خطا (MSE) است.

$$MSE = \frac{۱}{k}\sum_{k=۱}^{n}(T_k - \widehat{T_k})^۲$$

با تابع زیان ترکیبی زیر لایه‌های مشترک آموزش داده می‌شود:

$$L_{total} = L_۱ + w_۱ \cdot L_۲ + w_۲ \cdot L_۳$$

وزن‌های هر تابع زیان جزو ابر پارامتر هایی است که با استفاده از روش‌های جستجوی شبکه‌ای کالیبره می شوند. از الگوریتم بهینه سازی Adam برای بهینه کردن وزن‌های شبکه عصبی استفاده شده است. نرخ یادگیری در این مدل 0.001 در نظر گرفته شده است. ساختار شبیه‌سازی مدل در شکل ۱ نشان داده شده است.



شکل ۱: ساختار مدل تحلیل بقای عمیق چند وظیفه‌ای

۶ مثال کاربردی

داده‌های حوادث ترافیکی شامل ۳۹۷۶۰ حادثه با ویژگی‌های زمانی، جغرافیایی و حادثه است. داده‌ها مربوط به تصادف آزادراه‌های کشور کره جنوبی از سال ۲۰۱۴ تا ۲۰۱۹ می‌باشد. به طور تصادفی ۷۰ درصد از داده‌ها برای آموزش و اعتبارسنجی مدل و ۳۰ درصد از داده‌ها برای آزمون مدل استفاده شده است. اغلب برخی ویژگی‌های گمشده یا غیر واقعی در مجموعه داده‌ها وجود دارد که با روش‌های پیش‌پردازش داده‌ها برطرف می‌شوند. فقط از داده‌های کامل حادثه استفاده شده است که شامل تمام ویژگی‌های زمانی، جغرافیایی حادثه

و جریان ترافیک است با حذف نقاط پرت در نهایت ۱۶۹۲۹ داده برای توسعه مدل به کار گرفته شده است. تمام متغیرهای طبقه بندی در طول آموزش مدل به متغیرهای باینری تبدیل می‌شوند و متغیرهای مربوط به ترافیک پیوسته هستند و استاندارد می‌شوند. در این مثال، مدل زمان شکست شتابیده را با در نظر گرفتن توزیع‌های فرضی مختلف برای نرخ خطر پایه، به کار می گیرد .

مدل زمان شکست شتابیده (AFT)(توزیع وایبول)
مدل زمان شکست شتابیده (AFT)(توزیع لگ نرمال)
مدل زمان شکست شتابیده (AFT)(توزیع لگ لجستیک)

در این مطالعه مدل زمان شکست شتابیده با یک پارامتر تصادفی (RPAFT) برای در نظر گرفتن ناهمگنی مشاهده نشده به کار گرفته شد.از۲ مدل (GBM) و (RF) با دو هدف پیش‌بینی میانگین زمان پاکسازی و براورد تابع بقا برای بررسی عملکرد مدل هدف استفاده شده است. ابرپارامترهای اصلی با استفاده از تکنیک جستجوی شبکه‌ای (K-fold) تنظیم شده اند .

برای جنگل تصادفی (RF) چهار ابرپارامتر از جمله تعداد درخت، حداکثر عمق هر درخت، نرخ نمونه‌گیری برای آموزش هر درخت و تعداد متغیرهای مورد استفاده در هر گره تقسیم وجود دارد. برای ماشین گرادیان تقویتی (GBM) نیز چهار ابر پارامتر از جمله تعداد درخت، حداکثر عمق هر درخت، نرخ یادگیری و تعداد متغیرهای در نظرگرفته‌شده در هر درخت وجود دارد.

مدل هدف، جداگانه با هر کدام از مدل‌های تک وظیفه‌ای شبکه عصبی عمیق (DNN) مقایسه می‌شود تا نشان‌دهنده بهبود عملکرد یادگیری چند وظیفه‌ای باشد. ترکیب ابر پارامترها از جمله تعداد لایه‌ها و نورون‌های هر شبکه، نرخ یادگیری، نرخ dropout و وزن‌ها برای هر زیان به طور گسترده تنظیم شده‌اند تا معماری مدل نهایی را انتخاب کنند. میانگین خطای قدر مطلق (MAE) و میانگین درصد خطای قدر مطلق (MAPE) و شاخص تطابق وابسته به زمان (c-td) هر ترکیب از ابر پارامترها را ارزیابی می‌کنند و ترکیبی که عملکرد هماهنگ در این معیارها را به دست آورد، انتخاب می شود. عملکرد مدل برای پیش‌بینی میانگین مدت زمان رویداد با میانگین خطای قدرمطلق (MAE) و میانگین درصد خطای قدرمطلق (MAPE) ارزیابی می‌شود.

$$MAE = \frac{1}{k}\sum_{k=1}^{K} |T_k - \hat{T_k}|$$

$$MAPE = \frac{1}{k}\sum_{k=1}^{K} |\frac{T_k - \hat{T_k}}{T_k}| \times 100$$

عملکرد برای وظیفه برآورد توزیع زمان بقا (تابع بقا) با شاخص (C-td) ارزیابی می‌شود :

$$B_ij = 1 if \hat{H}(t_i|(X_{I,t_i}, X_{T,t_i})) > \hat{H}(t_j|(X_{I,t_j}, X_{T,t_j}))$$

$$C^{td} = \frac{\sum_{i \neq j} A_{i,j} \cdot B_{i,j}}{\sum_{i \neq j} A_{i,j}}$$

این شاخص توانایی مدل را در پیش‌بینی صحیح ترتیب وقوع رویدادها، بر اساس این فرض اندازه‌گیری می‌کند که رویدادهایی که مدت بیشتری طول کشیده‌اند، باید مخاطره کمتری نسبت به رویدادهایی داشته باشند که مدت کوتاه‌تری دوام آورده‌اند. شاخص تطابق وابسته

به زمان تمام بازه‌های زمانی مخاطره تجمعی را ارزیابی می‌کند، بنابراین می‌تواند مواردی را که اثرات عوامل تاثیرگذار در طول زمان تغییر می‌کنند (وابسته به زمان هستند) را نیز ارزیابی کند. این ویژگی به مدل اجازه می‌دهد تا در شرایط پویا و پیچیده‌تر، مانند شبکه‌های ترافیکی که متغیرهای مستقل ممکن است در زمان‌های مختلف تأثیرات متفاوتی داشته باشند، عملکرد بهتری داشته باشد.

بحث و نتیجه‌گیری

جداول زیر عملکرد پیش‌بینی مدل هدف را در مقایسه با سایر مدل‌ها نشان می‌دهد.

- مدل‌های آماری

مدل‌های آماری	MAE	MAPE	ctd	AIC
مدل زمان شکست شتابیده (AFT) (توزیع وایبول)	۱۸.۸۶۵	۵۴.۴۶۵		۲۲۵۵۲.۸۴۴
مدل زمان شکست شتابیده (AFT) (توزیع لگ نرمال)	۱۸.۳۳۵	۵۱.۳۹۱		۱۸۶۱۳.۵۳۶
مدل زمان شکست شتابیده (AFT) (توزیع لگ لجستیک)	۱۸.۴۶۱	۵۲.۱۹۸		۱۸۷۰۰.۴۲۰
مدل شکنندگی (RPAFT) (توزیع لگ لجستیک)	۱۷.۱۱۴	۴۲.۰۶۶		۱۸۲۷۳.۵۸۳

از بین مدل‌های آماری فوق مدل زمان شکست شتابیده با توزیع لگ لجستیک از مدل‌های دیگر AFT ازنظر MAE و MAPE و AICبهتر عمل می‌کند. همچنین مدل شکنندگی با یک پارامتر تصادفی RPAFT بهتر از مدل زمان شکست شتابیده با توزیع فرضی لگ لجستیک عمل می‌کند. این نشان‌دهنده وجود ناهمگنی مشاهده نشده در بین داده‌ها می‌باشد.

- مدل‌های یادگیری ماشین

مدل‌های یادگیری ماشین	MAE	MAPE	ctd	AIC
RF-Survival	۱۷.۱۱۱	۴۰.۴۹۱	۰.۶۰۲	
RF-Regression	۱۸.۱۷۸	۵۳.۵۵۸		
GBM-Survival	۱۶.۷۶۸	۴۰.۵۷۸	۰.۶۰۷	
GBM-Regression	۱۶.۷۹۶	۴۲.۰۰۵		

همه مدل‌های یادگیری ماشین و شبکه‌های عصبی عمیق عملکرد بهتری نسبت به مدل زمان شکست شتابیده با توزیع لگ لجستیک دارند و به این معنی که در نظر گرفتن یک رابطه تعاملی و غیرخطی عملکرد پیش‌بینی را بهبود می‌بخشد. مدل GBM-Regression و DNN-Regression و مدل هدف بهتر از مدل شکنندگی RPAFT عمل می‌کنند. عملکرد هردو مدل جنگل تصادفی (RF) و مدل ماشین گرادیان تقویتی(GBM) با تغییر وظایف‌شان از پیش‌بینی میانگین زمان پاکسازی به برآورد تابع بقا، بهبود یافته است.

- مدل‌های شبکه های عصبی عمیق

مدل‌های شبکه‌های عصبی عمیق	MAE	MAPE	ctd	AIC
DNN-Survival	۱۷.۵۵۳	۴۳.۱۹۵	۰.۶۲۴	
DNN-Regression	۱۶.۸۰۸	۴۱.۴۵۲		

برخلاف مدل‌های یادگیری ماشین، شبکه عصبی عمیق با هدف برآورد تابع بقا عملکرد پیش‌بینی کننده بهتری نسبت به شبکه عصبی عمیق با هدف پیش‌بینی میانگین زمان پاکسازی حادثه، ندارد.

- مدل هدف

مدل هدف	MAE	MAPE	ctd	AIC
Multi task Deep Survival Analysis Model	۱۶.۵۵۱	۳۹.۶۰۹	۰.۶۳۷	

برخلاف مدل‌های دیگر که به طور جداگانه وظایف پیش‌بینی مدت زمان حادثه و براورد تابع بقا را انجام می‌دهند مدل هدف این وظایف را به طور همزمان انجام می‌دهد. مدل هدف بهترین عملکرد را هم برای پیش‌بینی میانگین زمان پاکسازی (از نظر MAE و MAPE) و هم برای برآورد تابع بقا(از نظر c-td) نشان می‌دهد.

قدردانی و تشکر

نویسندگان از داوران محترم به خاطر دقت نظر و صرف وقت در مطالعه مقاله و ارائه پیشنهادات برای بهبود آن، تقدیر و تشکر می‌کنند.

مراجع

Nam, D. and Mannering, F. (2000).*An exploratory hazard based analysis of highway incident duration.*

Jones, B., Janssen, L., and Mannering, F. (1991). *Analysis of the frequency and duration of freeway accidents in Seattle.* Accident, Analysis and Prevention, 23(4), 239–255. https://doi.org/10.1016/0001-4575(91)90003-N

Chung, Y. (2010). *Development of an accident duration prediction model on the Korean Freeway Systems.* Accident; Analysis and Prevention, 42(1), 282–289. https://doi.org/10.1016/j.aap.2009.08.005

Tavassoli Hojati, A., Ferreira, L., Washington, S., and Charles, P. (2013). *Hazard based models for freeway traffic incident duration.* Accident; Analysis and Prevention, 52, 171–181. https://doi.org/10.1016/j.aap.2012.12.037

Wei, C., Lee, Y.(2007). *Sequential forecast of incident duration using artifical neural networks models.* Accident; Analysis and Prevention, 39(5), 944–954. https://doi.org/ 10.1016/j.aap.2006.12.017

Hamad, K., Al-Ruzouq, R., Zeiada, W., Abu Dabous, S., and Khalil, M. A. (2020). *Predicting incident duration using random forests.* Transportmetrica A: Transport Science, 16(3), 1269–1293. https://doi.org/10.1080/23249935.2020.1733132

Ma, X., Ding, C., Luan, S., Wang, Y., Wang, Y. (2017). *Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method.* IEEE Transactions on Intelligent Transportation Systems, 18(9), 2303–2310. https://doi.org/10.1109/ TITS.2016.2635719

Li, L., Sheng, X., Du, B., Wang, Y., and Ran, B. (2020). *A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction. Engineering Applications of Artificial Intelligence, 93(April), 103686.https://doi. org/10.1016/j.engappai.2020.103686*

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y., and Huang, H. (2020).Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. Analytic Methods in Accident Research, 27, 100123. https://doi.org/10.1016/j.amar.2020. 100123

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008).Random survival forests. The Annals of Applied Statistics, 2(3), 841–860. https://doi.org/10.1214/ 08-AOAS169

Hothorn, T., Buhlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006).Survival ensembles. Biostatistics (Oxford, England), 7(3), 355–373. https://doi.org/10.1093/ biostatistics/kxj011

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018).DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology, 18(1), 24. https://doi.org/10.1186/s12874-018-0482-1

Lee, C., Zame, W. R., Yoon, J., and Van Der Schaar, M. (2018).DeepHit: A deep learning approach to survival analysis with competing risks. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 2314–2321. https://doi. org/10.1609/aaai.v32i1.11842