مدل‌بندی داده‌های شمارشی با استفاده از رگرسیون پواسون آماسیده در صفر با یادگیری عمیق

ترنم ترابی نعمان، مهدی عمادی، محمد آرشی

گروه آمار، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد

چکیده: در چشم‌انداز وسیع تحلیل داده‌های مدرن، مواجهه با داده‌های شمارشی که با مشکل فراوانی بیش از حد صفر درگیر هستند، یک معضل تحلیلی بنیادین به شمار می‌رود. مدل‌های کلاسیک چون رگرسیون پواسون، در برابر این پدیده از خود ضعف نشان می‌دهد، زیرا فاقد توانایی برای تفکیک صفرهای ساختاری (ناشی از فرایندهای قطعی) از صفرهای تصادفی (ناشی از فرایندهای تصادفی) هستند. اگرچه مدل‌های رگرسیون پواسون آماسیده در صفر گامی مهم در جهت حل این چالش برداشته‌اند، اما کارایی آن‌ها در عصر داده‌های بزرگ و با ابعاد بسیار بالا، با محدودیت‌های جدی روبه‌رو است. این پژوهش، با نگاهی فراتر از چارچوب‌های سنتی، چارچوبی نوین مبتنی بر شبکه‌های عصبی عمیق را برای تقویت و متحول ساختن مدل‌های رگرسیون پواسون آماسیده در صفر معرفی می‌کند. این رویکرد ترکیبی، با بهره‌گیری از قدرت بی‌نظیر یادگیری عمیق در استخراج ویژگی‌های غیرخطی و پیچیده، قادر است فرایند دوگانه تولید داده‌ها (فرایند تولید صفر و فرایند شمارش) را با دقتی بی‌سابقه مدل‌سازی کند. نتایج حاصل از شبیه‌سازی‌های گسترده و کاربرد بر روی مجموعه داده‌های واقعی، برتری چشمگیر این معماری نوین را در سناریوهای با ابعاد بالا و روابط متغیرهای پیچیده، به وضوح آشکار می‌سازد.

واژه‌های کلیدی: پرسپترون چند لایه، داده شمارشی، رگرسیون پواسون آماسیده در صفر، رگرسیون پواسون عمیق، شبکه‌های عصبی.

مقاله پذیرفته شده

۱ مقدمه

در حوزه‌هایی چون اقتصاد، اپیدمیولوژی، مهندسی و علوم اجتماعی، داده‌های شمارشی روزمره رصد می‌شوند. اگر سهم صفرها زیاد شود، مدل‌سازی دشوار می‌گردد؛ این وضعیت صرفاً یک ویژگی آماری نیست، بلکه پیامد دو فرایند متفاوت است: بخشی از صفرها قطعی و بخشی تصادفی پدید می‌آیند. رگرسیون پواسون کلاسیک این دو را تفکیک نمی‌کند و برآورد و پیش‌گویی را دچار خطا می‌سازد تلاش برای رفع این مشکل از دهه ۱۹۷۰ در آمار صنعتی آغاز شد، هنگامی که پژوهشگران در داده‌های نقص تولید با فراوانی صفر روبه‌رو شدند. لامبرت (۱۹۹۲) نخستین چارچوب منسجم را با معرفی رگرسیون پواسون آماسیده در صفر (ZIP) ارائه کرد. سپس بوهینگ و همکاران (۱۹۹۹)، هال (۲۰۰۰) و ریداوت و همکاران (۲۰۰۱) این مدل را به داده‌های طولی و زیست‌آماری و به آزمون مقایسه ZIP با نسخه دو جمله‌ای منفی (ZINB) گستراندند. با ظهور داده‌های پیچیده و چندبعدی، محدودیت‌های ZIP و ZINB آشکارتر شد؛ این مدل‌ها همچنان بر ساختارهای خطی تکیه دارند و روابط غیرخطی و ساختارهای سلسله‌مراتبی را درنمی‌یابند. بنابراین نیاز به رویکردی انعطاف‌پذیر احساس می‌شود که ویژگی‌های پنهان و وابستگی‌های پیچیده را با خطای کمتر و دقت بالاتر مدل کند. شبکه عصبی به‌عنوان یک رویکرد غیرخطی این خلأ را پر می‌کند. فلاح و همکاران (۲۰۱۱) نشان داد مفروضات خطی رگرسیون در بسیاری موارد دقت پایین‌تری دارند، در حالی‌که شبکه عصبی این مسئله را برطرف می‌کند. ریزوی و همکاران (۲۰۲۴) با تقریب نسبت درستنمایی توسط شبکه عصبی به نتایج بهتری رسیدند. مؤنتسینوس–لوپز و همکاران (۲۰۲۱) عملکرد برتر شبکه عصبی عمیق پواسون را نسبت به رگرسیون بیزی و پواسون کلاسیک گزارش کردند.

همسو با این یافته‌ها، مطالعه حقانی و همکاران (۲۰۱۷) در کاربرد داده‌های اهدای خون نشان داد که شبکه عصبی مصنوعی با کاهش چشمگیر خطای میانگین مربعات نسبت به مدل‌های ZINB و ZIP،

آدرس الکترونیکی نویسنده مسئول مقاله: مهدی عمادی، emadi@um.ac.ir
کد موضوع‌بندی ریاضی (۲۰۲۰): 62J12, 68T07.

عملکردی به‌مراتب برتر در پیش‌گویی داده‌های شمارشی با تعداد بالای صفر دارد. مرور مطالعات مشابه در یادگیری ماشین و هوش مصنوعی را می‌توان در مددی و مترجم (۱۴۰۳) یافت. کاربرد شبکه‌های عصبی در داده‌های سانسور شده و زمان بقای گسسته در شارف و تسوکوس (۲۰۱۵) و شبکه‌های عصبی بیزی در سلامت در رودریگو (۲۰۱۷) و مدل‌سازی بیزی غیرخطی داده‌های شمارشی در رودریگو و تسوکوس (۲۰۲۰) و شبکه‌های عمیق در مدل‌های شکنندگی نیمه‌پارامتری در لی و همکاران (۲۰۲۳) قابل مشاهده است.

در این پژوهش مدلی ترکیبی به نام رگرسیون شبکه عصبی پواسون آماسیده در صفر معرفی می‌شود که چارچوب ZIP را با توان یادگیری شبکه عصبی ادغام می‌کند. این مدل هم‌زمان دو فرایند تولید صفرهای قطعی و شمارش مقادیر غیرصفر را در قالبی انعطاف‌پذیر مدل‌سازی می‌کند تا دقت پیش‌گویی افزایش یابد و درک عمیق‌تری از ساختار داده‌های شمارشی حاصل گردد. اهمیت این مدل‌سازی فراتر از افزایش دقت پیش‌گویی است و به ماهیت استنتاج آماری در عصر داده‌های پیچیده بازمی‌گردد. در مدل‌های احتمالی کلاسیک مانند ZIP، فرض بر آن است که روابط بین متغیرها از پیش مشخص و اغلب خطی هستند؛ اما در دنیای واقعی، پدیده‌های شمارشی همچون تولید علمی یا وقوع حوادث، تحت تأثیر شبکه‌ای از تعاملات غیرخطی و عوامل پنهان قرار دارند که مدل‌های پارامتریک سنتی قادر به ثبت آن‌ها نیستند. نادیده گرفتن این پیچیدگی‌های ساختاری نه‌تنها باعث کاهش دقت پیش‌گویی می‌شود، بلکه می‌تواند منجر به تفسیرهای غلط از علل پدیده‌ها و در نتیجه تصمیم‌گیری‌های ناکارآمد شود. بنابراین، گذار از مدل‌های ثابت و خطی به سمت مدل‌های انعطاف‌پذیر و غیرخطی، یک انتخاب تکنیکی نیست، بلکه یک ضرورت شناخت‌شناسانه برای درک صحیح داده‌های امروزی است.

از سوی دیگر، ادغام چارچوب‌های تفسیرپذیر آماری (مانند توزیع پواسون آماسیده در صفر) با قدرت محاسباتی شبکه‌های عصبی، پلی میان دو دنیای تفسیرپذیری و دقت ایجاد می‌کند. در حالی که شبکه‌های عصبی خالص اغلب به عنوان جعبه‌های سیاه مورد انتقاد قرار می‌گیرند، مدل پیشنهادی با حفظ ساختار احتمالی ZIP، اجازه می‌دهد که پارامترهای فیزیکی و معنادار مانند احتمال وقوع صفرهای قطعی، همچنان استخراج و تفسیر شوند. این رویکرد ترکیبی، ابهامات موجود در روش‌های یادگیری ماشین را کاهش داده و اعتمادپذیری نتایج را برای کاربران حوزه‌های علوم اجتماعی و پزشکی که به تفسیر پارامترها نیاز دارند،

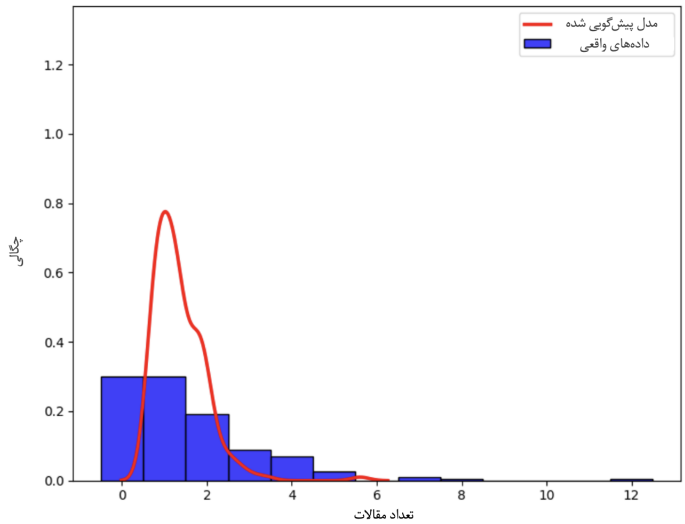
افزایش می‌دهد. بدین ترتیب، این پژوهش تنها یک ابزار محاسباتی ارائه نمی‌دهد، بلکه چارچوبی را توسعه می‌دهد که در آن دقت الگوریتم‌های هوش مصنوعی در خدمت استنتاج‌های علمی دقیق و قابل فهم قرار می‌گیرد. دو دیدگاه مدنظر است. نخست آنکه یادگیری تابع درستنمایی و برآورد پارامتر در توزیع‌های پیچیده دشوار است، هرچند در کاربردهای کلاسیک و بیزی با چنین توزیع‌هایی سروکار داریم. دوم آنکه حتی در توزیع‌هایی که برآوردگر بسته یا کارا دارند، چون ماکسیمم‌سازی تابع درستنمایی هدف است، کمینه‌سازی خطای برآورد لحاظ نمی‌شود؛ در حالی‌که در یادگیری عمیق می‌توان هم تابع درستنمایی را بیشینه کرد و هم خطای برآورد را کاهش داد که یک رویکرد جدید آماری–بهینه‌سازی محسوب می‌شود.

هدف اصلی این مقاله توسعه رگرسیون پواسون به رگرسیون شبکه عصبی پواسون آماسیده در صفر است تا عملکرد مدل در مجموعه‌ای از داده‌های تعداد مقالات دانشجویان تحصیلات تکمیلی ارتقا یابد. چالش موجود در این داده‌ها به‌عنوان یک مثال کاربردی در بخش بعدی تشریح می‌شود.

۱.۱ مثال انگیزشی

برای افزایش دقت پیش‌گویی تعداد مقالات منتشرشده توسط دانشجویان دکتری شیمی بیولوژیک، از مجموعه داده‌ای عمومی و در دسترس با عنوان بیوشیمی بهره‌گیری شده است. این داده‌ها که در بسته آماری pscl نرم‌افزار R قرار دارد، حاصل یک پیمایش طولی از دانشجویان تحصیلات تکمیلی رشته شیمی بیولوژیک در ایالات متحده است و اطلاعاتی درباره تعداد مقالات منتشرشده در دوره دکتری و عوامل مؤثر بر فعالیت‌های پژوهشی ارائه می‌دهد. نمونه مورد بررسی شامل ۹۱۵ دانشجوی دکتری پیش از دهه ۱۹۸۰ است. متغیر وابسته تعداد مقالات منتشرشده و متغیرهای توضیحی جنسیت، وضعیت تأهل، تعداد فرزندان زیر پنج سال، سطح تحصیلات دانشگاهی و تعداد مشاوران علمی در نظر گرفته شده‌اند. مقادیر متغیرهای عددی با کسر میانگین و تقسیم بر انحراف‌معیار استاندارد شده و متغیرهای دسته‌ای با کدگذاری دودویی وارد مدل شده‌اند؛ این فرآیند در تمام مدل‌های پواسون و پواسون آماسیده در صفر به‌طور یکنواخت اعمال شده است.

شکل ۱ توزیع فراوانی تعداد مقالات واقعی و پیش‌گویی‌شده توسط مدل ZIP را نشان می‌دهد؛ یک قله برجسته در صفر و دم کشیده به سمت مقادیر بزرگ‌تر وجود دارد که گویای دو گروه است: افرادی که به صورت ساختاری هیچ مقاله‌ای منتشر نمی‌کنند و افرادی که فعالیت علمی دارند اما تعداد مقالاتشان



شکل ۱: مقایسه مقادیر واقعی و مقادیر پیش‌گویی‌شده در داده‌های تعداد مقالات علمی

تصادفی است. این ساختار، مدل‌های سنتی پواسون را با چالش روبه‌رو می‌سازد و استفاده از رویکردهای پیشرفته‌تر را ضروری می‌نماید؛ ازاین‌رو در ادامه، ترکیب شبکه عصبی با توزیع‌های آماری برای کشف الگوهای عمیق‌تر تولید علم مورد استفاده قرار خواهد گرفت.

ساختار مقاله به این صورت تنظیم شده است: بخش ۲ رگرسیون شبکه عصبی پواسون را توضیح می‌دهد؛ بخش‌های ۳ و ۴ جزئیات بهینه‌سازی شبکه عمیق را ارائه می‌کنند؛ شبیه‌سازی مدل پیشنهادی در بخش ۵ آمده است؛ تحلیل داده‌های واقعی در بخش ۶ انجام شده و خلاصه نتایج در بخش ۷ ارائه شده است.

۲ رگرسیون شبکه عصبی پواسون

فرض کنید یک مجموعه داده آموزشی از بردارهای ورودی $\{x_n\}$، که $n = ۱, ..., N$، همراه با مقادیر هدف مربوطه $\{y_n\}$ داده شده است. فرض می‌کنیم مشاهدات y_n به شرط x_n مستقل باشند. در رگرسیون پواسون، فرض می‌شود که هر y_n دارای توزیع پواسون با پارامتر λ_n باشد که تابع جرم احتمال (PMF)

آن به صورت $P(Y_n = y_n) = \frac{e^{-\lambda_n} \cdot \lambda_n^{y_n}}{y_n!}, \quad y_n = ۰, ۱, ۲, ...,$ است. در این مدل، لگاریتم پارامتر λ_n به عنوان یک تابع خطی از ورودی x_n در نظر گرفته می‌شود: $\ln(\lambda_n) = x_n^\top \beta \Rightarrow \lambda_n = e^{x_n^\top \beta}$. بنابراین، $E[y_n|x_n] = e^{x_n^\top \beta}$ این بدین معناست که خروجی مدل (یعنی پیش‌گویی تعداد رویدادها) یک تابع نمایی از ورودی‌ها است. وقتی رابطه بین ورودی‌ها و خروجی‌ها پیچیده‌تر می‌شود، می‌توان از مدل‌های غیرخطی مانند شبکه‌های عصبی MLP استفاده کرد. در این حالت، به جای فرض رابطه خطی بین x_n و λ_n، از یک شبکه عصبی برای پیش‌گویی λ_n استفاده می‌شود. خروجی شبکه برای نمونه n–ام به صورت $t_k = \hat{\lambda}_n$ در نظر گرفته می‌شود. یک ساختار MLP دو لایه به صورت

$$t_k = \hat{\lambda}_n = \varphi_۰\big(\alpha_k + \sum_{j \to k} \omega_{jk} \cdot \varphi_h\big(\alpha_j + \sum_{i \to j} \omega_{ij} \cdot x_n\big)\big). \qquad (۱)$$

نوشته می‌شود، که در آن x_i مقادیر ورودی، t_k مقدار(های) خروجی، ω_{ij} وزن‌های اتصال لایه ورودی به لایه پنهان، ω_{jk} وزن‌های اتصال لایه پنهان به لایه خروجی، φ_h تابع فعال‌سازی لایه پنهان مثلاً ReLU یا Tanh، $\varphi_۰$ تابع فعال‌سازی لایه خروجی اغلب Softplus یا نمایی و α_j, α_k مقادیر اریبی لایه‌ها است.

برای آموزش این مدل، از معیار منفی لگاریتم درستنمایی[1] (NLL) استفاده می‌شود. این تابع زیان به صورت $E = -\sum_{n=۱}^{N} [-t_n + y_n \log(t_n) - \ln(y_n!)]$, تعریف می‌شود. اما جمله آخر ($\ln(y_n!)$) فقط به مقادیر هدف بستگی دارد و در طول آموزش ثابت است. بنابراین، این جمله را می‌توان حذف کرد و تابع زیان ساده شده به صورت $E = \sum_{n=۱}^{N} (t_n - y_n \log(t_n))$, خواهد بود. این تابع به ما کمک می‌کند تا وزن‌های شبکه را به گونه‌ای تنظیم کنیم که خروجی‌های پیش‌گویی‌شده t_n به مقادیر واقعی y_n نزدیک شوند.

پارامترهای شبکه عصبی با استفاده از الگوریتم پس‌انتشار خطا راملهارت و همکاران (۱۹۸۶) و قاعده زنجیره‌ای به‌روزرسانی شدند. در مرحله انتشار پیشرو، خروجی لایه‌های پنهان با استفاده از تابع فعال‌سازی ReLU و خروجی لایه نهایی با استفاده از تابع Softplus به صورت $\hat{y} = \log\left(۱ + e^{Z^{(out)}}\right)$ محاسبه می‌شود. سپس، گرادیان تابع درستنمایی رگرسیون پواسون نسبت به خروجی شبکه به صورت

[1] Negative log-likelihood

$\delta^{(out)} = \left(\frac{y}{\hat{y}} - ۱\right) \frac{۱}{۱+e^{-Z^{(out)}}}$ محاسبه می‌شود. خطای لایه‌های پنهان نیز با استفاده از قاعده زنجیره‌ای از لایه خروجی به سمت لایه‌های قبلی منتقل شده و گرادیان وزن‌ها و اریبی‌ها محاسبه می‌شود. در نهایت، پارامترهای شبکه مطابق روابط

$$w^{(h)} \leftarrow w^{(h)} - \eta \frac{\partial L}{\partial w^{(h)}}, \qquad b^{(h)} \leftarrow b^{(h)} - \eta \frac{\partial L}{\partial b^{(h)}},$$

به‌روزرسانی می‌شوند. که در آن η نرخ یادگیری است. این فرآیند تا همگرایی تابع هدف تکرار می‌شود.

۳ مدل‌سازی رگرسیون و شبکه‌های عصبی تحت توزیع ZIP

در مدل ZIP، متغیر تصادفی Y دارای دو مؤلفه است: یک متغیر تصمیم‌گیرنده دو وضعیتی که مشخص می‌کند آیا مشاهده صفر است یا خیر و یک متغیر توزیع پواسون که مقادیر غیرصفر را مدل می‌کند. بنابراین گوییم متغیر تصادفی Y دارای توزیع ZIP است اگر MPF به صورت

$$P(Y = y) = \begin{cases} \pi + (۱ - \pi)e^{-\lambda}, & y = ۰ \\ (۱ - \pi)\dfrac{\lambda^y e^{-\lambda}}{y!}, & y > ۰ \end{cases}, \tag{۲}$$

تعریف شود که در آن، $\lambda > ۰$ پارامتر میانگین توزیع پواسون برای مقادیر غیر صفر و $\pi \in [۰, ۱]$ احتمال اضافه شدن صفرها است. این مدل اجازه می‌دهد تا دو نوع صفر حاصل از عدم وقوع رویداد و صفر حاصل از وقوع تصادفی رویداد را تشخیص دهیم. در برآورد ماکسیمم درستنمایی فرض کنید داده‌های شمارشی $y_۱, \ldots, y_n$ به صورت مستقل و هم‌توزیع از مدل ZIP با پارامترهای π و λ گرفته شده‌اند. PMF برای هر مشاهده به صورت فرمول (۲) است. بنابراین لگاریتم تابع درستنمایی نمونه برابر

$$\mathscr{L}(\pi, \lambda) = \sum_{i=۱}^{n} I(y_i = ۰) \log \left[\pi + (۱ - \pi)e^{-\lambda}\right]$$

$$+\sum_{i=1}^{n} I(y_i > 0)\left[\log(1-\pi) - \lambda + y_i \log\lambda - \log(y_i!)\right], \qquad (3)$$

است، که در آن $\mathbb{I}(\cdot)$ تابع نشانگر است.

۴ مدل‌سازی غیرخطی در مدل ZIP عمیق

می‌توان از شبکه‌های عصبی MLP برای برآورد کردن پارامترهای λ_k و π_k به شیوه‌ای غیرخطی استفاده کرد. در این حالت، شبکه دارای دو خروجی مستقل برای $\hat{\lambda}_k$ و دیگری برای $\hat{\pi}_k$ است. یک ساختار MLP دو لایه برای برآورد پارامترهای λ_k و π_k به صورت

$$\hat{\lambda}_k = \varphi_0^{(\lambda)}\big(\alpha_k^{(\lambda)} + \sum_{j\to k} \omega_{jk}^{(\lambda)} \cdot \varphi_h\big(\alpha_j + \sum_{i\to j} \omega_{ij} \cdot x_i\big)\big),$$
$$\hat{\pi}_k = \varphi_0^{(\pi)}\big(\alpha_k^{(\pi)} + \sum_{j\to k} \omega_{jk}^{(\pi)} \cdot \varphi_h\big(\alpha_j + \sum_{i\to j} \omega_{ij} \cdot x_i\big)\big),$$

است، که در آن، i شماره مشاهده، j شماره نورون در لایه پنهان و k شماره نورون در لایه خروجی (در اینجا یکی برای λ و یکی برای π) است. وزن‌های به کاربرده شده شامل w_{ij} وزن بین نورون j–ام در لایه ورودی و نورون i–ام، $w_{jk}^{(\lambda)}$ وزن بین نورون j–ام در لایه پنهان و نورون k–ام در لایه خروجی برای λ و $w_{jk}^{(\pi)}$ وزن بین نورون j–ام در لایه پنهان و نورون k–ام در لایه خروجی برای v است. همچنین α_j اریبی نورون j–ام، $\alpha_k^{(\lambda)}$ اریبی نورون k–ام در لایه خروجی برای λ، $\alpha_k^{(\pi)}$ اریبی نورون k–ام در لایه خروجی برای π، φ_h تابع فعال‌سازی لایه پنهان مانند ReLU یا tanh، $\varphi_0^{(\lambda)}, \varphi_0^{(\pi)}$ توابع فعال‌سازی لایه خروجی اغلب Softplus یا نمایی برای λ_k و سیگموئید برای π_k هستند.

برای آموزش این مدل، به منظور بهینه‌سازی پارامترهای مدل بر اساس مشاهدات، از معیار منفی لگاریتم درستنمایی استفاده می‌شود. این معیار به عنوان تابع زیان، اندازه‌گیری خطا و عدم تطابق میان پیش‌گویی‌های مدل و داده‌های مشاهده‌شده را به صورت عددی نشان می‌دهد. برای محاسبه تابع زیان، باید

به تابع جرم احتمال توزیع ZIP توجه کرد. بنابر فرمول‌های (۲) و (۳) تابع زیان به صورت

$$E = -\sum_{i=1}^{n} \log P(Y_i = y_i \mid x_i),$$

تعریف می‌شود. با توجه به دو حالت موجود در فرمول (۲) ، تابع زیان را می‌توان به صورت

$$\begin{aligned} E = -\sum_{i=1}^{n} \Big[& \mathbb{I}(y_i = 0) \log \big(\pi_i + (1 - \pi_i) e^{-\lambda_i}\big) \\ & + \mathbb{I}(y_i > 0) \log \big((1 - \pi_i) \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}\big) \Big], \end{aligned} \tag{۴}$$

بازنویسی کرد که در آن $\mathbb{I}(\cdot)$ تابع نشانگر است.

برای بهینه‌سازی وزن‌های شبکه عصبی در مدل ZIP، می‌توان از روش‌های بهینه‌سازی با استفاده از مشتقات مرتبه اول یا دوم استفاده کرد. در اینجا، نحوه محاسبه مشتق خطا نسبت به وزن‌ها را با استفاده از الگوریتم R-انتشار بیان می‌شود. R-انتشار یک روش کارآمد برای محاسبه مشتقات جهت‌دار مرتبه دوم بدون محاسبه مستقیم ماتریس هسین است. این روش برای بهینه‌سازی مدل‌های پیچیده مانند شبکه‌های عصبی با خروجی ZIP مناسب است. معادلات پس‌انتشار برای یک شبکه MLP با یک لایه پنهان به صورت

$$t_j^{(1)} = \sum_i w_{ji} x_i, \quad z_j = g(t_j^{(1)}), \quad t_k^{(2)} = \sum_j w_{kj} z_j,$$

تعریف می‌شوند که در آن $t_j^{(1)}$ خروجی قبل از فعال‌سازی در لایه اول، z_j خروجی لایه پنهان پس از اعمال تابع فعال‌سازی $g(\cdot)$، $t_k^{(2)}$ خروجی قبل از فعال‌سازی در لایه دوم است.

۱.۴ معادلات R-انتشار و توابع گرادیان

بیان ریاضی مشتقات دوم تابع زیان با عملگر $R[\cdot]$، به صورت گام به گام انجام می‌شود و دو ویژگی کلیدی شکل تابع زیان و پایداری عددی را آشکار می‌کند. این موضوع می‌تواند مشکلات احتمالی از همگرایی

مقاله پذیرفته شده

الگوریتم بهینه‌سازی را پیش‌گویی کند. فرض کنید $R[w] = \pi$، که در آن $R[\cdot]$ یک عملگر دیفرانسیلی خطی و w بردار وزن‌های شبکه است. روابط R-انتشار به صورت

$$R[t_j^{(۱)}] = \sum_i \pi_{ji} x_i,$$
$$R[z_j] = g'(t_j^{(۱)}) \cdot R[t_j^{(۱)}],$$
$$R[t_k^{(۲)}] = \sum_j w_{kj} \cdot R[z_j] + \sum_j \pi_{kj} \cdot z_j,$$

بیان می‌شود که در آن π_{ji} عنصری از بردار v مربوط به وزن w_{ji} و $g'(\cdot)$ مشتق تابع فعال‌ساز است. این معادلات اجازه می‌دهند تا عملگر دیفرانسیلی را به صورت پیشرونده از طریق شبکه منتشر کنیم و مشتقات مرتبه دوم را به درستی برآورد کند.

در مدل ZIP، خروجی‌ها دارای دو پارامتر مستقل λ_k، پارامتر نرخ (شبیه به میانگین) و π_k، احتمال آماسیده در صفر، هستند. بنابراین، تابع خطا همان‌طور که در فرمول (۴) نتیجه‌گیری کردیم، به صورت

$$E = -\sum_{k=۱}^{K} \left[\mathbb{I}(y_k = ۰) \log\left(\pi_k + (۱-\pi_k)e^{-\lambda_k}\right) + \mathbb{I}(y_k > ۰) \log\left(۱-\pi_k\right) \frac{\lambda_k^{y_k} e^{-\lambda_k}}{y_k!}\right],$$

تعریف می‌شود. در این مرحله لازم است مشتق‌های خطا را نسبت به پارامترهای λ_k و π_k برای مرحله پس‌انتشار و به روز کردن وزن‌های شبکه محاسبه گردد. برای این منظور روابط

$$\begin{cases} R[\delta_k^{(\lambda)}] = R[t_k^{(۲)}], & \text{برای } \lambda_k, \\ R[\delta_k^{(\pi)}] = R[t_k^{(۲)}], & \text{برای } \pi_k, \end{cases}$$

بیان می‌شود. با استفاده از عملگر $R[\cdot]$، مشتقات مرتبه دوم به صورت

$$R[\delta_j^{(\lambda)}] = g''(t_j^{(1)})R[t_j^{(1)}]\sum_k w_{kj}\delta_k^{(\lambda)} + g'(t_j^{(1)})\sum_k v_{kj}\delta_k^{(\lambda)} + g'(t_j^{(1)})\sum_k w_{kj}R[\delta_k^{(\lambda)}],$$

$$R[\delta_j^{(\pi)}] = g''(t_j^{(1)})R[t_j^{(1)}]\sum_k w_{kj}\delta_k^{(\pi)} + g'(t_j^{(1)})\sum_k v_{kj}\delta_k^{(\pi)} + g'(t_j^{(1)})\sum_k w_{kj}R[\delta_k^{(\pi)}],$$

و بدون محاسبه ماتریس هسین کامل، به دست می‌آیند. مشتقات مرتبه اول تابع خطا نسبت به وزن‌ها به صورت

$$\frac{\partial E}{\partial w_{ji}} = \delta_j x_i, \qquad \text{(برای لایه اول)}$$

$$\frac{\partial E}{\partial w_{kj}} = \delta_k^{(\lambda)} z_j + \delta_k^{(\pi)} z_j, \qquad \text{(برای لایه دوم)}$$

هستند. برای محاسبه عناصر حاصل‌ضرب بردار–هسین $v^\top H$، از روابط

$$R\left\{\frac{\partial E}{\partial w_{ji}}\right\} = x_i R[\delta_j],$$

$$R\left\{\frac{\partial E}{\partial w_{kj}}\right\} = R[\delta_k^{(\lambda)}]z_j + \delta_k^{(\lambda)}R[z_j] + R[\delta_k^{(\pi)}]z_j + \delta_k^{(\pi)}R[z_j],$$

استفاده می‌گردد. این چارچوب به ما اجازه می‌دهد تا اطلاعات مرتبه دوم را بهره‌ورانه پیش‌گویی کنیم که برای بهینه‌سازی مدل‌های پیچیده مانند شبکه‌های عصبی با خروجی ZIP ضروری است.

۵ مطالعات شبیه‌سازی

در این بخش، به‌منظور ارزیابی و مقایسه عملکرد مدل‌های رگرسیونی در تحلیل داده‌های شمارشی با تعداد زیادی صفر، یک مطالعه شبیه‌سازی انجام شده است. هدف اصلی، بررسی توانایی یک مدل رگرسیونی

مقاله پذیرفته شده

شبکه عصبی پواسون آماسیده در صفر، در پیش‌گویی داده‌های شمارشی و مقایسه آن با مدل رگرسیونی ZIP و رگرسیون پواسون و رگرسیون شبکه عصبی پواسون است. این مقایسه به‌گونه‌ای طراحی شده که ضمن حفظ شرایط یکسان برای هر سه مدل، تفاوت‌های ذاتی در ساختار مدل‌سازی، انعطاف‌پذیری و توانایی در شناسایی الگوهای پیچیده داده‌ها مورد کاوش قرار گیرد. مقایسه مدل‌های به کار گرفته شده، با توجه به معیار RMSE با رابطه $\mathrm{RMSE} = \sqrt{\frac{۱}{n}\sum_{i=۱}^{n}\left(y_i - \hat{y}_i\right)^۲}$ و معیار MAE با رابطه $\mathrm{MAE} = \left|\frac{۱}{n}\sum_{i=۱}^{n}\left(y_i - \hat{y}_i\right)^۲\right|$ انجام شد.

داده‌های شبیه‌سازی شده با ساختار صفر بیش از حد و روابط غیرخطی بین متغیرهای پیش‌گو و متغیر پاسخ تولید شده است. داده‌ها از یک فرآیند ترکیبی بر اساس توزیع ZIP شبیه‌سازی شده‌اند که در آن هر مشاهده y_i $(i = ۱, ۲, \ldots, n)$ با احتمالی مشخص از یک فرآیند صفر ساختاری ناشی می‌شود و در غیر این صورت از یک توزیع پواسون با میانگین وابسته به متغیرهای توضیحی پیروی می‌کند. این ساختار به‌خوبی پدیده‌هایی را مدل می‌کند که در آن‌ها تعداد زیادی صفر به دلایل ساختاری (مانند عدم تمایل یا دسترسی)، نه به‌عنوان یک پیامد تصادفی از یک فرآیند شمارشی، رخ می‌دهند.

فرآیند تولید داده به این صورت است که ابتدا $n = ۲۰۰۰$ نمونه از یک توزیع یکنواخت در بازه $[-۱, ۱]$ برای $p = ۳$ متغیر مستقل تولید می‌شوند، یعنی $\mathbf{x}_{ip} = [x_{i۱}, x_{i۲}, x_{i۳}]^\top \sim \mathrm{U}(-۱, ۱)$. سپس دو مؤلفه اصلی مدل ZIP، یعنی احتمال صفر ساختاری π_i و میانگین توزیع پواسون λ_i، به‌صورت غیرخطی بر حسب این متغیرها تعریف می‌شوند. برای مدل‌سازی میانگین پواسون، از یک تابع لگاریتمی با روابط غیرخطی استفاده شده است. به عنوان مثال داریم:

$$\log(\lambda_i) = ۰/۵x_{i۱} + ۰/۸\sin(\pi x_{i۲}) - ۰/۳x_{i۳}^۲,$$

که پس از آن با تبدیل نمایی، میانگین توزیع پواسون به‌دست می‌آید:

$$\lambda_i = \exp\left(۰/۵x_{i۱} + ۰/۸\sin(\pi x_{i۲}) - ۰/۳x_{i۳}^۲\right).$$

این روابط غیرخطی به شبیه‌سازی شرایطی واقعی کمک می‌کنند که در آن‌ها تأثیر متغیرها بر پاسخ، خطی

نیست. هم‌زمان، احتمال اینکه مشاهده y_i یک صفر ساختاری باشد، از طریق یک مدل رگرسیون لجستیک به صورت

$$\text{logit}(\pi_i) = -۰/۶ + ۱/۲x_{i۱} - ۰/۹\cos(\pi x_{i۲}),$$

تعریف می‌شود، که با تبدیل سیگموئیدی به احتمال مربوطه تبدیل می‌شود:

$$\pi_i = \frac{۱}{۱ + \exp\big(-(-۰/۶ + ۱/۲x_{i۱} - ۰/۹\cos(\pi x_{i۲}))\big)}.$$

در ادامه، برای هر نمونه i، یک متغیر پنهان $z_i \sim \text{Bernoulli}(\pi_i)$ تولید می‌شود که تعیین می‌کند آیا مشاهده از مؤلفه صفر ساختاری ($z_i = ۱$) یا از مؤلفه پواسونی ($z_i = ۰$) می‌آید. سپس یک مقدار $c_i \sim \text{Poisson}(\lambda_i)$ از توزیع پواسون تولید شده و مقدار نهایی متغیر پاسخ به صورت $y_i = (۱ - z_i)\cdot c_i$ محاسبه می‌شود. این معادل است با:

$$y_i = \begin{cases} ۰ & \text{با احتمال } \pi_i \\ c_i & \text{با احتمال } ۱ - \pi_i \end{cases},$$

که توزیع توام آن به‌صورت تابع جرم احتمال (۲) تعریف می‌شود. این ساختار ترکیبی از دو فرآیند، ضمن ایجاد تورم مصنوعی در مقادیر صفر، امکان مدل‌سازی روابط پیچیده و غیرخطی بین متغیرهای ورودی و پاسخ را فراهم می‌کند. تمامی محاسبات با کنترل تکرارپذیری انجام شده‌اند تا نتایج قابل بازتولید باشند.

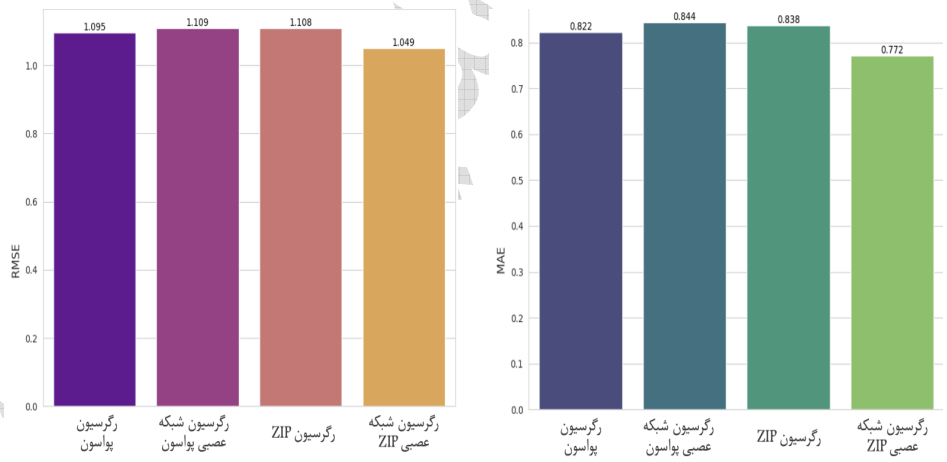
پس از تولید داده‌های شبیه‌سازی شده با مدل ZIP، داده‌ها به‌طور تصادفی به مجموعه‌های آموزش و آزمایش تقسیم شدند و برای استفاده در مدل‌های مبتنی بر شبکه عصبی، به فرمت تنسور تبدیل شدند. بررسی داده‌ها نشان داد که ۶۵/۹۵ درصد از مقادیر متغیر پاسخ برابر با صفر است. این عدد بالا نشان‌دهنده وجود فرایند آماسیده در صفر است؛ یعنی بیشتر افراد یا واقعه‌ها هیچ مقداری نداشته‌اند، نه به خاطر شانس، بلکه به دلیل اینکه اصلا وارد فرآیند شمارش نشده‌اند (مثل کسی که هرگز خرید نمی‌کند). این ویژگی، داده‌ها را به چالشی واقعی برای مدل‌ها تبدیل می‌کند، چون مدل باید بتواند بین صفرهای واقعی و صفرهای

تصادفی تفارت قائل شود. این ویژگی، اهمیت استفاده از مدل‌هایی مانند ZIP یا شبکه‌های عصبی ترکیبی را برجسته می‌کند، چون مدل‌های ساده نمی‌توانند چنین داده‌هایی را به درستی پیش‌گویی کنند.

نتایج مقایسه عملکرد مدل‌ها در جدول ۱ نشان می‌دهد که هیچ یک از مدل‌های رگرسیون پواسون نمی‌توانند به‌خوبی داده‌های شبیه‌سازی شده با چالش تعداد بالای صفر را پیش‌گویی کنند.

جدول ۱: مقایسه عملکرد مدل‌های مختلف بر اساس معیارهای خطای MAE و RMSE

مدل	MAE	RMSE
رگرسیون پواسون	۰٫۸۲۰	۱٫۰۹۰
رگرسیون شبکه عصبی پواسون	۰٫۸۴۰	۱٫۱۱۰
رگرسیون ZIP	۰٫۸۳۸	۱٫۱۰۸
رگرسیون شبکه عصبی ZIP	۰٫۷۷۲	۱٫۰۴۹



(الف) (ب)

شکل ۲: مقایسه عملکرد مدل‌های مختلف بر اساس معیارهای خطای الف: MAE و ب: RMSE

مدل رگرسیون پواسون ساده با MAE حدود ۰٫۸۲ و RMSE برابر با ۱٫۰۹، از کارایی مناسبی برخوردار است، اما از آنجا که این مدل فرض می‌کند تمام صفرها ناشی از تصادف و شدت کم رویداد هستند، نمی‌تواند با دقت بالا رفتار داده‌ها را بیان کند. مدل شبکه عصبی پواسونی که از قابلیت‌های غیرخطی شبکه‌های عصبی استفاده می‌کند، نسبت به مدل ساده کمی ضعیف‌تر عمل کرده و خطاهای آن

کمی بالاتر است (RMSE: ۱/۱۱، MAE: ۰/۸۴) . این نشان می‌دهد که اگرچه شبکه عصبی قادر به یادگیری روابط غیرخطی است، اما بدون در نظر گرفتن فرایند آماسیده در صفر در داده‌ها، همچنان محدودیت‌هایی دارد. در مقابل، مدل ZIP که به‌طور اختصاصی برای داده‌های آماسیده در صفر طراحی شده، عملکرد بهتری نسبت به دو مدل قبلی دارد (RMSE: ۱/۱۰۸ ، MAE: ۰/۸۳۸)، چون توانسته است مؤلفه صفر ساختاری را به‌طور جداگانه مدل‌سازی کند.

بهترین عملکرد را مدل رگرسیونی شبکه عصبی پواسون آماسیده در صفر داشته است که با MAE معادل ۰/۷۷۲ و RMSE معادل ۱/۰۴۹، کمترین خطا را ثبت کرده است. این نتیجه نشان می‌دهد که ترکیب قابلیت‌های یادگیری عمیق با ساختار آماری ZIP، به بهترین شکل از دو جنبه روابط غیرخطی و فرایند آماسیده در صفر، پشتیبانی می‌کند و می‌تواند پیش‌گویی دقیق‌تری ارائه دهد. همچنین برای ارزیابی دقت و کارایی مدل شبکه عصبی پواسون آماسیده در صفر پیشنهادی، معیارهای عملکرد پارامترهای λ و π محاسبه شد.

جدول ۲: آماره‌های توصیفی، معیارهای خطا و ضریب همبستگی مقادیر واقعی و پیش‌گویی‌شده پارامترها

پارامتر	وضعیت	میانگین	انحراف معیار	اریبی	RMSE	خطای استاندارد	ضریب همبستگی
λ	واقعی	۱/۰۴۵۳	۰/۷۷۹۲	−۰/۱۴۲	۰/۳۷۹	۰/۳۵۱	۰/۸۲۰۵
	پیش‌گویی	۰/۹۴۴۲	۰/۷۸۶۴				
π	واقعی	۰/۴۲۸۳	۰/۲۳۷۰	−۰/۰۰۷	۰/۱۳۱	۰/۱۳۲	۰/۸۵۴۷
	پیش‌گویی	۰/۳۵۳۳	۰/۲۷۱۴				

نتایج ارائه‌شده در جدول ۲ نشان می‌دهد که مدل در برآورد پارامتر π عملکرد مطلوبی داشته است؛ به‌طوری‌که مقدار اریبی بسیار ناچیز و برابر با ۰/۰۰۷− و خطای استاندارد برابر با ۰/۱۳۲ است. این امر بیانگر آن است که مدل توانسته است به‌طور دقیق احتمال وقوع صفرهای اضافی در داده‌ها را شناسایی کند.در مورد پارامتر λ، اگرچه مقدار اریبی کمی بیشتر است (۰/۱۴۲−)، اما با توجه به مقدار خطای استاندارد (۰/۳۵۱) و RMSE که برابر با ۰/۳۷۹ است، برآورد همچنان در محدوده قابل قبولی قرار دارد. همچنین، نزدیکی مقدار RMSE به خطای استاندارد در هر دو پارامتر، نشان‌دهنده ثبات و پایداری مدل در برابر نوسانات تصادفی داده‌ها است. در مجموع، این شواهد عددی تایید می‌کند که مدل پیشنهادی از دقت مناسبی در پیش‌گویی پارامترهای ساختاری توزیع برخوردار است.برای ارزیابی عمیق‌تر عملکرد مدل پیشنهادی، تطابق آماری میان مقادیر برآوردی و مقادیر واقعی پارامترهای λ و π مورد بررسی قرار گرفت.

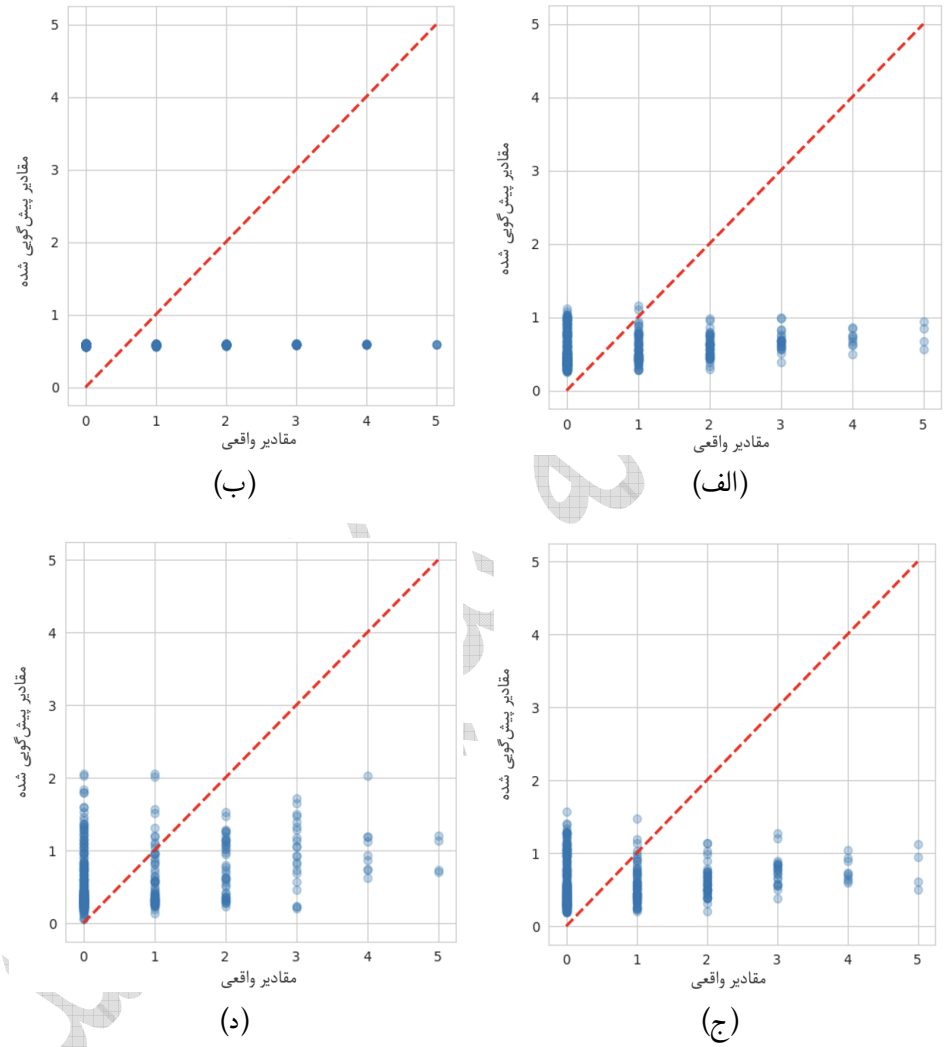
جدول ۲ آماره‌های توصیفی این پارامترها را در دو وضعیت واقعی و پیش‌گویی‌شده نشان می‌دهد. نتایج نشان می‌دهد که میانگین مقادیر پیش‌گویی‌شده برای پارامتر λ (۰٫۹۴۴۲) در مقایسه با میانگین واقعی (۱٫۰۴۵۳) و برای پارامتر π (۰٫۳۵۳۳) در مقایسه با مقدار واقعی (۰٫۴۲۸۳) بیانگر آن است که مدل تمایل به برآورد کمی پایین‌تر از مقادیر واقعی دارد، هرچند این انحراف در دامنه قابل قبولی قرار دارد. انحراف معیار مقادیر پیش‌گویی‌شده برای λ (۰٫۷۸۶۴) و π (۰٫۲۷۱۴) به‌ترتیب به انحراف معیار مقادیر واقعی (۰٫۷۷۹۲ و ۰٫۲۳۷۰) بسیار نزدیک است که نشان‌دهنده توانایی مدل در حفظ پراکندگی طبیعی داده‌هاست.همچنین بر اساس ستون آخر در جدول ۲، ضریب همبستگی پیرسون میان مقادیر واقعی و پیش‌گویی‌شده برای λ برابر با ۰٫۸۲۰۵ و برای π برابر با ۰٫۸۵۴۷ به‌دست آمد که حاکی از همبستگی قوی و معنادار بین برآوردهای مدل و مقادیر واقعی است. به‌عنوان نمونه، جدول ۳ ده مشاهده اول از مقادیر واقعی و پیش‌گویی‌شده را نشان می‌دهد که تطابق نسبی مناسبی را میان برآوردهای مدل و مقادیر واقعی تأیید می‌کند. این یافته‌ها، همراه با مقادیر اریبی و RMSE گزارش‌شده در جدول ۲، مؤید آن است که مدل رگرسیونی شبکه عصبی پواسون آماسیده در صفر نه‌تنها در پیش‌گویی خروجی، بلکه در برآورد ساختار پارامتریک توزیع نیز از دقت و پایایی مطلوبی برخوردار است.

جدول ۳: نمونه‌هایی از مقادیر واقعی و پیش‌گویی‌شده پارامترها (۱۰ مشاهده اول)

مشاهده	λ واقعی	λ پیش‌گویی	π واقعی	π پیش‌گویی
۱	۰٫۱۷۴۱	۰٫۱۸۹۶	۰٫۰۷۵۹	۰٫۰۰۱۲
۲	۰٫۴۷۷۹	۰٫۶۳۸۱	۰٫۶۶۷۰	۰٫۷۳۶۷
۳	۰٫۷۷۲۸	۰٫۶۷۸۱	۰٫۰۷۴۵	۰٫۰۰۸۵
۴	۰٫۸۸۵۲	۰٫۶۳۰۱	۰٫۱۸۳۸	۰٫۰۰۹۴
۵	۱٫۷۵۴۵	۱٫۵۳۴۵	۰٫۴۹۰۲	۰٫۴۸۶۱
۶	۰٫۴۱۷۲	۰٫۳۲۶۹	۰٫۳۰۳۸	۰٫۲۲۰۰
۷	۱٫۷۱۱۰	۱٫۲۶۳۲	۰٫۵۹۵۴	۰٫۶۹۵۹
۸	۱٫۱۳۰۸	۱٫۹۱۵۴	۰٫۸۵۶۷	۰٫۷۹۵۷
۹	۱٫۰۲۶۷	۰٫۸۸۶۰	۰٫۳۳۹۵	۰٫۰۳۸۱
۱۰	۰٫۳۹۲۸	۰٫۲۷۴۷	۰٫۳۲۱۲	۰٫۲۹۰۲

نمودارهای شکل ۳، عملکرد چهار مدل مختلف را در پیش‌گویی مقادیر شمارشی نشان می‌دهند. در هر نمودار، محور افقی شامل مقادیر واقعی y و محور عمودی شامل مقادیر پیش‌گویی شده $\hat{y}$ است. خط‌چین با شیب یک ($y = x$) نشان‌دهنده حالت ایده‌آل است که در آن تمام پیش‌گویی‌ها دقیقا با مقادیر واقعی مطابقت دارند. در رگرسیون پواسون ساده بیشتر نقاط در نزدیکی صفر قرار دارند و در مقادیر بالاتر (مثل ۲،

مقاله پذیرفته شده



شکل ۳: مقایسه مقادیر پیش‌گویی شده در مقابل مقادیر واقعی در مدل الف: رگرسیون پواسون، ب: رگرسیون شبکه عصبی پواسون، ج: رگرسیون ZIP و د: رگرسیون شبکه عصبی ZIP.

۳، ۴، ۵) مجموعه‌ای از نقاط پراکنده و نسبتاً پایین‌تر از خط ایده‌آل مشاهده می‌شوند. این نشان می‌دهد که مدل پواسون ساده، هرچند در پیش‌گویی صفرها به نسبت خوب عمل می‌کند، اما در پیش‌گویی مقادیر مثبت بالاتر ضعیف است.

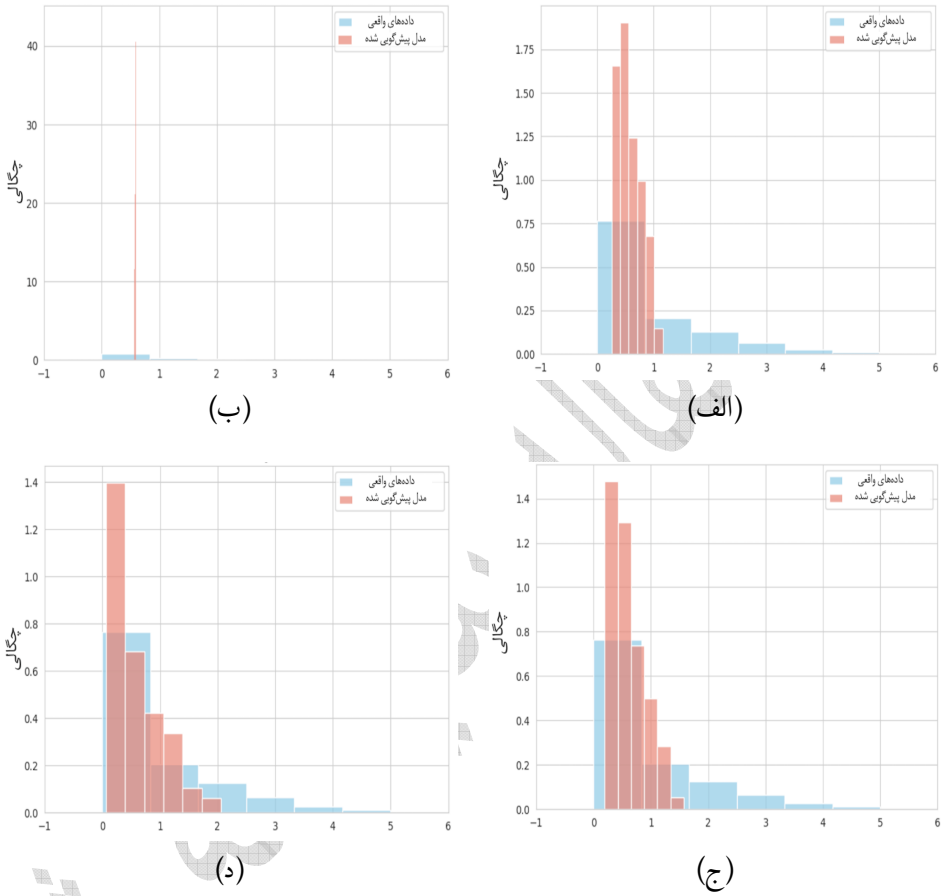
در رگرسیون شبکه عصبی پواسون نسبت به مدل ساده، رفتار منظم‌تری دارد، اما هم‌چنان در مقادیر بالاتر از ۱، پیش‌گویی‌های بسیار پایین و ثابت است. بیشتر نقاط در محدوده ۰/۵ تا ۰/۷ قرار دارند و همگی به سمت یک مقدار ثابت کشیده شده‌اند. این نشان‌دهنده این است که شبکه عصبی بدون در نظر گرفتن ساختار ZIP، قادر به یادگیری توزیع واقعی داده نیست و ممکن است به دلیل از بین رفتن انعطاف‌پذیری در پیش‌گویی مقادیر بالاتر، به سمت یک مقدار میانگین جمع‌آوری شود.

در مدل ZIP به‌طور قابل توجهی بهتر از دو مدل قبل عمل می‌کند. نقاط در اطراف خط ایده‌آل جمع شده‌اند و برای مقادیر ۰ تا ۳، پیش‌گویی‌ها نسبتا دقیق هستند. این نشان‌دهنده این است که مدل ZIP توانسته است مؤلفه صفر ساختاری را به‌طور جداگانه مدل‌سازی کند و به‌طور کلی توزیع داده را بهتر درک کند. با این حال، در مقادیر بالاتر (مانند ۴ و ۵)، همچنان نوساناتی وجود دارد که نشان‌دهنده محدودیت‌های مدل ZIP در یادگیری روابط غیرخطی است.

در مدل رگرسیونی شبکه عصبی ZIP بهترین عملکرد مشاهده می‌شود. نقاط به‌طور گسترده‌ای در اطراف خط ایده‌آل قرار دارند و هم در صفرها و هم در مقادیر بالاتر (حتی تا ۵) پیش‌گویی‌های دقیقی ارائه می‌دهد. همچنین، توزیع نقاط در هر کلاس از مقادیر واقعی (مثلاً ۰، ۱، ۲، ...) نشان می‌دهد که مدل توانسته است هر دو مؤلفه احتمال صفر ساختاری و توزیع پواسونی برای مقادیر مثبت را به‌صورت هم‌زمان و با دقت بالا مدل‌سازی کند. این رفتار نشان‌دهنده توانایی شبکه عصبی در یادگیری روابط غیرخطی و هم‌زمان با در نظر گرفتن ساختار داده است.

نمودارهای شکل ۴، توزیع داده‌های واقعی (با رنگ آبی) و توزیع پیش‌گویی‌های هر یک از چهار مدل (با رنگ قرمز) را با استفاده از هیستوگرام‌های چگالی [۲] مقایسه می‌کنند. محور افقی شامل مقادیر شمارشی ($y = ۰, ۱, ۲, \ldots$) و محور عمودی نشان‌دهنده چگالی نسبی [۳] است. هدف این نمودارها، بررسی این است که آیا مدل‌ها توانسته‌اند توزیع واقعی داده‌ها را به خوبی شبیه‌سازی کنند یا خیر. این نمودارها به‌وضوح نشان می‌دهند که مدل‌های ساده (مثل پواسون) و شبکه‌های عصبی بدون ساختار ZIP، نمی‌توانند توزیع واقعی داده‌هایی با فرایند آماسیده در صفر را شبیه‌سازی کنند. به طور دقیق‌تر می‌توان گفت در رگرسیون پواسون ساده توزیع پیش‌گویی مدل (قرمز) به‌طور قابل توجهی از توزیع واقعی (آبی) فرق می‌کند. مدل

[2] density histograms
[3] relative density



شکل ۴: مقایسه توزیع داده‌های واقعی و پیش‌گویی شده در مدل الف: رگرسیون پواسون، ب: رگرسیون شبکه عصبی پواسون، ج: رگرسیون ZIP و د: رگرسیون شبکه عصبی ZIP.

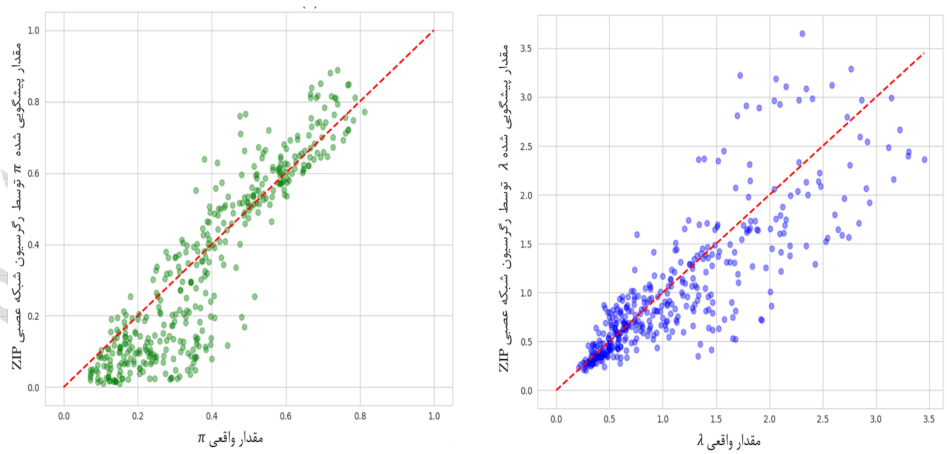
پیش‌گویی‌های خود را بیشتر در بازه [۰ , ۱] متمرکز کرده و از مقادیر بالاتر (مثل ۲، ۳، ۴) به طور کلی کمتر پیش‌گویی می‌کند. این نشان می‌دهد که مدل پواسون ساده نمی‌تواند فرایند آماسیده در صفر را به درستی مدل‌سازی کند و در نتیجه، احتمال صفر را بالاتر از واقعیت و احتمال مقادیر مثبت را پایین‌تر از واقعیت پیش‌گویی می‌کند. این مشکل ناشی از عدم در نظر گرفتن مؤلفه ساختاری صفر است.

رگرسیون شبکه عصبی پواسون از بدترین عملکرد در این نمودار برخوردار است. در مقادیر مثبت تقریبا پیش‌گویی نمی‌کند. این نشان‌دهنده این است که مدل به‌طور کامل به سمت پیش‌گویی صفر متمایل شده و

قادر به تولید مقادیر مثبت نیست. این رفتار معمولا ناشی از کاهش انعطاف‌پذیری شبکه در حضور داده‌های آماسیده در صفر است.

مدل ZIP به‌طور قابل توجهی بهتر از دو مدل قبل عمل می‌کند. توزیع پیش‌گویی‌ها در مقادیر ۰ و ۱ نسبتا نزدیک به توزیع واقعی است و حتی در مقادیر بالاتر (مثل ۲ و ۳)، مقادیر پیش‌گویی شده وجود دارد. این مسئله نشان می‌دهد که مدل توانسته است مؤلفه صفر ساختاری را به‌درستی مدل‌سازی کند و در نتیجه، توزیع پیش‌گویی‌ها به‌طور کلی با داده واقعی هم‌سویی دارد. با این حال، در مقادیر بالاتر، تفاوت‌هایی هم‌چنان وجود دارد که نشان‌دهنده محدودیت‌های مدل در یادگیری روابط غیرخطی است.

رگرسیون شبکه عصبی ZIP دارای بهترین تطابق با توزیع داده واقعی است. همان‌طور که در شکل ۴ مشاهده می‌شود، توزیع پیش‌گویی‌ها به‌طور گسترده‌ای در تمام دامنه‌ها (صفر تا ۵) با توزیع واقعی هم‌سویی دارد. این مدل نه تنها توانسته است تعداد صفرها را به‌درستی پیش‌گویی کند، بلکه در مقادیر مثبت نیز توزیع دقیقی ارائه می‌دهد. این مسئله نشان‌دهنده این است که شبکه عصبی ZIP توانسته است احتمال صفر ساختاری را به‌طور جداگانه پیش‌گویی کند و هم‌زمان، توزیع پواسونی برای مقادیر مثبت را با دقت بالا مدل‌سازی کند.



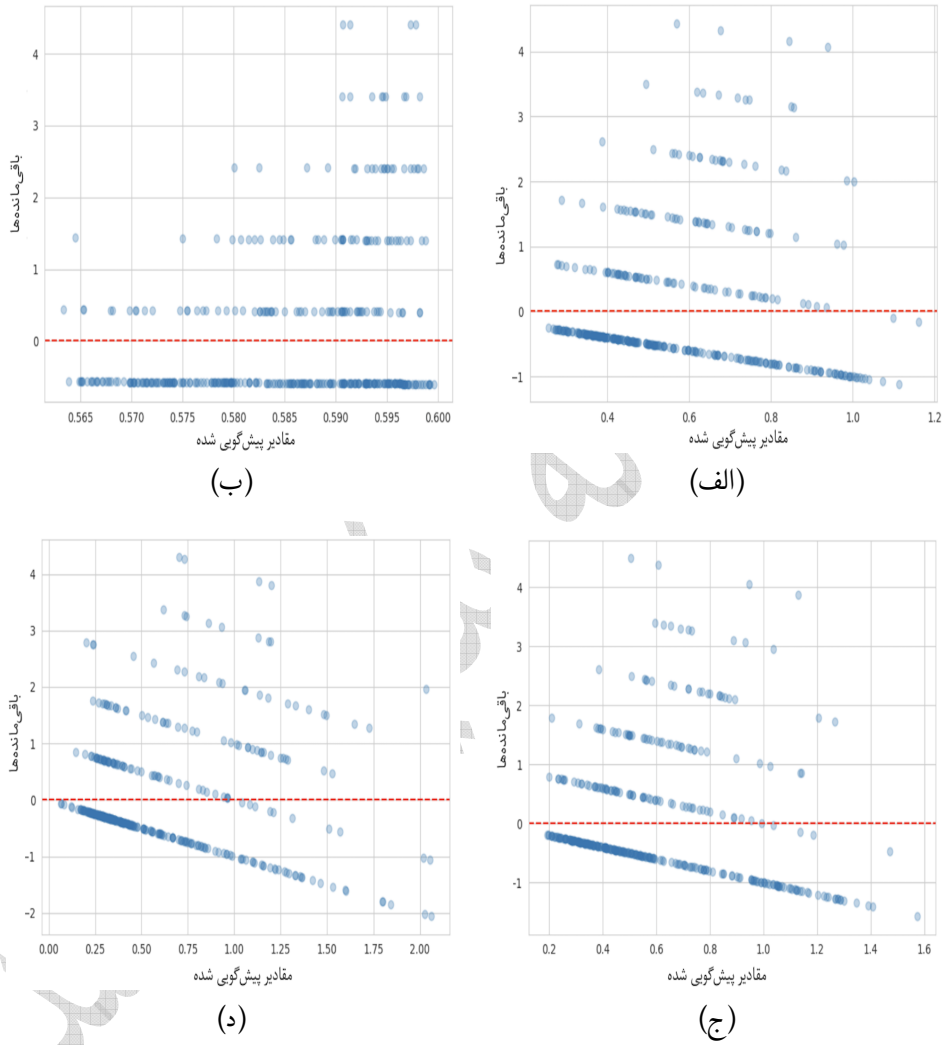
(الف) (ب)

شکل ۵: یادگیری پارامترها در مدل رگرسیونی شبکه عصبی پواسون آماسیده در صفر در مقابل مقدار واقعی الف: λ و ب: π.

در شکل ۵ عملکرد شبکه عصبی پواسون آماسیده در صفر در یادگیری دو پارامتر کلیدی مدل ZIP یعنی λ (میانگین توزیع پواسون) و π (احتمال صفر ساختاری) را با مقایسه با مقادیر واقعی [۴] که در فرآیند شبیه‌سازی داده استفاده شدند، نشان می‌دهند. هر دو نمودار، مقادیر واقعی را در محور افقی و مقادیر پیش‌گویی‌شده توسط مدل در محور عمودی قرار داده‌اند. خط‌چین با شیب یک ($y = x$) نشان‌دهنده حالت ایده‌آل است که در آن تمام پیش‌گویی‌ها دقیقا با مقادیر واقعی مطابقت دارند. در نمودار چپ یعنی یادگیری λ، بیشتر نقاط در نزدیکی خط ایده‌آل قرار دارند، اما با وجود تغییراتی در دامنه بالاتر (به‌ویژه در $\lambda > ۲$)، می‌توان گفت که مدل توانسته است رابطه بین ویژگی‌ها و میانگین پواسون را به‌طور کلی یاد بگیرد. در دامنه‌های کمتر از ۱/۵، تطابق بسیار خوبی مشاهده می‌شود، اما در مقادیر بالاتر، بعضی از پیش‌گویی‌ها از خط ایده‌آل دورتر هستند. این مسئله نشان می‌دهد که شبکه عصبی توانسته است رابطه غیرخطی بین متغیرهای ورودی و λ را یاد بگیرد، اما در برخی موارد، کمی از دقت خود را از دست داده است. در نمودار راست یعنی یادگیری π توزیع نقاط نشان می‌دهد که بیشتر مقادیر پیش‌گویی‌شده نزدیک به خط ایده‌آل هستند، به‌خصوص در دامنه‌های متوسط و بالای π. حتی در مقادیر پایین (مثل $\pi < ۰/۲$) و بالا (مثل $\pi > ۰/۷$)، تطابق مناسبی وجود دارد. این نشان‌دهنده این است که شبکه توانسته است رابطه لجستیکی بین ویژگی‌ها و احتمال صفر ساختاری را به‌طور دقیق یاد بگیرد. نمودارهای باقی‌مانده‌ها در شکل ۶ نشان می‌دهند که چگونه هر مدل در پیش‌گویی داده‌ها عمل کرده است. هرچه پراکندگی نقاط دورتر از خط صفر و با الگوی مشخص باشد، مدل ضعیف‌تر است. درواقع باقی‌مانده‌ها در تمام مدل‌های ارائه شده حول محور صفر متمرکز شده‌اند و هیچ روند یا الگوی سیستماتیک مشخصی مانند رفتار سینوسی در توزیع آن‌ها مشاهده نمی‌شود. این امر بیانگر آن است که بخش خطای مدل دارای ماهیت تصادفی است. علاوه بر این، عدم وجود همبستگی معنادار در باقی‌مانده‌ها نشان‌دهنده فرض صحیح استقلال مشاهدات است. این تحلیل‌ها تایید می‌کند که مدل‌ها از برازش مناسبی برخوردار است و خطاهای آن در محدوده قابل قبول آماری قرار دارند.

به طور دقیق‌تر مدل پواسون ساده و رگرسیون شبکه عصبی پواسون هر دو الگوهای واضحی از خطا دارند بدان معنا که یا همیشه صفر را بیشتر یا کمتر از واقعیت پیش‌گویی می‌کنند، یا تمام پیش‌گویی‌ها را به

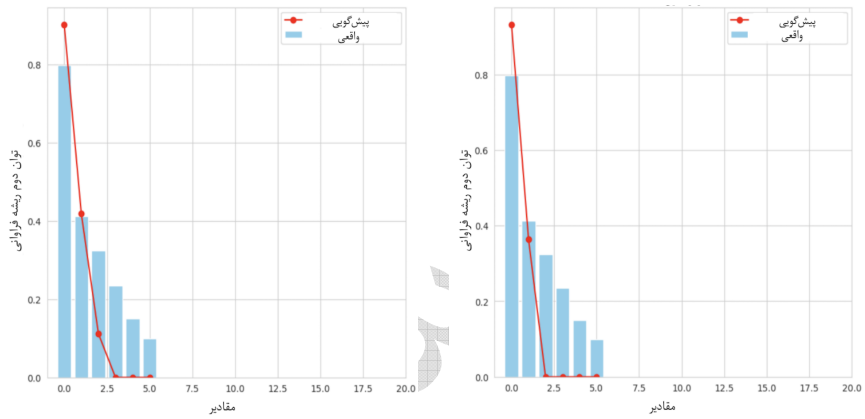
[4] True



شکل ۶: باقی‌مانده‌ها در مقابل مقادیر پیش‌گویی شده در مدل الف: رگرسیون پواسون، ب: رگرسیون شبکه عصبی پواسون، ج: رگرسیون ZIP و د: رگرسیون شبکه عصبی ZIP.

یک عدد ثابت می‌رسانند. این نشان می‌دهد که این مدل‌ها نمی‌توانند طبیعت واقعی داده با تعداد زیادی صفر و مقادیر مثبت را درست بفهمند. با دقت بیشتر در مقیاس عددی محور افقی، مدل ZIP بهتر است، اما هنوز کمی در پیش‌گویی مقادیر کوچک دچار خطاست. در مقابل، رگرسیون شبکه عصبی ZIP بهترین

عملکرد را دارد. خطاها به‌صورت تصادفی و بدون هیچ الگوی مشخصی در اطراف صفر پراکنده شده‌اند. این یعنی مدل هم صفرها را درست تشخیص داده، هم مقادیر مثبت را خوب پیش‌گویی کرده و به طور کلی رفتار داده را درست فهمیده است.



(ب) (الف)

شکل ۷: نمودارهای روتوگرام: توزیع پیش‌گویی شده و واقعی در مدل الف: رگرسیون ZIP و ب: رگرسیون شبکه عصبی ZIP.

در شکل ۷ هدف نمودار این است که ارزیابی شود مدل چقدر توانسته است توزیع واقعی داده را شبیه‌سازی کند. اگر خط قرمز (پیش‌گویی) دقیقا روی ستون‌های آبی (واقعی) قرار گیرد، یعنی مدل تمامی فرکانس‌ها را به درستی پیش‌گویی کرده است. مدل ZIP در پیش‌گویی صفرها خوب عمل می‌کند، اما در مقادیر بالاتر (مثل ۲، ۳، ۴) کمتر از واقعیت پیش‌گویی می‌کند و از تعداد واقعی آن‌ها عقب می‌ماند. در مقابل، رگرسیون شبکه عصبی ZIP تقریبا در همه جا به خوبی با داده واقعی هماهنگ است. هم صفرها را درست پیش‌گویی می‌کند، هم مقادیر مثبت را. این نشان می‌دهد که این مدل نه‌تنها عدد درستی می‌گوید، بلکه الگوی کلی داده را هم درست فهمیده است.

۶ مدل‌بندی داده‌های تعداد مقالات

داده‌ها به‌صورت تصادفی به دو بخش مجموعه‌های آموزشی و آزمایشی تقسیم شدند که برای توسعه و برازش مدل‌های و ارزیابی عملکرد مدل‌ها و سنجش قابلیت تعمیم‌پذیری آن‌ها استفاده شد. همچنین، استفاده از مجموعه آزمایشی مستقل، امکان مقایسه منصفانه بین مدل‌های مختلف را فراهم می‌سازد و اطمینان حاصل می‌شود که یافته‌های تحقیق نه تنها در داده‌های فعلی، بلکه در شرایط جدید و مشابه نیز قابل تعمیم هستند.

به طور دقیق می‌توان گفت فرآیند برآورد پارامترهای مدل، متشکل از وزن‌ها و اریبی‌ها، با دقت و بر پایه اصول یادگیری عمیق اجرا شد. به‌منظور بهینه‌سازی این پارامترها، از الگوریتم پس‌انتشار خطا در کنار روش گرادیان خطا بهره برده شد تا تابع زیان منفی لگاریتم درست‌نمایی که ملاک دقت مدل است، به حداقل برسد. آغاز فرآیند آموزش با مقداردهی اولیه تصادفی پارامترها صورت گرفت تا از افتادن در بهینه‌های محلی جلوگیری شود. در هر گام تکرار، پارامترها به‌گونه‌ای تنظیم شدند که خطای مدل کاهش یابد. برای تضمین تعمیم‌پذیری و پرهیز از پدیده بیش‌برازش، استراتژی توقف زودهنگام به کار گرفته شد. بدین منظور، مجموعه داده به دو بخش آموزش (۸۰ درصد) و اعتبارسنجی (۲۰ درصد) تقسیم گردید و فرآیند آموزش در نقطه‌ای متوقف شد که خطای مجموعه اعتبارسنجی روند صعودی گرفت. علاوه بر این، جهت همگرایی پایدار و دقیق پارامترها به مقادیر بهینه سراسری، از نرخ یادگیری تطبیقی‌پذیر استفاده شد. نتایج نهایی حاکی از آن بود که پارامترهای مدل پس از عبور از مراحل همگرایی به ثبات رسیدند و مدل پیشنهادی موفق شد ساختار پیچیده و نهفته در داده‌های واقعی را با کارایی بالا بیاموزد. کدهای مربوط به پیاده‌سازی مدل روی داده‌های واقعی در دسترس پژوهشگران قرار گرفته است[۵] همچنین کدهای بخش شبیه‌سازی نیز به طور جداگانه منتشر شده است[۶]. نتایج حاصل از اجرای فرایند مدل‌سازی و ارزیابی در جدول ۴، به‌طور مشخص نشان‌دهنده برتری مدل رگرسیونی شبکه عصبی پواسون آماسیده در صفر نسبت به دو مدل پایه رگرسیون پواسون و مدل ZIP در پیش‌گویی تعداد مقالات منتشرشده توسط دانشجویان دکتری شیمی بیولوژیک است.

[5] https://github.com/TaranomTN/Zero-Inflated-Poisson-Neural-Network-bioChemists

[6] https://github.com/TaranomTN/ZIP-NN-Simulation

تحلیل متغیر وابسته تعداد مقالات نشان داد که ۳۰٫۰۵ درصد از مشاهدات صفر هستند. این مقدار بالا از صفرها، نشان‌دهنده وجود یک فرایند دوگانه در داده است. یک گروه از افراد که به‌طور ساختاری هیچ مقاله‌ای منتشر نمی‌کنند و گروه دیگری که فعالیت علمی دارند اما تعداد مقالات آن‌ها تصادفی است.

جدول ۴: مقایسه عملکرد مدل‌های مختلف در پیش‌گویی تعداد مقالات منتشرشده

مدل	MAE	MSE	RMSE
رگرسیون شبکه عصبی ZIP	۱٫۲۲	۲٫۹۵	۱٫۷۲
رگرسیون پواسون	۱٫۳۰	۳٫۰۵	۱٫۷۵
رگرسیون ZIP	۱٫۳۰	۳٫۲۸	۱٫۸۱

نتایج حاصل از مقایسه مدل‌های مختلف نشان می‌دهد که مدل رگرسیونی شبکه عصبی ZIP در تمامی معیارهای ارزیابی عملکرد بهتری نسبت به مدل‌های رگرسیون پواسون و ZIP دارد؛ به‌گونه‌ای که با داشتن کمترین مقدار MAE (۱٫۲۲)، MSE (۲٫۹۵) و RMSE (۱٫۷۲)، نه تنها دقت بالاتری در پیش‌گویی میانگین تعداد مقالات دارد، بلکه توانسته است خطاهای بزرگ را به‌طور مؤثرتری کنترل کند. این برتری به‌ویژه در معیارهای RMSE و MSE آشکار می‌شود که نسبت به خطاها با اندازه زیاد حساس‌تر هستند و نشان می‌دهند که مدل پیشنهادی قادر است توزیع واقعی داده به‌ویژه در بخش صفرها و دم توزیع را با دقت بیش‌تری بازتولید کند. در حالی‌که مدل‌های کلاسیک با فرض‌های خطی به مرزهای خود محدود می‌شوند، مدل رگرسیونی شبکه عصبی پواسون آماسیده در صفر با بهره‌گیری از قابلیت شبکه‌های عصبی در یادگیری روابط غیرخطی و اثرات تعاملی پیچیده، بدون نیاز به تعریف صریح این روابط، توانسته است ساختار پنهان داده را به‌خوبی کشف کند. این یافته‌ها اهمیت تلفیق مدل‌های آماری معتبر با قابلیت‌های یادگیری ماشین را در تحلیل داده‌های شمارشی با فرایند وجود صفرهای زیاد تایید می‌کند.

۷ بحث و نتیجه‌گیری

این پژوهش با هدف رفع محدودیت‌های مدل‌های کلاسیک در تحلیل داده‌های شمارشی با صفرهای ساختاری، چارچوبی نوین بر پایه شبکه‌های عصبی عمیق برای رگرسیون پواسون آماسیده در صفر ارائه داد. نتایج حاصل از شبیه‌سازی‌ها و کاربرد داده‌های واقعی نشان داد که این معماری ترکیبی، عملکرد به

مراتب دقیق‌تری در مقایسه با رگرسیون پواسون کلاسیک و رگرسیون شبکه عصبی پواسون دارد.اهمیت این مطالعه در ایجاد پیوند میان ساختار تفکیک‌پذیر مدل‌های آماری و قدرت پیش‌گویی هوش مصنوعی است. این رویکرد با عبور از فرض محدودکننده خطی‌بودن، کارایی بالایی در تحلیل داده‌های پیچیده حوزه‌های اقتصاد، اپیدمیولوژی و علوم اجتماعی نشان می‌دهد.کارایی محاسباتی و مدیریت محدودیت‌هاپیچیدگی محاسباتی الگوریتم به‌روزرسانی پارامترها در این مدل، مشابه سایر شبکه‌های عصبی، با تعداد پارامترها و حجم داده‌ها رابطه‌ای خطی دارد؛ به طوری که پیچیدگی زمانی آن در هر مرحله از مرتبه $O(N \times P)$ است (که در آن N تعداد مشاهدات و P مجموع وزن‌ها و اریبی‌های شبکه است). با این حال، استفاده از روابط بازگشتی برای محاسبات مرتبه دوم، نیاز به محاسبه مستقیم ماتریس هسین را برطرف کرده و کارایی محاسباتی مدل را برای داده‌های با ابعاد بالا تضمین می‌کند.از سوی دیگر، چالش‌هایی نظیر ماهیت جعبه‌سیاه لایه‌های پنهان، وابستگی به تنظیم ابرپارامترها و نیاز به حجم نمونه بزرگ در این معماری وجود دارد. برای تعدیل این محدودیت‌ها در شرایط کمبود داده، می‌توان با به‌کارگیری تکنیک‌های منظم‌سازی (مانند حذف نورون‌ها و جریمه‌سازی تابع زیان) و ساده‌سازی معماری شبکه، پیچیدگی مدل را مدیریت کرد، از بیش‌برازش جلوگیری نمود و تعمیم‌پذیری مدل را ارتقا داد.

این مطالعه مسیرهای جدیدی را برای ادامه پژوهش‌ها باز می‌کند. نخست، می‌توان چارچوب ارائه‌شده را به مدل‌های دیگر مانند دو جمله‌ای منفی آماسیده در صفر برای داده‌های دارای بیش‌پراکنش گسترش داد. همچنین استفاده از معماری‌های پیشرفته‌تر مانند شبکه‌های بازگشتی برای داده‌های سری زمانی یا شبکه‌های پیچشی برای داده‌های مکانی، می‌تواند دامنه کاربرد این روش را گسترش دهد.

تقدیر و تشکر

نویسندگان بدین‌وسیله مراتب تشکر و قدردانی عمیق خود را از سردبیر محترم مجله، داوران گرامی و ویراستار ارجمند ابراز می‌دارند. پیشنهادات علمی دلسوزانه و دقت نظر داوران محترم، همراه با تلاش‌های ارزشمند سردبیر و ویراستار در ویرایش ادبی و ساختاری مقاله، نقشی بی‌بدیل در ارتقای کیفیت علمی و رفع ابهامات این پژوهش داشته است که صمیمانه مورد سپاسگزاری نویسندگان می‌باشد.

مراجع

مددی، م. و مترجم، ک. (۱۴۰۳)، تحلیل داده‌های بقا با روش‌های مختلف یادگیری آماری، مجله علوم آماری، **۱۸**(۲)، ۴۱۵-۴۳۴.

Ashraf, M., Zaman, M., and Ahmed, M. (2020), An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering Approaches, *Procedia Computer Science*, **167**, 1471–1483.

Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L., and Kirchner, U. (1999), The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **162**(2), 195–209.

Fallah, N., Gu, H., Mohammad, K., Seyyedsalehi, S. A., Nourijelyani, K., and Eshraghian, M. R. (2009), Nonlinear Poisson Regression Using Neural Networks: A Simulation Study, *Neural Computing and Applications*, **18**(8), 945–954.

Fallah, N., Mitnitski, A., and Rockwood, K. (2011), Applying Neural Network Poisson Regression to Predict Cognitive Score Changes, *Journal of Applied Statistics*, **38**(9), 2051–2062.

Haghani, S., Sedehi, M., and Kheiri, S. (2017), Artificial Neural Network to Modeling Zero-Inflated Count Data: Application to Predicting Number of Return to Blood Donation, *Journal of Research in Health Sciences*, **17**(3), 392.

Hall, D. B. (2000), Zero-inflated Poisson and Binomial Regression with Random Effects: A Case Study, *Biometrics*, **56**(4), 1030–1039.

Lambert, D. (1992), Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing, *Technometrics*, **34**(1), 1–14.

Lee, H., Ha, I. D., and Lee, Y. (2023), Deep Neural Networks for Semiparametric Frailty Models via H-likelihood, *arXiv preprint arXiv:2307.06581*.

Montesinos-López, O. A., Montesinos-López, J. C., Salazar, E., Barrón, J. A., Montesinos-López, A., Buenrostro-Mariscal, R., and Crossa, J. (2021), Application of a Poisson Deep Neural Network Model for the Prediction of Count Data in Genome-Based Prediction, *The Plant Genome*, **14**(2), e20118.

Montesinos-López, O. A., et al. (2022), Comparing Gradient Boosting Machine and Bayesian Threshold BLUP for Genome-Based Prediction of Categorical Traits in Wheat Breeding, *The Plant Genome*, **15**(1), e20214.

Ridout, M., Hinde, J., and Demétrio, C. G. B. (2001), A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives, *Biometrics*, **57**(1), 219–223.

Rizvi, S., Pettee, M., and Nachman, B. (2024), Learning Likelihood Ratios With Neural Network Classifiers, *Journal of High Energy Physics*, **2024**(2), 136.

Rodrigo, H., and Tsokos, C. (2020), Bayesian Modelling of Nonlinear Poisson Regression with Artificial Neural Networks, *Journal of Applied Statistics*, **47**(5), 757–774.

Rodrigo, H. S. (2017), *Bayesian Artificial Neural Networks in Health and Cybersecurity*, Doctoral Dissertation, University of South Florida.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986), Learning Representations by Back-Propagating Errors, *Nature*, **323**(6088), 533–536.

Sharaf, T. (2015), *Statistical Learning with Artificial Neural Network Applied to Health and Environmental Data*, Doctoral Dissertation, University of South Florida.

Sharaf, T., and Tsokos, C. P. (2015), Two Artificial Neural Networks for Modeling Discrete Survival Time of Censored Data, *Advances in Artificial Intelligence*, **2015**, 270165.

مقاله پذیرفته شده

Count Data Modeling Using Zero-Inflated Poisson Regression with Deep Learning

Torabi Neman, T. , Emadi, M., Arashi, M.

[1]*Department of Statistics, Faculty of Mathematical Science, Ferdowsi University, Mashhad, Iran.*

Abstract: In the broad landscape of modern data analysis, dealing with count data affected by excessive zeros represents a fundamental analytical challenge. Classical models such as Poisson regression exhibit notable weaknesses in this context, as they cannot distinguish between structural zeros (arising from deterministic processes) and random zeros (arising from stochastic processes). Although zero-inflated Poisson regression models have taken a significant step toward addressing this issue, their performance is seriously limited in the era of high-dimensional and large-scale data. This research, looking beyond traditional frameworks, introduces an innovative deep neural network–based framework designed to enhance and transform zero-inflated Poisson regression models. By leveraging the remarkable capacity of deep learning to extract nonlinear and complex features, this hybrid approach can model the dual data-generating processes, the zero-generation process, and the count-generation process with unprecedented precision.

Keywords: Zero-Inflated Poisson Regression,Deep Poisson Regression, Neural Networks, Multilayer Perceptron, Count Data.

Mathematics Subject Classification (2020): 62J12, 68T07.