مجله مدیریت اطلاعات سلامت
سال هجدهم/ شماره اول/ فروردین و اردیبهشت ۱۴۰۰
Published by *Vesnu Publications*
DOI: 10.22122/him.v18i1.4194

جایگذاری مقادیر گمشده در مجموعه داده‌های دیابت و سرطان سینه با استفاده از شبکه عصبی پرسپترون دو لایه

الهام پورجانی[۱]، سارا نجف‌زاده[۲]، نادر جعفرنیا دابانلو[۳]

مقاله پژوهشی

چکیده

مقدمه: جایگذاری مقادیر گمشده در مجموعه داده‌های اطلاعاتی پزشکی، یکی از چالش‌های مهم در مسایل داده‌کاوی به شمار می‌رود. بنابراین، پژوهش حاضر با هدف جایگذاری مقادیر گمشده برخی از ویژگی‌های مجموعه داده‌های دیابت و سرطان سینه انجام شد.

روش بررسی: در این مطالعه توصیفی، از مجموعه داده سرطان سینه شامل ۶۹۹ نمونه که ۴۵۸ نمونه خوش‌خیم و ۲۴۱ نمونه بدخیم و مجموعه داده دیابت شامل ۷۶۸ نمونه که ۵۰۰ نمونه فاقد بیماری دیابت و ۲۶۸ نمونه دیگر دارای بیماری دیابت بودند، استفاده گردید. برای جایگذاری مقادیر گمشده در این دو مجموعه داده، مدلی بر پایه شبکه عصبی پرسپترون دو لایه طراحی شد. به منظور ارزیابی، ماشین بردار پشتیبان SVM (Support Vector Machine) و آزمون t مورد استفاده قرار گرفت.

یافته‌ها: میزان میانگین مربعات خطا MSE (Mean Squared Error) به دست آمده در مدل شبکه عصبی پرسپترون دو لایه در مجموعه داده دیابت، حدود ۰/۰۳ و در مجموعه داده سرطان سینه، حدود ۰/۰۴ کمتر از MSEهای به دست آمده در روش جایگذاری با مقدار میانگین گزارش گردید. مقادیر جایگذاری شده با استفاده از مدل نسبت به مقادیر جایگذاری شده با مقدار میانگین، به مقدار واقعی نزدیک‌تر بود. صحت و حساسیت طبقه‌بندی بیماری در حالتی که مقادیر گمشده توسط شبکه عصبی پرسپترون جایگذاری شده بود، در مقایسه با دو روش مرسوم مقدار میانگین و روش حذف مقادیر گمشده در مجموعه داده دیابت به ترتیب در حدود ۲، ۴، ۲ و ۴ درصد و در مجموعه داده سرطان سینه به ترتیب در حدود ۱، ۳، ۲، ۵ درصد بیشتر شد. تفاوت معنی‌داری بین دو روش جایگذاری مقادیر گمشده با مقدار میانگین و جایگذاری مدل وجود داشت.

نتیجه‌گیری: جایگذاری مقایر گمشده در مجموعه داده‌های پزشکی توسط شبکه عصبی پرسپترون دو لایه نسبت به دو روش جایگذاری با مقدار میانگین و روش حذف مقادیر گمشده، نتایج بهتری در طبقه‌بندی بیماری نشان می‌دهد.

واژه‌های کلیدی: داده‌کاوی؛ مدل‌های شبکه عصبی؛ ماشین بردار پشتیبان

پیام کلیدی: مدل پیشنهاد شده در مطالعه حاضر برای جایگذاری مقادیر گمشده در مجموعه داده‌های آزمایشگاهی و پزشکی ارایه گردید که به منظور پیش‌بینی و تشخیص دقیق‌تر بیماری دیابت و سرطان سینه قابل استفاده است؛ بدین معنی که مقادیر جایگذاری شده توسط الگوریتم ارایه شده در روش داده‌کاوی، می‌تواند جایگزین مناسبی برای مقادیر گمشده باشد.

دریافت مقاله: ۱۳۹۹/۵/۲۹ پذیرش مقاله: ۱۴۰۰/۱/۱۱ تاریخ انتشار: ۱۴۰۰/۱/۱۵

ارجاع: پورجانی الهام، نجف‌زاده سارا، جعفرنیا دابانلو نادر. **جایگذاری مقادیر گمشده در مجموعه داده‌های دیابت و سرطان سینه با استفاده از شبکه عصبی پرسپترون دو لایه.** مدیریت اطلاعات سلامت ۱۴۰۰؛ ۱۸ (۱): ۶-۱

مقدمه

یکی از چالش‌های اساسی در کار با داده‌ها، وجود مقادیر گمشده در مجموعه داده‌های اطلاعاتی می‌باشد. فقدان این اطلاعات، تجزیه و تحلیل ویژگی‌ها را دچار نقصان می‌نماید. داده‌های از دست رفته، یک مشکل رایج در بیشتر زمینه‌های تحقیقاتی است و عنصر ابهام را در تجزیه و تحلیل داده‌ها وارد می‌کند. آن‌ها می‌توانند به دلایل مختلف به وجود آیند که از آن جمله می‌توان به عدم استفاده صحیح از نمونه‌ها، خطای اندازه‌گیری و حذف اشتباه نمونه‌ها اشاره نمود (۱). در مجموعه داده‌های پزشکی و بیمارستانی، سوابق بیمار برای تشخیص و پیش‌بینی جمع‌آوری می‌شود و فقدان برخی از شاخص‌های مرتبط با بیمار، روند تشخیص بیماری را دچار مشکل می‌کند. در این‌گونه موارد، مشکلی تحت عنوان مقادیر گمشده وجود دارد. ویژگی‌های از دست رفته می‌توانند مقدار قابل توجهی از اریبی (Bias) را نشان دهند و مدیریت و تجزیه و تحلیل داده‌ها را دشوارتر کنند و باعث کاهش بهره‌وری شوند (۲).

استراتژی مقابله با مشکل مقادیر گمشده، انجام محاسبات داده به عنوان فرایندی به منظور تعیین مقادیر گمشده در یک مجموعه داده می‌باشد که توسط مقادیر مناسب محاسبه و تخمین زده می‌شوند. به عبارت دیگر، محاسبه داده قادر به پر کردن شکاف داده‌ها با اشتباهات عدم پاسخ است که مجموعه‌ای کامل از داده‌ها تولید کند. Folguera و همکاران با یک روش مبتنی بر نگاشت

مقاله حاصل پایان‌نامه مقطع کارشناسی ارشد می‌باشد که با حمایت دانشگاه آزاد اسلامی، واحد علوم و تحقیقات انجام شده است.

۱- دانشجوی کارشناسی ارشد، هوش مصنوعی و رباتیک، گروه مهندسی کامپیوتر، دانشکده مکانیک، برق و کامپیوتر، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
۲- استادیار، شبکه، گروه کامپیوتر، دانشکده مهندسی برق، واحد یادگار امام (ره)، دانشگاه آزاد اسلامی، شهرری، ایران
۳- دانشیار، الکترونیک، گروه مهندسی برق، دانشکده علوم و فن‌آوری‌های پزشکی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
نویسنده طرف مکاتبه: الهام پورجانی؛ دانشجوی کارشناسی ارشد، هوش مصنوعی و رباتیک، گروه مهندسی کامپیوتر، دانشکده مکانیک، برق و کامپیوتر، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
Email: e.pourjani@gmail.com

خودسازمانده، محاسبه داده را تحت مفهوم فاصله جسم در هر وزن برای پیش‌بینی شاخص‌های فیزیکی و شیمیایی نمونه‌های آب در یک مجموعه داده شامل ۲۷۰ نمونه که غلظت آنالیت‌های مختلف آن از دست رفته بود، بررسی و نتایج را با روش فاصله اقلیدسی مقایسه کردند که دقت نتایج حاصل از مطالعه بهتر بود (۳). Purwar و Singh به منظور جایگذاری مقادیر گمشده، مدلی از ترکیب خوشه‌بندی k-mean و شبکه پرسپترون چند لایه را ارایه کردند. مدل پیشنهادی آن‌ها در مقایسه با مدل‌های فازی، درخت تصمیم و رگرسیون، بالاترین صحت و حساسیت را داشت (۴). de Silva و Perera برای جایگذاری مقادیر گمشده، الگوریتم ژنتیک بهینه‌سازی k نزدیکترین همسایگی k-NN (k-Nearest Neighbors) را پیشنهاد و نتایج را با روش k-NN مقایسه نمودند. از بین این دو روش، الگوریتم ژنتیک بهینه‌سازی k-NN خطای کمتری داشت (۵). Jea و همکاران الگوریتمی را برای جایگذاری مقادیر گمشده ارایه دادند. این الگوریتم قواعدی را برای محاسبه ویژگی‌های از دست رفته توسط قانون انجمنی (Association Rule) ایجاد می‌کند و سپس از تابع فاصله جهت تنظیم قانون برای پر کردن مقادیر مناسب استفاده می‌کند. صحت روش ارایه شده از الگوریتم‌های C4.5 و k-NN، ۳ تا ۵ درصد بیشتر می‌باشد (۶).

Duan و همکاران یک مدل یادگیری عمیق به نام Denoising Autoencoders انباشته را برای محاسبه ویژگی‌های ترافیکی پیشنهاد کردند. آن‌ها یک الگوریتم برای تحقق کارامد یادگیری عمیق جهت محاسبه ویژگی‌های ترافیکی به وسیله آموزش مدل به صورت سلسله مراتبی با استفاده از مجموعه کامل ویژگی‌ها از تمام ایستگاه‌های شناسایی خودرو را توسعه و نتایج را با دو مدل شبکه عصبی BP (Backpropagation) و ARIMA (Autoregressive Integrated Moving Average) مقایسه کردند. صحت الگوریتم پیشنهادی نسبت به دو مدل BP و ARIMA بین ۵ تا ۵۰ درصد بیشتر است (۷). Deb و Liew الگوریتمی شبیه به دیگر الگوریتم‌های محاسبه مبتنی بر درخت تصمیم ارایه و نتایج را با الگوریتم‌های k-NN و C4.5 مقایسه کردند. صحت روش پیشنهادی از دو روش دیگر بالاتر و خطای روش پیشنهادی کمتر از دو روش دیگر بود (۸). Silva-Ramirez و همکاران برای جایگذاری مقادیر مفقود شده، مدلی بر پایه شبکه پرسپترون چند لایه ارایه نمودند و نتایج حاصل از جایگذاری، با روش‌های جایگذاری با مقدار میانگین، مد، رگرسیون و Hot-deck مقایسه گردید. از بین این روش‌ها، مدل پیشنهادی نسبت به سایر روش‌ها دقت بالاتری در جایگذاری مقادیر گمشده داشت (۹).

با توجه به این که مقادیر گمشده در مجموعه داده‌های پزشکی، روند پیشگیری و تشخیص بیماری را دچار مشکل می‌کند، ارایه مدل‌هایی به منظور پیش‌بینی مقادیر گمشده، از اهمیت بالایی برخوردار است. از سوی دیگر، داده‌کاوی مجموعه، روش‌هایی برای ارایه مدل‌های پیش‌بینی مقادیر گمشده دارد. بنابراین، در پژوهش حاضر ضمن بررسی مدل‌های پیش‌بینی مقادیر گمشده، مدلی با دقت بالاتر ارایه گردید.

روش بررسی

این مطالعه از نوع توصیفی بود و مجموعه داده‌های سرطان سینه و دیابت، از مخزن یادگیری ماشین UCI دریافت شد. مجموعه داده سرطان سینه، از بیمارستان‌های دانشگاه ویسکانسین جمع‌آوری گردید و ویژگی‌ها از طریق یک تصویر دیجیتالی با یک سوزن ظریف FNA (Fine Needle Aspiration) که برای برداشتن نمونه از توده پستان استفاده می‌شود، به دست آمد. آن‌ها ویژگی‌های هسته سلولی موجود در تصویر را توصیف می‌کنند و شامل ۶۹۹ نمونه (۴۵۸ نمونه خوش‌خیم و ۲۴۱ نمونه بدخیم) بود. مجموعه داده دیابت توسط انستیتوی ملی دیابت و گوارش و بیماری‌های کلیوی آمریکا جمع‌آوری شد که سوابق بیماران مبتلا به دیابت از دو منبع دستگاه ثبت الکترونیکی و ثبت کاغذی به دست آمد. دستگاه اتوماتیک دارای یک ساعت داخلی برای برچسب زنی وقایع بود؛ در حالی که سوابق کاغذی فقط اسلات‌های «زمان منطقی» (صبحانه، ناهار، شام، زمان خواب) را ارایه می‌کرد. برای سوابق کاغذی، زمان‌های مشخص به صبحانه (۰۸:۰۰) ، ناهار (۱۲:۰۰)، شام (۱۸:۰۰) و زمان خواب (۲۲:۰۰) اختصاص داده شد. این مجموعه شامل ۷۶۸ نمونه (۵۰۰ نمونه فاقد بیماری دیابت و ۲۶۸ نمونه دیگر دارای بیماری دیابت) بود. در مرحله پیش‌پردازش، مجموعه داده‌ها به روش Min-Max در نرم‌افزار Excel نسخه ۲۰۱۶ نرمال شد و مورد استفاده قرار گرفت. در مجموعه داده دیابت، به میزان ۲۰ درصد نمونه‌ها و بر روی چهار متغیر «تست تحمل گلوکز، عضلات سه‌سر، انسولین و شاخص جرم» و در مجموعه داده سرطان سینه نیز به میزان ۲۰ درصد نمونه‌ها و بر روی چهار متغیر «یکنواختی اندازه سلول، یکنواختی شکل سلول، اندازه سلول مخاطی تک و کروماتین مطلوب» به صورت کاملاً تصادفی داده گمشده ایجاد شده است. در جدول ۱ به تشریح این مجموعه داده‌ها پرداخته شده است.

مدلی بر پایه شبکه عصبی پرسپترون دو لایه طراحی شد. شبکه دو لایه پیچیدگی کمتر و قدرت تعمیم‌دهی بیشتری دارد. بر طبق محاسبات، برای لایه پنهان شش نورون انتخاب شد. مقادیر Epoch = ۱۰۰۰ و Mu = ۱/۰۰e + ۱۰ و از دو تابع فعال‌ساز Tansig برای لایه پنهان و Purelin برای لایه خروجی استفاده گردید. این دو تابع به پیشنهاد نرم‌افزار Matlab برگزیده شد.

توابع فعال‌سازی در قالب روابط ۱ و ۲ ارایه شده است.

$$tansig(n) = 2/(1+exp(2*n))-1 \quad \text{رابطه ۱}$$

$$purelin(n) = n \quad \text{رابطه ۲}$$

ویژگی‌های فاقد مقدار، خروجی شبکه و ویژگی‌های دارای مقدار به همراه برچسب کلاس (سالم یا بیمار) به عنوان ورودی شبکه عصبی پرسپترون دو لایه قرار می‌گیرد. برای آموزش شبکه، ابتدا نمونه‌های حاوی مقادیر گمشده از مجموعه داده‌ها حذف می‌شود. شبکه با استفاده از نمونه‌های کامل آموزش داده می‌شود (۷۰ درصد داده آموزش، ۱۵ درصد داده اعتبارسنجی و ۱۵ درصد داده تست) و سپس نمونه‌های حاوی مقادیر گمشده به شبکه ارایه می‌گردد. در نمونه‌های حاوی مقادیر گمشده، ویژگی‌های دارای مقدار به همراه برچسب کلاس به عنوان ورودی شبکه و ویژگی‌های فاقد مقدار به عنوان خروجی شبکه قرار می‌گیرد. بدین ترتیب، مقادیر گمشده توسط شبکه جایگذاری می‌شود. از آن‌جایی که مقادیر گمشده به صورت کاملاً تصادفی بر روی مجموعه داده‌های کامل ایجاد شده است، مقدار واقعی متغیرها در اختیار می‌باشد. از این‌رو، برای بررسی و مقایسه مقادیر جایگذاری شده توسط مدل با مقادیر واقعی، از میانگین مربعات خطا MSE (Mean Squared Error) بهره گرفته شده است. نتایج حاصل نشان دهنده این است که مقادیر جایگذاری شده در مجموعه داده دیابت و سرطان سینه به مقادیر واقعی نزدیک می‌باشند. نزدیک بودن مقادیر جایگذاری شده توسط مدل به مقادیر واقعی، باعث کاهش مقدار MSE می‌شود. به عبارت دیگر، هرچه مقدار MSE نزدیک به صفر باشد، مقدار جایگذاری شده توسط مدل به مقدار واقعی نزدیک‌تر است. نتایج به دست آمده از مدل با روش رایج جایگذاری با مقدار میانگین مقایسه می‌شود (۱۰).

جدول ۱: ویژگی‌های استفاده شده از مجموعه داده دیابت و سرطان سینه در ایجاد مدل پیش‌بینی مقادیر گمشده

مجموعه داده	نام ویژگی	بیشترین مقدار	کمترین مقدار	ویژگی‌های دارای مقدار گمشده
دیابت	تعداد دفعات بارداری	۱۷	۰	–
	تست تحمل گلوکز	۱۹۹	۰	گمشده به صورت کاملاً تصادفی
	فشار خون انبساطی	۱۲۲	۰	–
	عضلات سه سر	۹۹	۰	گمشده به صورت کاملاً تصادفی
	انسولین	۸۴۶	۰	گمشده به صورت کاملاً تصادفی
	شاخص جرم	۶۷/۱	۰	گمشده به صورت کاملاً تصادفی
	نژاد	۲/۴۲	۰/۰۷۸	–
	سن	۸۱	۲۱	–
	متغیر کلاس	۱	۰	–
سرطان سینه	ضخامت توده	۱۰	۱	–
	یکنواختی اندازه سلول	۱۰	۱	گمشده به صورت کاملاً تصادفی
	یکنواختی شکل سلول	۱۰	۱	گمشده به صورت کاملاً تصادفی
	چسبندگی حاشیه	۱۰	۱	–
	اندازه سلول مخاطی تک	۱۰	۱	گمشده به صورت کاملاً تصادفی
	شاخص حجم	۱۰	۱	–
	کروماتین مطلوب	۱۰	۱	گمشده به صورت کاملاً تصادفی
	هسته میان سلول نرمال	۱۰	۱	–
	میتوزها	۱۰	۱	–
	متغیر کلاس	۲	۱	–

برای ارزیابی عملکرد مدل با استفاده از طبقه‌بندی ماشین بردار پشتیبان SVM (Support Vector Machine)، صحت و حساسیت طبقه‌بندی (سالم یا بیمار) در حالتی که ویژگی‌های گمشده در مجموعه داده دیابت و سرطان سینه با استفاده از مدل جایگذاری شده‌اند، با حالتی که داده‌های گمشده حذف شده‌اند و روش متداول جایگذاری با مقدار میانگین مقایسه شد تا مشخص گردد کدام روش در مواجهه با مقادیر گمشده در مجموعه داده دیابت و سرطان سینه بهتر است (حذف نمونه دارای مقدار گمشده، جایگذاری مقدار گمشده توسط مدل شبکه پرسپترون دو لایه و یا جایگذاری توسط مقدار میانگین). همچنین، به منظور بررسی وجود تفاوت معنی‌دار بین روش‌های کلاسیک و مدل‌های مبتنی بر شبکه عصبی، از آزمون t استفاده شد. پیاده‌سازی مدل در نرم‌افزار Matlab نسخه R2018b و آزمون t در نرم‌افزار Excel صورت گرفت. به منظور پیاده‌سازی روش استفاده شده بر روی داده‌های غیر عددی، لازم است ابتدا داده غیر عددی به عدد تبدیل شود و سپس توسط مدل جایگذاری گردد. مطالعه حاضر از نوع توصیفی و داده‌ها تنها در راستای هدف پژوهشی و پاسخ به سؤال مطالعه استفاده شد.

یافته‌ها

نتایج حاصل از جایگذاری مقادیر گمشده توسط مدل با مقدار واقعی مقایسه شد و بین آن‌ها MSE محاسبه گردید. در جدول ۲ مقادیر میانگین MSEها، انحراف معیار MSEها و خطای کل در حالتی که مقادیر گمشده توسط مدل جایگذاری شده است را در ۲۰ بار اجرا برای مجموعه داده دیابت و سرطان سینه نشان می‌دهد.

در روش جایگذاری مقادیر گمشده با مقادیر میانگین در مجموعه داده‌های دیابت و سرطان سینه، نتایج حاصل از جایگذاری با مقدار واقعی متغیرها مقایسه و بین آن‌ها MSE محاسبه گردید. در جدول ۳ مقادیر میانگین MSEها، انحراف معیار MSEها و خطای کل در ۲۰ بار اجرا برای مجموعه داده‌های دیابت و سرطان سینه ارایه شده است.

جدول ۲: نتایج با ۲۰ بار اجرا، جایگذاری مقادیر مفقود شده توسط مدل برای داده‌های دیابت و سرطان سینه

مجموعه	نام ویژگی	انحراف معیار MSEها	میانگین MSEها	خطای کل (میانگین ± انحراف معیار)
داده‌های دیابت	تست تحمل گلوکز	۰/۰۱۰۹	۰/۰۲۲۹	۰/۰۲۲۹ ± ۰/۰۱۰۹
	عضلات سه سر	۰/۰۰۳۹	۰/۰۲۳۳	۰/۰۲۳۳ ± ۰/۰۰۳۹
	انسولین	۰/۰۰۵۸	۰/۰۱۲۳	۰/۰۱۲۳ ± ۰/۰۰۵۸
	شاخص جرم	۰/۰۰۴۳	۰/۰۱۰۵	۰/۰۱۰۵ ± ۰/۰۰۴۳
داده‌های سرطان سینه	یکنواختی اندازه سلول	۰/۰۱۲۱	۰/۰۲۳۹	۰/۰۲۳۹ ± ۰/۰۱۲۱
	یکنواختی شکل سلول	۰/۰۰۷۹	۰/۰۲۴۵	۰/۰۲۴۵ ± ۰/۰۰۷۹
	اندازه سلول مخاطی تک	۰/۰۱۳۱	۰/۰۱۸۹	۰/۰۱۸۹ ± ۰/۰۱۳۱
	کروماتین مطلوب	۰/۰۰۷۸	۰/۰۲۰۲	۰/۰۲۰۲ ± ۰/۰۰۷۸

MSE: Mean Squared Error

جدول ۳: نتایج با ۲۰ بار اجرا، جایگذاری مقادیر مفقود شده با مقدار میانگین برای داده‌های دیابت و سرطان سینه

مجموعه	نام ویژگی	انحراف معیار MSEها	میانگین MSEها	خطای کل (میانگین ± انحراف معیار)
داده‌های دیابت	تست تحمل گلوکز	۰/۰۰۷۲	۰/۰۴۰۵	۰/۰۴۰۵ ± ۰/۰۰۷۲
	عضلات سه سر	۰/۰۰۹۱	۰/۰۴۱۴	۰/۰۴۱۴ ± ۰/۰۰۹۱
	انسولین	۰/۰۱۰۳	۰/۰۴۲۸	۰/۰۴۲۸ ± ۰/۰۱۰۳
	شاخص جرم	۰/۰۰۵۹	۰/۰۲۶۹	۰/۰۲۶۹ ± ۰/۰۰۵۹۰
داده‌های سرطان سینه	یکنواختی اندازه سلول	۰/۰۱۲۸	۰/۰۷۳۵	۰/۰۷۳۵ ± ۰/۰۱۲۸
	یکنواختی شکل سلول	۰/۰۲۴۸	۰/۰۸۳۸	۰/۰۸۳۸ ± ۰/۰۲۴۸
	اندازه سلول مخاطی تک	۰/۰۱۶۶	۰/۰۶۳۳	۰/۰۶۳۳ ± ۰/۰۱۶۶
	کروماتین مطلوب	۰/۰۱۸۳	۰/۰۶۷۶	۰/۰۶۷۶ ± ۰/۰۱۸۳

MSE: Mean Squared Error

صحت و حساسیت طبقه‌بندی (سالم یا بیمار) در حالتی که داده‌های گمشده از مجموعه داده‌های دیابت و سرطان سینه حذف شده بود، با استفاده از مقدار میانگین و مدل شبکه عصبی پرسپترون دو لایه جایگذاری گردید و در پنج اجرا به دست آمد. نتایج و میانگین آن‌ها در جدول ۴ گزارش شده است. صحت و حساسیت طبقه‌بندی بیماری در حالتی که مقادیر گمشده توسط شبکه عصبی پرسپترون جایگذاری شده بود، در مقایسه با دو روش مرسوم مقدار میانگین و روش حذف مقادیر گمشده در مجموعه داده دیابت به ترتیب حدود ۲، ۴، ۲ و ۴ درصد و در مجموعه داده سرطان سینه به ترتیب حدود ۱، ۳، ۲، ۵ درصد بیشتر شد.

محاسبه مقدار P بین MSEهای به دست آمده از دو روش جایگذاری با مقدار میانگین و جایگذاری توسط مدل بر روی دو مجموع داده دیابت و سرطان سینه، نشان دهنده وجود تفاوت معنی‌دار ($P < ۰/۰۰۱$) میان مقادیر جایگذاری شده توسط این دو روش بود. در جدول ۵ مقادیر Pهای به دست آمده بر روی چهار متغیر تست تحمل گلوکز، عضلات سه سر و شاخص جرم در مجموعه داده دیابت و چهار متغیر یکنواختی اندازه سلول، یکنواختی شکل سلول، اندازه سلول مخاطی تک و کروماتین مطلوب در مجموعه داده سرطان سینه ارایه شده است.

بحث

در پژوهش حاضر، با به کار گرفتن روش‌های داده‌کاوی بر روی مجموعه داده‌های دیابت و سرطان سینه، مدل شبکه پرسپترون دو لایه به منظور پیش‌بینی مقادیر گمشده طراحی گردید. این مدل کمک می‌کند که مقادیر از دست رفته در مجموعه داده‌های پزشکی جایگذاری شود تا روند تشخیص و پیش‌بینی بیماری به درستی انجام گیرد. کارایی مدل پیشنهاد شده در مطالعه حاضر، از لحاظ معیارهای ارزیابی در مقایسه با نتایج مطالعه Silva-Ramirez و همکاران (۹)، دارای مقادیر پایین MSE در محدوده ۰/۰۱ تا ۰/۰۴ و بالاترین میزان دقت بود. پایین بودن میزان MSEها در تحقیق حاضر نسبت به نتایج پژوهش Silva-Ramirez و همکاران که در محدوده ۰/۰۴ تا ۰/۱ است (۹)، نشان دهنده نزدیک بودن مقادیر جایگذاری شده توسط مدل به مقادیر واقعی و میزان اعتماد به مدل ارایه شده در مطالعه می‌باشد. همچنین، مدل ارایه شده توسط Deb و Liew از نظر صحت، مقدار کمتر و از نظر میزان خطا نسبت به مطالعه حاضر، مقدار بیشتری (۰/۰۹ تا ۰/۱) را نشان داد (۸). این مدل از نظر معیار صحت در مقایسه با الگوریتم به کار رفته در تحقیق Purwar و Singh مقدار بالاتری را نشان داد (۴). میزان خطای نزدیک به صفر در پژوهش حاضر، حکایت از عملکرد خوب شبکه پرسپترون دو لایه در جایگذاری مقادیر گمشده در مجموعه داده‌های پزشکی دارد. بنابراین، مدل پیشنهاد شده می‌تواند در مقداردهی مقادیر گمشده در مجموعه داده‌های آزمایشگاهی و پزشکی به منظور پیش‌بینی و تشخیص درست بیماری مورد استفاده قرار گیرد. مطالعه حاضر محدودیت‌هایی داشت که از آن جمله می‌توان به تعداد محدود ویژگی‌ها و محدودیت جغرافیایی به منظور جمع‌آوری داده‌ها اشاره نمود که می‌تواند در تحقیقات آینده مورد توجه قرار گیرد.

جدول ۴: نتایج حاصل از جایگذاری مقادیر گمشده در ۵ بار اجرا (به درصد) برای داده‌های دیابت و سرطان سینه

مجموعه		حذف مفقود شده‌ها		جایگذاری با مقدار میانگین		جایگذاری توسط مدل	
		صحت	حساسیت	صحت	حساسیت	صحت	حساسیت
داده‌های دیابت	اجرای اول	۷۶/۳	۵۵/۰	۷۵/۴	۵۳/۰	۷۷/۵	۵۷/۰
	اجرای دوم	۷۶/۹	۵۵/۰	۷۵/۱	۵۲/۰	۷۷/۱	۵۶/۰
	اجرای سوم	۷۴/۰	۵۱/۰	۷۵/۳	۵۶/۰	۷۷/۹	۵۸/۰
	اجرای چهارم	۷۲/۴	۵۳/۰	۷۶/۰	۵۳/۰	۷۷/۲	۵۷/۰
	اجرای پنجم	۷۶/۷	۵۵/۰	۷۵/۹	۵۴/۰	۷۸/۵	۵۸/۰
	میانگین	۷۵/۳	۵۳/۸	۷۵/۵	۵۳/۶	۷۷/۶	۵۷/۲
داده‌های سرطان سینه	اجرای اول	۹۴/۹	۹۱/۰	۹۵/۹	۹۳/۰	۹۶/۷	۹۶/۰
	اجرای دوم	۹۴/۵	۹۱/۰	۹۵/۴	۹۲/۰	۹۶/۷	۹۶/۰
	اجرای سوم	۹۵/۲	۹۲/۰	۹۴/۸	۹۲/۰	۹۶/۹	۹۷/۰
	اجرای چهارم	۹۴/۰	۹۱/۰	۹۶/۴	۹۵/۰	۹۶/۹	۹۷/۰
	اجرای پنجم	۹۵/۰	۹۲/۰	۹۵/۹	۹۴/۰	۹۶/۷	۹۷/۰
	میانگین	۹۴/۷	۹۱/۴	۹۵/۷	۹۳/۲	۹۶/۷	۹۶/۶

جدول ۵: نتایج حاصل از آزمون t در دو مجموعه داده‌های دیابت و سرطان سینه

مجموعه داده دیابت		مجموعه داده سرطان سینه	
متغیر	مقدار P	متغیر	مقدار P
تست تحمل گلوکز	< ۰/۰۰۱	یکنواختی اندازه سلول	< ۰/۰۰۱
عضلات سه سر	< ۰/۰۰۱	یکنواختی شکل سلول	< ۰/۰۰۱
انسولین	< ۰/۰۰۱	اندازه سلول مخاطی تک	< ۰/۰۰۱
شاخص جرم	< ۰/۰۰۱	کروماتین مطلوب	< ۰/۰۰۱

نتیجه‌گیری

مقادیر جایگذاری شده توسط الگوریتم ارایه شده، می‌تواند جایگزین مناسبی برای مقادیر گمشده باشد. عملکرد روش پیشنهادی در پژوهش حاضر، باعث بهبود صحت و حساسیت طبقه‌بندی در تشخیص بیماری دیابت و سرطان سینه نسبت به روش حذف مقادیر گمشده و جایگذاری میانگین شده است. در مطالعه حاضر، برتری روش‌های جایگذاری مقادیر گمشده مبتنی بر ماشین‌های یاد گیرنده نسبت به روش‌های کلاسیک، مشهود می‌باشد.

پیشنهادها

به منظور جایگذاری مقادیر گمشده، می‌توان از انواع روش‌های یادگیری و طبقه‌بندی‌ها و همچنین، داده‌های سایر بیماری‌ها استفاده نمود.

تشکر و قدردانی

بدین وسیله از جناب آقای دکتر سید مسعود امینی، استاد دانشکده علوم ریاضی دانشگاه تربیت مدرس تشکر و قدردانی به عمل می‌آید.

تضاد منافع

در انجام پژوهش حاضر، نویسندگان هیچ‌گونه تضاد منافعی نداشته‌اند.

References

1. Ispirova G, Eftimov T, Seljak BK. Evaluating missing value imputation methods for food composition databases. Food Chem Toxicol 2020; 141: 111368.
2. Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E, Lix LM. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health Qual Life Outcomes 2019; 17(1): 106.
3. Folguera L, Zupan J, Cicerone D, Magallanes JF. Self-organizing maps for imputation of missing data in incomplete data matrices. Chemometr Intell Lab Syst 2015; 143: 146-51.
4. Purwar A, Singh SK. Hybrid prediction model with missing value imputation for medical data. Expert Syst Appl 2015; 42(13): 5621-31.
5. de Silva H, Perera AS. Missing data imputation using Evolutionary k- Nearest neighbor algorithm for gene expression data. Proceedings of the 16th International Conference on Advances in ICT for Emerging Regions (ICTer); 2016 Sep 1-3; Negombo, Sri Lanka.
6. Jea K, Hsu C, Tang L. A missing data imputation method with distance function. Proceddings of the 2018 International Conference on Machine Learning and Cybernetics (ICMLC); 2018 Jul 15-18; Chengdu, China.
7. Duan Y, Lv Y, Liu YL, Wang FY. An efficient realization of deep learning for traffic data imputation. Transp Res Part C Emerg Technol 2016; 72: 168-81.
8. Deb R, Liew AW-C. Missing value imputation for the analysis of incomplete traffic accident data. Inf Sci 2016; 339: 274-89.
9. Silva-Ramirez EL, Pino-Mejias R, Lopez-Coello M, Cubiles-de-la-Vega MM-D. Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Netw 2011; 24(1): 121-9.
10. de Goeij MC, van DM, Jager KJ, Tripepi G, Zoccali C, Dekker FW. Multiple imputation: Dealing with missing data. Nephrol Dial Transplant 2013; 28(10): 2415-20.

Health Information Management

Published by ***Vesnu Publications***

Vol 18/ No 1/ April & May 2021

DOI: 10.22122/him.v18i1.4194

Imputing of Missing Values in Diabetes and Breast Cancer Datasets through a Two-Layer Perceptron Neural Network

Elham Pourjani[1], Sara Najafzadeh[2], Nader Jafarnia-Dabanloo[3]

Original Article

Abstract

Introduction: Imputation of missing values in a medical data set is one of the important challenges in data mining. Therefore, this study was performed with the aim of imputation the missing values of some features of the diabetes and breast cancer datasets.

Methods: In this descriptive study, a breast cancer dataset consisting of 699 specimens including 458 benign and 241 malignant specimens, along with a diabetes dataset consisting of 768 specimens including 500 non-diabetic specimens and 268 other specimens with diabetes, were used. For the purpose of the imputation of missing values in these two datasets, a model based on a two-layer perceptron neural network was developed, and for the purpose of assessment, support vector machine (SVM) and t test were used.

Results: The mean squared errors (MSEs) obtained in the two-layer perceptron neural network model, in the diabetes dataset about 0.03 and in the breast cancer dataset about 0.04, were less than the MSEs obtained in the imputation method with the mean value. The values imputed by the model were closer to the actual value than the values imputed with the mean value. Accuracy and sensitivity of disease classification in the case of missing values imputed by the perceptron neural network increased in comparison with the two conventional methods of mean value and the method of deleting missing values, about 2, 4, 2, and 4 percent in the diabetes dataset, and about 1, 3, 2, 5 percent in the dataset breast cancer, respectively. There was a significant difference between the two methods of imputation of missing values with the mean value and imputation by the model.

Conclusion: The imputation of the missing values in the medical data set by the two-layer perceptron neural network showed better results in the classification of the disease than the two methods of imputation with the mean value and the method of deleting missing values.

Keywords: Data Mining; Neural Network Models; Support Vector Machine

Received: 20 Aug., 2020 Accepted: 31 Mar., 2021 Published: 04 Apr., 2021

Citation: Pourjani E, Najafzadeh S, Jafarnia-Dabanloo N. **Imputing of Missing Values in Diabetes and Breast Cancer Datasets through a Two-Layer Perceptron Neural Network.** Health Inf Manage 2021; 18(1): 1-6.

Article resulted from MSc thesis funded by Islamic Azad University, Science and Research Branch.

1- MSc Student, Artificial Intelligence and Robotic, Department of Computer Engineering, School of Mechanics, Electrical and Computer, Science and Research Branch, Islamic Azad University, Tehran, Iran

2- Assistant Professor, Network, Department of Computer, School of Electrical Engineering, Yadegar-e-Imam Branch, Islamic Azad University, Shahr-e-Rey, Iran

3- Associate Professor, Electronic, Department of Electrical Engineering, School of Science and Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

Address for correspondence: Elham Pourjani; MSc Student, Artificial Intelligence and Robotic, Department of Computer Engineering, School of Mechanics, Electrical and Computer, Science and Research Branch, Islamic Azad University, Tehran, Iran; Email: e.pourjani@gmail.com