دوره شانزدهم، شماره بهار و تابستان ۱۴۰۲

مجله فناوری اطلاعات در طراحی مهندسی

Information Technology in Engineering Design

http://ited.sinaweb.net

جبران داده‌های مفقود پزشکی با ترکیب شبکه بیزین و ماشین یادگیری مفرط

الهه صباحی [۱] سیدمحمدحسین معطر*[۲] رضا شیبانی [۳]

(۱) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

(۲) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران*

(۳) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

(تاریخ دریافت :۱۴۰۲/۰۴/۳۱ تاریخ پذیرش: ۱۴۰۲/۰۷/۰۵)

چکیده

شبکه بیزین یکی از توانمندترین روش‌ها در تخمین داده‌های مفقود است. از طرفی ماشین یادگیری مفرط به‌طور تحلیلی وزن‌های خروجی‌های بهینه را محاسبه می‌کند و این امیدواری وجود دارد که در مورد داده‌های مفقود، به‌عنوان یک مدل خوب عمل کند. مهمترین چالش بسیاری از روش‌های تخمین مقادیر مفقود این است که ماهیت این روش‌ها عمدتا برای داده‌ها با مقادیر پیوسته یا داده‌ها با مقادیر گسسته مناسب است. شبکه بیزین و ماشین یادگیری مفرط نیز از این قاعده مستثنا نیستند و به ترتیب برای پر کردن مقادیر مفقود گسسته و مقادیر مفقود پیوسته مناسب‌ترند. لذا در این پژوهش از ترکیب این دو مدل جهت تخمین داده‌های مفقود مخلوط در مجموعه داده هپاتیت استفاده شده و در نهایت دسته بندی بر اساس شبکه بیزین جهت تخمین کلاس خروجی انجام شده است. روش پیشنهادی بر اساس معیارهای دقت، فراخوانی، صحت و ریشه میانگین مربعات خطا مورد ارزیابی قرار گرفت. نتایج نشان داد، جبران داده‌های مفقود با ترکیب *BN-ELM* و طبقه‌بندی بر اساس شبکه بیزین صحت بالایی کسب کرده است. همچنین روش پیشنهادی با سایر روش‌های تخمین داده، بر اساس طبقه بندهای *BN*، *ELM* و *KNN* مورد مقایسه قرار گرفت و نتایج برتری روش پیشنهادی را نشان می‌دهد.

کلمات کلیدی: تخمین مقادیر مفقود، تشخیص بیماری هپاتیت، شبکه بیزین، ماشین یادگیری مفرط

*عهده‌دار مکاتبات:

سیدمحمدحسین معطر

نشانی: گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

پست الکترونیکی:**moattar@mshdiau.ac.ir**

۱- مقدمه

پژوهش در زمینه داده‌کاوی و پیش‌بینی برای کاربردهای پزشکی یک موضوع در حال رشد است. به‌طورکلی، یک پزشک دانش خود را از علائم بیمار جمع‌آوری می‌کند و درنهایت بیماری را تشخیص می‌دهد؛ تشخیص معمولاً یا با ارزیابی نتایج آزمایش‌ها فعلی از بیماران و یا با مراجعه به تصمیم‌گیری‌های قبلی در مورد سایر بیماران با نتایج آزمایش‌ها مشابه، گرفته می‌شود. با توجه به افزایش تعداد بیماران و از طرف دیگر، توسعه روش‌های محاسباتی جدید و ابزار، باعث می‌شود تصمیم‌گیری از پایگاه داده اختصاص داده شده از سوابق الکترونیکی بیمار[۱] آسان‌تر گردد. داده‌های بالینی اغلب برای شناسایی ویژگی‌های که می‌تواند در طبقه‌بندی بیماران کمک کند، استفاده می‌شود. شناسایی کلاس‌های بیماران با دوره‌های مختلف بالینی یا پاسخ به درمان خاص اجازه می‌دهد تا روش مناسب برای مدیریت هر بیمار [۲] و [۳] طراحی شود. مجموعه داده‌های بالینی اغلب دارای حجم نمونه نسبتاً کوچک و حجم بسیاری از اطلاعات ازدست‌رفته است. در این شرایط، یک روش مدیریت داده‌های گمشده، حذف داده‌های ناموجود و تجزیه و تحلیل بر روی داده‌های باقیمانده است. در مقابل، اگر مکانیزمی باشد که اجازه دهد تا داده‌های ناقص را کامل کنیم، از کاهش حجم نمونه اجتناب می‌شود، و علاوه بر این کامل کردن داده‌های مفقود ممکن است نتایج تشخیص و پیش‌بینی بهبود یابد[۴].

یکی از موضوعاتی که غالباً در بسیاری از مسائل مختلف آماری مورد توجه قرار گرفته، بحث داده‌های مفقود است؛ وجود داده‌های مفقود در اغلب بررسی‌ها یا مطالعات اقتصادی، اجتماعی، پزشکی و غیره امری انکارناپذیر است [۵] و [۶]. داده‌های مفقود، یکی از مشکلاتی است که اغلب محققان و تحلیلگران در هنگام کار با مجموعه داده‌ها، با آن روبرو هستند. در مواردی که داده‌های مفقود در مسائل آماری تأثیر قابل‌توجه‌ای در نتایج داشته باشند، مواجه‌شدن با آن‌ها حساسیت بیشتری را ایجاد می‌کند و باعث می‌شود موضوع گمشدگی داده‌ها، همواره به‌عنوان یکی از مهم‌ترین مباحث علم آمار موردتوجه قرار گیرد و پر کردن مقادیر مفقود اهمیت بسزایی در داده‌کاوی دارند. داده‌های مفقود معمولاً اطلاعات مهمی را نمایان می‌کنند که در صورت عدم در دسترس بودن، تحلیل و تفسیر دقیق داده‌ها دچار اشتباه می‌شود. موجودیت داده‌های مفقود می‌تواند به اندازه نمونه اثر بگذارد. اگر تعداد زیادی داده مفقود داشته باشیم، این ممکن است به کاهش دقت تحلیل‌های ما منجر شود. داده‌های مفقود ممکن است تصمیم‌گیری‌ها و تحلیل‌های ما را تحت تأثیر قرار دهند. این تأثیر ممکن است به خصوص در مواردی که داده‌های مفقود به صورت نامتعادل در متغیرها یا زمان‌ها واقع شده باشند، مهم باشد.

ادبیات روش‌های پر کردن مقادیر مفقود طیف متنوعی از روشها را شامل می‌شوند. در روشهای پر کردن متغیرهای مفقود با میانگین یا مد، مقدار مفقود را با میانگین یا مد متغیر مربوطه جایگزین می‌کنند. این روش ساده است اما ممکن است اطلاعات مهمی را از دست بدهد. در روش‌های احتمالاتی مثل رگرسیون و بیزین، مقادیر مفقود را با توجه به ارتباط با متغیرهای دیگر پیش‌بینی می‌کنند. در روش‌های مبتنی بر یادگیری ماشین شبکه‌های عصبی از یادگیری بر روی نمونه یا روش‌های خودآموزی برای

پر کردن داده‌های مفقود استفاده می‌شود. انتخاب روش پر کردن داده‌های مفقود بر اساس ویژگی‌های داده‌ها، میزان داده‌های مفقود، و هدف تحلیل انجام می‌شود. همچنین، توجه به معیارهای ارزیابی روش‌های پر کردن داده‌های مفقود نیز حائز اهمیت است.

یکی از مهمترین تفاوت‌های روش‌های پر کردن مقادیر مفقود تفاوت در نوع متغیرها اعم از گسسته یا پیوسته است. متغیرهای پیوسته ممکن است مقادیری در بازه‌های پیوسته و با تعداد نامعینی داشته باشند، در حالی که متغیرهای گسسته مقادیری دارند که از یک مجموعه محدود و معین انتخاب می‌شوند. در پر کردن مقادیر مفقود داده‌های پیوسته و گسسته، تفاوت‌هایی وجود دارد. در داده‌های پیوسته، معمولاً از روش‌های مدل‌سازی احتمالی مانند رگرسیون خطی یا غیرخطی برای پیش‌بینی مقادیر مفقود استفاده می‌شود. در این روشها می‌توان از توزیع‌های احتمالی مثل توزیع نرمال برای مدل‌سازی داده‌های پیوسته استفاده کرد. همچنین روش‌های یادگیری ماشین از قبیل روش‌های خودآموزی نیز می‌توانند برای پیش‌بینی داده‌های پیوسته مورد استفاده قرار بگیرند. در مقابل، در داده‌های گسسته، ممکن است از روش‌های مخصوص به داده‌های گسسته مثل روش‌های کلاس بندی یا احتمالات چندجمله‌ای برای پر کردن مقادیر مفقود استفاده شود. همچنین تخمین توزیع احتمالات گسسته (مثل توزیع احتمالی چندجمله‌ای) ممکن است برای مقادیر مفقود مورد استفاده قرار گیرد. به عنوان مثال، در پر کردن مقادیر مفقود داده‌های پیوسته، ممکن است از تقریب‌های خطی استفاده شود، در حالی که در داده‌های گسسته ممکن است از تخمین احتمالات توزیع داده‌ها استفاده شود.

شبکه بیزین[۱] (BN) و ماشین یادگیری مفرط[۲] (ELM) هر دو از روش‌های قدرتمند در پر کردن داده‌های مفقود هستند. ماشین یادگیری مفرط به طور تحلیلی وزن‌های بهینه خروجی را محاسبه می‌کند و امیدواری وجود دارد که در مواردی که تعداد زیادی داده مفقود وجود دارد و انتساب دقیق داده‌ها ضروری است، به عنوان یک مدل مؤثر عمل کند. یکی از چالش‌های اساسی در روش‌های تخمین داده‌های مفقود، تطبیق مدل‌ها به نوع داده‌هاست، به این معنا که روش‌ها عمدتاً برای داده‌های پیوسته یا داده‌های گسسته مناسب هستند. شبکه بیزین و ماشین یادگیری مفرط نیز از این نظر استثناء نیستند. در این پژوهش، ما از ترکیب این دو مدل برای تخمین داده‌های مفقود مخلوط[۳] در مجموعه داده هپاتیت استفاده نموده‌ایم. منظور از داده‌های مخلوط، داده‌هایی است که هر دو جنس متغیر گسسته و پیوسته را شامل می‌شود. عملکرد روش پیشنهادی براساس معیارهای دقت، فراخوانی، صحت و میانگین مربعات خطا مورد ارزیابی قرار گرفته‌است و نتایج نشان از دقت بالای روش ترکیبی پیشنهادی در پر کردن داده‌های مفقود دارد. ساختار این مقاله به این صورت است: پس از این بخش و در بخش دوم به بررسی کارهای مرتبط پرداخته شده است. در بخش سوم روش پیشنهادی بیان شده است. در فصل چهارم نتایج ارزیابی‌ها و درنهایت در بخش پنجم نتیجه‌گیری کلی از این مقاله بیان گردیده است.

۲- کارهای پیشین

[۱] Bayesian Network

[۲] Extreme Learning Machine

[۳] Mixed

در سال‌های اخیر سیستم‌های تشخیص پزشکی کمک شایانی به پزشکان کرده و داده‌کاوی یکی از علوم جدید و تأثیرگذار می‌باشد که به کمک علم پزشکی آمده است. در این راستا تحقیقات گسترده‌ای انجام شده است ازجمله چن و همکارانش [۷] از یک مجموعه سخت و طبقه بند ماشین بردار پشتیبانی برای تشخیص سرطان پستان استفاده کردند. مطالعه آن‌ها باهدف ایجاد یک سیستم تشخیصی خودکار برای تشخیص تومور خوش خیم پستان از بدخیم می‌باشد و در مقایسه با سایر روش‌های طبقه‌بندی شامل جنگل تصادفی، شبکه‌های عصبی، عملکرد و تعمیم بهتری داشته است. ژنگ و همکارانش [۸] ترکیبی از الگوریتم K-means و ماشین بردار پشتیبانی را توسعه دادند. الگوریتم K-means برای به رسمیت شناختن الگوهای پنهان از تومورهای بدخیم و خوش خیم مجموعه داده‌های ویسکانسین مورد استفاده قرار گرفت، سپس از SVM[۱] برای به دست آوردن یک طبقه بند جدید برای تشخیص تومورها استفاده شد. هدف از این پژوهش، تشخیص سرطان پستان بر اساس استخراج ویژگی‌های تومور بود. در پژوهش آن‌ها K-SVM عملکرد بهتری نسبت SVM داشت. مگلوگیانیس و همکارانش [۹]، یک سیستم تشخیص خودکار برای شناسایی بیماری‌های دریچه قلب مبتنی بر طبقه بند ماشین بردار پشتیبان بر اساس صداهای قلب پیشنهاد کردند. این سیستم کار تشخیص را که حتی برای پزشک با تجربه بسیار دشوار است، انجام می‌دهد. این مطالعه از یک مجموعه داده جهانی که از ۱۹۸ سیگنال صدا قلب به کار گرفته شده، استفاده می‌کند. در ابتدا صداهای قلب با استفاده از SVM به دو دسته عادی و یا بیماری دسته‌بندی می‌شود، سپس موارد بیمار به عنوان سیستولیک یا دیاستولیک طبقه‌بندی شده‌اند. روش پیشنهاد شده با روش‌های شبکه‌های عصبی پس انتشار و K نزدیک‌ترین همسایه مقایسه شده و نتایج نشان‌دهنده برتری روش پیشنهادشده، بوده است. با توجه به مطالعات انجام شده که در زمینه تشخیص پزشکی وجود دارد، دریافتیم یکی از مشکلاتی که در طراحی سیستم‌های پزشکی وجود دارد، داده‌های مفقود است، این داده‌ها باید به‌صورت مناسب تخمین زده شوند تا تشخیص دقیق‌تر باشد. در دهه‌های گذشته، محققین روش‌های مختلفی را برای تخمین این داده‌ها ارائه کرده‌اند [۱۰] تا [۱۲]، اما روش‌های قدیمی‌تر متکی به تصحیح مجموعه داده‌ها با صرف‌نظر کردن از موردهای دارای مقادیر گم‌شده و یا جایگزینی مقادیر تخمینی با مقادیر گم‌شده بودند. در ادامه این بخش به بررسی مقالات مرتبط در زمینه تخمین داده پرداخته شده است.

۲-۱- روش‌های احتمالاتی برای تخمین مقادیر مفقود

رانکویتا و همکارانش [۴] از شبکه بیزین برای پیدا کردن وابستگی میان متغیرهای آزمایشگاهی و درنتیجه پر کردن مقادیر گمشده استفاده کردند. در این روش پس از حدس مقادیر گمشده، به درخت بقا اجازه داده شد تا بر روی دیتاست کامل اعمال شود. در این روش شبکه‌ی بیزین به‌طور مستقیم از یک داده ناقص با استفاده از الگوریتم EM[۲] آموزش داده شد، برای این منظور تنها برای حدس مقادیر گمشده بدون هیچ دانشی از درخت بقا استفاده شده است و پس‌ازآن درخت بقا بر روی یک دیتاست کامل اعمال گردیده است.

[۱] Support Vector Machine

[۲] Expectation Maximization

سـویلج و همکارانش [۱۳] مسـئله عمومی رگرسیون تحت سناریو داده‌های گم‌شده را مورد بررسی قرار دادند. به این منظور تخمین‌های معتبر برای تابع رگرسیون (تقریبی) ارائه‌شده و یک روش جدید بر اساس مدل مخلوط گوسی و ماشین یادگیری مفرط توسـعه یافته اسـت. مدل مخلوط گوسـی برای مدل توزیع داده‌ها اسـتفاده می‌شود که مسئولیت رسیدگی به مقادیر ازدست‌رفته را دارد، درحالی‌که ماشـین یادگیری مفرط قادر به طراحی یک اسـتراتژی انتساب چندگانه برای برآورد نهایی است. مراحل کلی این روش به این صـورت اسـت: ۱) برازش مدل مخلوط گوسـی در یک مجموعه داده‌ها با مقادیر ازدسـت‌رفته. ۲) تولید مجموعه داده‌های جدید از طریق انتسـاب چندگانه بر اساس مدل مخلوط گوسی از مرحله اول. ۳) ساخت ماشین یادگیری مفرط برای هر مجموعه داده تولید شده در مرحله اول. ۴) ترکیب تمام ماشین‌های یادگیری مفرط برای تخمین نهایی.

سهگال و همکارانش [۱۴] یک الگوریتم جدید برای کامل کردن داده‌های ازدست‌رفته به نام CMVE[۱] ارائه کردند؛ این روش از چند ماتریس مبتنی بر کوواریانس برای پیش‌بینی نهایی ارزش ازدست‌رفته استفاده کرده است. ماتریس محاسبه‌شده با استفاده از رگرسـیون حداقل مربعات و روش برنامه‌ریزی خطی بهینه‌سـازی شـده اسـت. الگوریتم CMVE با روش‌های تخمینی موجود ازجمله تجزیه‌وتحلیل مؤلفه‌های اصـلی بیزی، حداقل نسـبت دادن مربع و K نزدیک‌ترین همسـایه مقایسه شده است. نتایج نشان داده CMVE به‌طورقطع قابلیت برآورد برتر و قوی از داده‌های دست رفته را در مقایسه با روش‌های دیگر دارد.

فولگورا و همکارانش [۱۷] از شبکه عصبی خودسازمانده برای حل چالش داده‌های مفقود استفاده کردند. نقشه خودسازمانده[۲] (SOM) روشـی چند متغیره قوی برای تجزیه‌وتحلیل داده‌ها می‌باشـد و قادر به مدل‌سازی خطی در یک سیستم است و درنتیجه برآورد خوبی از ارزش داده‌های ازدسـت‌رفته در مجموعه داده‌های زیست‌محیطی ارائه می‌دهد و در مقایسه با روش‌های دیگر داده محور مانند پرسـپترون چندلایه شـبکه‌های عصـبی مصنوعی و روش احتمال بر اساس تجزیه‌وتحلیل رگرسیون، ثابت‌شده بهترین عملکرد را برای نسـبت دادن به ارزش‌های ازدسـت‌رفته ازنظر دو ضـریب همبسـتگی بین مشـاهده و پیش‌بینی ارزش و همچنین میانگین مربع خطا دارد. مزیت دیگر SOM این است که می‌توان همان نقشه را برای پیش‌بینی هر مقدار ازدست‌رفته دیگر در هر متغیر استفاده کرد، ولی مدل‌های رگرسیون و شبکه‌های عصبی چندلایه پرسپترون برای هر متغیر که شامل ارزش ازدست‌رفته باشد، توسعه داده می‌شود.

مقاله [۱۸] به مسئله انتقال داده‌های ناقص در ماتریس‌های داده‌ای می‌پردازد. روشی که در این مقاله ارائه شده است از ترکیب روش‌های رگرسـیون با تقریب‌های پایین‌رتبه استفاده می‌کند. برای بهبود تکمیل داده‌ها، یک تعمیم پیشنهاد شده است که در آن از تجزیه مقادیر ویژه اسـتفاده می‌شود و پارامتر تنظیمی این روش توسط اعتبارسنجی متقاطع تخمین زده می‌شود. در مقاله [۱۹]، دو شـبکه عصـبی عمیق نظارت شده، یعنی چندلایه پرسپترون (MLP) و شبکه‌های باور عمیق (DBN) برای جبران مقادیر گم‌شده مقایسـه می‌شوند. نتایج نشان می‌دهد که MLP و DBN به طریق قابل توجهی بهتر از روش‌های جبران مقدار مبتنی بر میانگین، KNN، CART و SVM عمل می‌کنند، و DBN بهترین عملکرد را ارائه می‌دهد.

[۱] Collateral missing value imputation

[۲] Self Organizing Map

در [۲۰]، برای پر کردن مقادیر مفقود، یک الگوریتم جدید جبران داده‌های سری زمانی ارائه شده است. در [۲۱]، تکنیک‌های افزایش داده برای بازسازی دنباله‌های زمانی مورد بررسی قرار گرفته است. در این روش ابتدا، یک خودرمزگذار تطبیقی به عنوان مدل اصلی جبران انتخاب شده است. سپس نرخ بهینه افزایش داده انتخاب می‌شود. همچنین [۲۲] یک روش نوین بر اساس گراف برای جبران مقادیر گم‌شده ارائه داده است، که اطلاعات ویژگی و ارتباطات بین نقاط داده را در نظر می‌گیرد تا عملکرد مدل را بهبود بخشد. این روش مقادیر گم‌شده را با استفاده از یک شبکه‌ی کانولوشنی گرافی[۱] (GCN) جبران می‌کند.

استفاده از روش‌های احتمالاتی برای تخمین مقادیر مفقود دارای مزایا و معایبی است. از جمله مزایای این روشها این است که روش‌های احتمالاتی بر مبنای اصول احتمالاتی و تئوری احتمالات ساخته شده‌اند. این به این معناست که آن‌ها می‌توانند توزیع احتمالاتی دقیقی برای مقادیر مفقود مدل کنند. این مدل‌ها به تحلیل دقیق‌تر داده‌های مفقود کمک می‌کنند. همچنین این روش‌ها معمولاً از اطلاعات موجود در داده‌ها به عنوان پارامترها یا توزیع‌های احتمالی برای مقادیر مفقود استفاده می‌کنند. این به این معناست که از اطلاعات موجود بهره‌برداری می‌شود و تخمین‌های نزدیکتری به واقعیت ارائه می‌دهد. ولی در مقابل، این روشها به مدل‌سازی دقیقی از توزیع داده‌ها نیاز دارند. همچنین این روش‌ها محاسبات پیچیده‌تری نسبت به روش‌های دیگر می‌طلبند و به فرضیات اولیه از مدل احتمالاتی حساس هستند.

۲-۲- روشهای مبتنی بر مد و میانگین برای تخمین مقادیر مفقود

از میان روش‌های مبتنی بر مد و میانگین، الگوریتم k نزدیکترین همسایه از محبوب‌ترین و متداول‌ترین روش‌های تخمین مقادیر مفقود است [۲۳]. از دیگر روشهای ناپارامتری که برای این منظور پیشنهاد شده است، می‌توان به جنگل تصادفی[۲] (RF) اشاره نمود [۲۴]. از آنجا که KNN، از همبستگی بین ویژگی‌ها صرف‌نظر می‌کند، در مقاله [۲۵] یک رویکرد ناپارامتری دیگر با استفاده از نزدیک‌ترین همسایه برای این منظور پیشنهاد شده است. در مقاله [۲۶] یک جنگل تصادفی وزندار لاپلاسی تطبیقی، برای تخمین مقادیر مفقود پیشنهاد شده است. در این رویکرد، وزن ویژگی به صورت پویا در هنگام ساخت مدل تنظیم می‌شود، که احتمال انتخاب ویژگی‌های با اهمیت بالا را افزایش می‌دهد.

پوروار و همکارانش [۲۷] یک مدل پیش‌بینی ترکیبی برای کامل کردن داده‌های ناقص ارائه کردند. در این مدل تکنیک‌های مختلف برای کامل کردن داده‌ها بکار رفته و از خوشه‌بندی K-means ساده استفاده شده است. مدل ترکیبی پیشنهاد شده برای اولین بار از ترکیب خوشه‌بندی K-means با پرسپترون چندلایه استفاده کرده است. برای ارزیابی این روش از سه مجموعه داده پزشکی شامل دیابت سرخپوستان، ویسکانسین سرطان پستان، و هپاتیت استفاده شده، همچنین از معیارهای دقت، حساسیت، صحت، کاپا و سطح زیر منحنی مشخصه عملکرد سیستم برای ارزیابی استفاده شده است. نتایج نشان داد روش ارائه شده برای پیش‌بینی داده‌های ناقص در دامنه‌های پزشکی بسیار مفید است به‌خصوص زمانی که تعداد مقادیر ازدست‌رفته در مجموعه داده‌ها زیاد هستند. مقاله [۲۸] به جلوگیری از مقادیر گم‌شده در ویژگی آلبومین داده‌های کبدی می‌پردازد و از روش جبران با استفاده از

[۱] Graph Convolutional Network

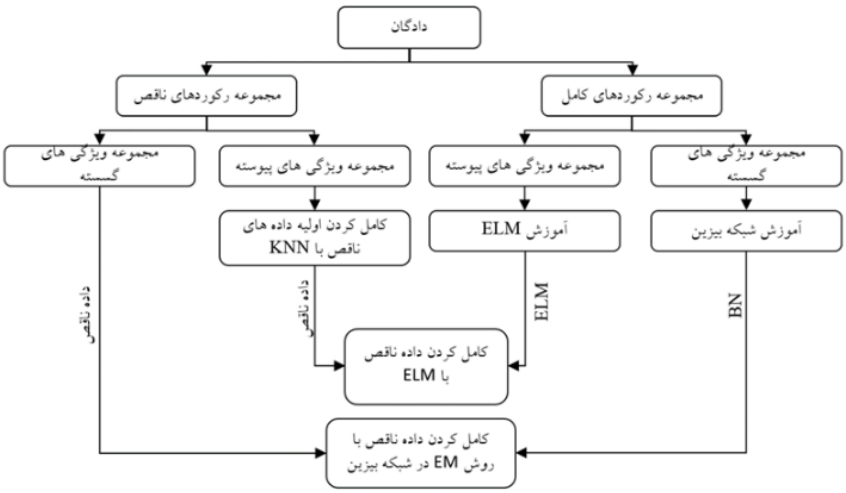
[۲] Random Forest

الگوریتم نزدیک‌ترین همسایه استفاده می‌کند. در این تحقیق، مقادیر K=3، K=5، K=7، K=9 و K=15 مورد استفاده قرار گرفته‌اند. برای ارزیابی دقت جبران، از معیار میانگین مربعات خطا (MSE) استفاده شده است. بر اساس نتایج، بهترین دقت محاسبات در موقعیتی به دست می‌آید که K=7 باشد.

این دسته از روش‌ها مزایا و معایب خود را دارند. از مزیت‌های این روشها، سادگی و قابلیت اجرای آسان آن‌هاست. همچنین استفاده از میانگین یا مد به عنوان تخمین‌های مقادیر مفقود، می‌تواند توزیع مرجع داده‌ها را حفظ کند. در مقابل این روش‌ها اطلاعات پراکندگی داده‌ها را لحاظ نمی‌کنند و همچنین ممکن است به مقادیر پرت حساس باشند. همچنین اگر داده‌ها تنوع زیادی داشته باشند، تخمین میانگین و مد ممکن است نادرست باشد و این روشها نتوانند روابط پیچیده بین متغیرها را مدل کنند که این، منجر به از دست رفتن اطلاعات مهم در داده‌ها می‌شود.

۳- روش پیشنهادی

در تحلیل داده‌ها برخی مشاهدات به دلایل گوناگون و روش‌های متفاوت، گمشده محسوب می‌شوند؛ چگونگی برخورد با این مشاهدات به دلیل اهمیت نتایج حاصل از آن‌ها به‌ویژه در تصمیم‌گیری‌های حساس، از اهمیت به‌سزایی برخوردار است. پیش‌ازاین، برای غلبه بر مشکل داده‌های گمشده مرسوم‌ترین روش، حذف داده‌های گمشده بود که منجر به داده‌هایی با کیفیت پایین و به تبع آن تحلیل و استخراج نتایج غیر مطلوبی می‌شد، اما امروزه با پیشرفت‌های علمی در حوزه‌های گوناگون و پیدایش روش‌های توانمند آماری، می‌توان پیش از مدل‌سازی داده‌های ناکامل، مقادیر گمشده را با مقادیر مناسب جایگذاری یا برآورد کرد. در این پژوهش، به تخمین داده‌های مفقود در مجموعه داده هپاتیت بر اساس ترکیبی از شبکه بیزین و ماشین یادگیری مفرط پرداخته شده است. مدل پیشنهادی در شکل ۱ نشان داده شده است؛ همان‌طور که مشاهده می‌شود، روش پیشنهادی در سه مرحله به تخمین داده‌های گمشده بیماری هپاتیت پرداخته است. در مرحله اول پیش‌پردازشی بر روی مجموعه داده انجام شده تا داده‌ها برای ورود به مدل اصلی آماده شوند، سپس داده‌های مفقود گسسته توسط شبکه بیزین و سایر داده‌های مفقود پیوسته توسط ELM تخمین زده شد. مراحل روش پیشنهادی در ادامه تشریح شده است.



شکل ۱- مدل پیشنهادی

۳-۱ آماده‌سازی مجموعه داده

در اولین مرحله از روش پیشـــنهادی، پیش‌پردازشـــی بر روی مجموعه داده بیماری هپاتیت اعمال می‌گردد؛ این پیش‌پردازش شـامل جداسـازی رکوردهای مجموعه داده با داده‌های مفقود از سـایر رکوردها اسـت، زیرا قصـد داریم از شـبکه بیزین و ماشین یـادگیری مفرط برای تخمین داده‌های مفقود اســتفاده کنیم، بنابراین باید این شـــبکه‌ها را با داده‌های کامل آموزش دهیم. پس از جداسازی رکوردها با داده‌های مفقود، ویژگی‌های پیوسته و گسسته مجموعه داده نیز جدا می‌گردد، زیرا قصد داریم با شبکه بیزین داده‌های مفقود گسسته و با ماشین یادگیری مفرط سایر داده‌های مفقود را تخمین بزنیم.

۳-۲ تخمین داده‌های مفقود با شبکه بیزین

در روش پیشنهادی از شبکه بیزین برای تخمین داده‌های مفقود گسسته استفاده شده است؛ یکی از روش‌های برآورد پارامترها و مقادیر داده‌های گمشـده در شـبکه بیزین اسـتفاده از الگوریتم امید ریاضـی- بیشینه‌سازی است [۲۹] تا [۳۱]؛ این الگوریتم یک روش تکرارشونده است که به دنبال یافتن برآوردی با بیشترین درست نمایی پارامترهای یک توزیع پارامتری است. الگوریتم EM به‌منظور تخمین داده‌های مفقود در هر تکرار دو گام را شـامل می‌شـود؛ در گام اول امید ریاضی داده‌های گمشده به شرط داده‌های مشـاهده شـده، محاسـبه می‌شود، سپس این امید ریاضی‌ها را به‌جای داده‌های گمشده قرار می‌دهند و پارامترهای موردنظر برآورد می‌شـوند. در گام ماکزیمم کردن بعد از جایگذاری اعداد اولیه به‌جای داده گمشـده، لگاریتم تابع درسـتنمایی حداکثر می‌شـود؛ بنابراین در این مرحله ابتدا شبکه بیزین بر اساس رکوردهای کامل ویژگی‌های گسسته ایجاد و آموزش می‌یابد، سپس مقادیر مفقود با استفاده از الگوریتم EM تخمین زده می‌شود. به بیان دیگر مراحل این فاز از روش پیشنهادی به صورت زیر است:

الف) ساختن مدل شبکه بیزین با استفاده از رکوردهای آموزشی

ب) شناسایی فیلدهای مفقود در داده‌های آزمایشی

ج) تخمین مقادیر مفقود بر اساس مدل بیزین اولیه و به کمک الگوریتم بیشینه‌سازی امید ریاضی (EM)

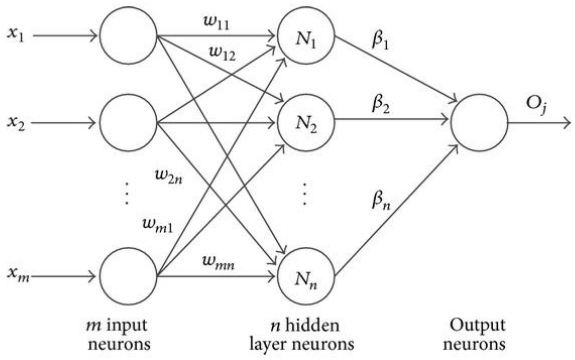
۳-۳ تخمین داده‌های مفقود با ماشین یادگیری مفرط

در روش پیشنهادی از ELM برای تخمین داده‌های مفقود پیوسته استفاده شده است؛ ماشین یادگیری مفرط، یک الگوریتم جدید در یادگیری ماشین است که برای شبکه عصبی تک لایه مخفی معرفی شده و به‌طور تحلیلی وزن‌های خروجی‌های بهینه را محاسبه می‌کند. در ELM وزن اتصال ورودی به گره‌های پنهان به‌صورت تصادفی است. این وزن بین خروجی و گره‌های پنهان در یک مرحله به دست می‌آید. این مدل می‌تواند عملکرد خوب و یادگیری هزار برابر سریع‌تر از شبکه‌های آموزش دیده با پس انتشار باشد[۳۲]. ELM یک فناوری ضروری است که عملکرد بسیار خوبی را برای هر دو مسائل طبقه‌بندی و رگرسیون با سرعت یادگیری بالا فراهم می‌کند، از دیگر مزایای ELM می‌توان سرعت یادگیری بالا، کارایی بالا، مناسب برای تمام توابع فعال‌سازی غیرخطی و مناسب برای توابع فعال‌سازی پیچیده را نام برد. الگوریتم ELM به صورت زیر است [۳۳] و [۳۴]:

با توجه به مجموعه آموزش زیر

$$\{x_i, t_i \mid x_i \in R^n, t_i \in R^m, i = 1, \ldots, N\}$$

یک تابع فعال سازی g(x) و تعداد نرونهای مخفی N را در نظر بگیرید. سپس این مراحل انجام می‌گردد: ۱) به طور تصادفی وزن-های w_i و بایاس b_i با توجه به تابع چگالی احتمال پیوسته تخصیص داده می‌شود. ۲) محاسبه لایه پنهان با ماتریس خروجی H. ۳) محاسبه وزن‌های خروجی. ساختار ELM به صورت شکل ۲ است.



شکل ۲- ساختار ماشین یادگیری مفرط[۳۲]

در بسیاری از کاربردها، تعداد نرون‌های مخفی خیلی کم‌تر از تعداد نمونه‌های یادگیری می‌باشند، در واقع $k \ll N$، و H یک ماتریس غیر مربعی است وممکن است $\{\beta_i, W_i, b_i\}_{i=1}^{k}$ که $H\beta = T$ وجود نداشته باشد. در روش ELM وزن‌های ورودی W_i و بایاس‌های لایه مخفی b_i، SLFNs به روزرسانی نمی‌شوند، بلکه به صورت تصادفی انتخاب می‌شوند و ثابت هستند. این معادل با نگاشت نمونه‌ها به یک فضای ویژگی تصادفی است؛ بنابراین، آموزش SLFNs معادل با یافتن راه حل خطای مربعی حداقل $\hat{\beta}$ در سیستم خطی $H\beta = T$ می‌باشد. کوچک‌ترین راه حل حداقل-مربعات واحد سیستم خطی به صورت $\hat{\beta} = H^{\dagger}T$

می‌باشد. که $H^{\dagger}$، معکوس تعمیم یافته مور-پنروز ماتریس H خروجی لایه مخفی است. این روش تمایل به رسیدن به یک عملکرد تعمیم یافته مناسب دارد.

در روش پیشنهادی از ELM برای تخمین داده‌های مفقود پیوسته استفاده شده است. به این منظور ابتدا شبکه ELM بر اساس مجموعه رکوردهای کامل آموزش می‌بیند. سپس مقادیر مفقود داده‌های آزمایشی بر اساس الگوریتم KNN مقداردهی اولیه می‌شوند. پس از این مرحله داده‌های تکمیل‌شده اولیه وارد مدل ELM می‌شوند. انتظار می‌رود خروجی مدل ELM تخمین دقیقتری از مقادیر ورودی اولیه ارائه کند. درنهایت خروجی ELM برای فیلد مفقود مبنای تخمین مقدار مفقود قرار می‌گیرد. مراحل این بخش از روش پیشنهادی به شرح زیر است:

الف) آموزش مدل ELM بر روی داده‌های بدون مقدار مفقود به نحوی که وزنهای شبکه W حاصل شود.

ب) پر کردن مقادیر مفقود به کمک روش KNN به عنوان مقادیر اولیه

ج) ورود رکورد با مقادیر پرشده اولیه به ELM و تخمین خروجی با استفاده از شبکه.

د) جایگذاری مقدار جدید با مقادیر پرشده اولیه

برای استفاده از ELM باید پارامترهای آن شامل تعداد نرون لایه مخفی و تابع فعال مقداردهی شود. در این پژوهش حالات مختلف با تعداد ۵، ۱۰ و ۲۰ نرون در لایه مخفی با توابع فعالسازی سیگموئید، سینوسی، Hardlim و Radbas [۳۵] مورد بررسی قرار گرفت.

۳-۴ طبقه‌بند پیشنهادی

طبقه‌بندها ابزاری هستند جهت شناسایی دقیق و کامل از یک ماهیت؛ به‌بیان‌دیگر طبقه‌بندی فرایندی برای پیدا کردن مدلی است که طبقه‌های موجود در داده‌ها را تعریف و متمایز می‌کند، با این هدف که بتوان از این مدل برای پیش‌بینی طبقه رکوردهایی که برچسب طبقه آن‌ها (متغیر هدف) ناشناخته است، استفاده نمود. در حقیقت در طبقه‌بندی برخلاف پیش‌بینی، هدف پیش‌بینی مقدار یک متغیر گسسته است. در روش پیشنهادی پس از تخمین داده‌های مفقود با ترکیب روش BN و ELM از طبقه بندهای مختلف برای تعیین میزان دقت روش پیشنهادی بر روی مجموعه داده هپاتیت استفاده شده است. در این پژوهش طبقه بندهای بیزین، ماشین یادگیری مفرط و نزدیک‌ترین همسایه مورد بررسی قرار گرفته است.

۴- آزمایش‌ها

در این بخش نتایج ارزیابی روش پیشنهادی بیان شده است. ابتدا مجموعه داده و معیارهای ارزیابی بیان شده، سپس به تحلیل نتایج پرداخته شده است. روش پیشنهادی در محیط متلب پیاده‌سازی شده است و آزمایشات و ارزیابی‌ها بر روی یک سیستم کامپیوتری با ویندوز ۷ و پردازنده ۴ هسته‌ای انجام شده است.

۴-۱ مجموعه داده

در این پژوهش از مجموعه داده هپاتیت که از سایت UCI دریافت شده، استفاده شده است [۳۶]. این مجموعه داده دارای ۱۵۵ نمونه (۳۲ نمونه از کلاس هپاتیت بدخیم و ۱۲۳ نمونه از کلاس هپاتیت خوش‌خیم) با ۱۹ ویژگی است، ۱۷,۶۳ درصد این مجموعه‌داده دارای مقادیر مفقود است. متغیرهای این مجموعه داده در جدول ۱ نشان داده شده است؛ همان‌طور که مشاهده می‌شود ۱۳ ویژگی این دادگان گسسته و ۶ ویژگی پیوسته است.

جدول ۱- ویژگی‌های مجموعه داده هپاتیت

ردیف	نام ویژگی	بازه مقادیر
۱	سن	[10-80]
۲	Bilirubin	[0.39-4]
۳	Alk Phosphate	[33-250]
۴	Sgot	[13-500]
۵	Albumin	[2.1-6]
۶	Protime	[10-90]
۷	Steroid	Yes, No
۸	Antivirals	Yes, No
۹	Fatigue	Yes, No
۱۰	Malaise	Yes, No
۱۱	Anorexia	Yes, No
۱۲	Liver Big	Yes, No
۱۳	Liver Firm	Yes, No
۱۴	Spleen Palpable	Yes, No
۱۵	Spiders	Yes, No
۱۶	Ascites	Yes, No

۱۷	Varices	Yes, No
۱۸	Histology	Yes, No
۱۹	جنسیت	مرد- زن

۴-۲ معیارهای ارزیابی

در این بررسی از معیارهای دقت، صحت، فراخوانی و جذر میانگین مربعات خطا برای ارزیابی روش پیشنهادی استفاده شده است. دقت مهم‌ترین معیار برای تعیین کارایی یک روش است؛ این معیار دقت کل یک روش را محاسبه می‌نماید و نشان‌دهنده این حقیقت است که روش پیشنهادی چند درصد از کل مجموعه رکوردهای آزمایشی را به‌درستی تشخیص داده است. دقت الگوریتم بر اساس مفاهیم مطرح شده در ماتریس درهم‌ریختگی با توجه به رابطه (۱) قابل محاسبه می‌باشد.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100 \qquad (۱)$$

معیار فراخوانی، دقت دسته‌بندی عدم بیماری را با توجه به کل رکوردها با برچسب عدم بیماری نشان می‌دهد (رابطه ۲). معیار صحت، دقت دسته‌بندی بیماری را با توجه به کل مواردی که دارای بیماری است، نشان می‌دهد (رابطه ۳).

$$\text{Recall} = \frac{\text{TN}}{\text{FN} + \text{TN}} \qquad (۲)$$

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}} \qquad (۳)$$

که در روابط فوق TP و TN به ترتیب مثبت صحیح و منفی صحیح هستند و FP و FN بیانگر آن دسته از رکوردهای کلاس منفی(مثبت) هستند که به اشتباه در کلاس مثبت (منفی) طبقه بندی شده اند. آزمون خطای جذر میانگین مربعات، تفاوت میان مقدار پیش‌بینی شده توسط مدل یا برآوردگر آماری و مقدار واقعی می‌باشد و یک ابزار خوبی است برای مقایسه خطاهای پیش‌بینی توسط یک مجموعه داده و به‌صورت رابطه (۴) محاسبه می‌گردد.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(actual(i) - predict(i))^2}{n}} \qquad (۴)$$

که در رابطه فوق actual و predict به ترتیب مقادیر واقعی و تخمین زده شده برای یک فیلد داده می‌باشند.

۴-۳ نتایج ارزیابی‌ها

در روش پیشنهادی بعد از کامل کردن داده‌های مفقود گسسته بر اساس شبکه بیزین، سایر داده‌های مفقود بر اساس ELM کامل گردید، درنهایت دقت طبقه‌بندی روش پیشنهادی بر اساس طبقه-بندهای BN، ELM، KNN-۱، KNN-۳ و KNN-۵ مورد بررسی قرار گرفت. برای کامل کردن داده‌های مفقود بر اساس ELM توابع فعال Sig، Hardlim،Sin و Radbas و تعداد ۵، ۱۰ و ۲۰ نرون در لایه مخفی موردبررسی قرار گرفت، همچنین برای ارزیابی از روش ۱۰-اعتبار سنجی متقابل استفاده شد.

جدول ۲- نتایج طبقه بندهای مختلف بر روی دیتاست هپاتیت با کامل کردن داده‌های ناقص بر اساس ترکیب BN و ELM

Activation Function	NO. of Hidden Neurons	Classifier	Accuracy	Precision	Recall
Sig	5	1-KNN	0.8237	0.7518	0.7269
		3-KNN	0.847	0.8271	0.7262
		5-KNN	0.864	0.8376	0.8149
		BN	0.851	0.7759	0.7562
	10	1-KNN	0.7959	0.6949	0.7007
		3-KNN	0.8823	0.8632	0.8142
		5-KNN	0.8750	0.8197	0.8150
		BN	0.8903	0.8346	0.8268
	20	1-KNN	0.8321	0.767	0.7595
		3-KNN	0.8425	NaN	0.7730
		5-KNN	0.8280	0.7687	0.7522
		BN	0.8645	0.7910	0.8105
Sin	5	1-KNN	0.8329	0.7763	0.7730
		3-KNN	0.8733	0.8017	0.8041
		5-KNN	0.8875	0.8439	0.82119
		BN	0.8774	0.8115	0.81872
	10	1-KNN	0.8250	0.7779	0.7474
		3-KNN	0.8599	0.8177	0.77211
		5-KNN	0.8567	0.8095	0.7656
		BN	0.8709	0.8010	0.8146
	20	1-KNN	0.8003	0.7044	0.7152
		3-KNN	0.8239	0.7613	0.7173
		5-KNN	0.8437	0.8084	0.7820
		BN	0.8838	0.8227	0.8227
hardlim	5	1-KNN	0.7812	0.7111	0.6567

		3-KNN	0.8250	0.7551	0.7240
		5-KNN	0.8144	0.7307	0.7266
		BN	0.8838	0.8227	0.8227
	10	1-KNN	0.8329	0.7862	0.7567
		3-KNN	0.8443	0.8091	0.7608
		5-KNN	0.8076	0.7118	0.6899
		BN	0.8774	0.8184	0.7956
	20	1-KNN	0.7705	0.7527	0.6878
		3-KNN	0.8505	NaN	0.7250
		5-KNN	0.8494	0.7663	0.7473
		BN	0.8903	0.8346	0.8268
radbas	5	1-KNN	0.8045	0.7246	0.7375
		3-KNN	0.8771	0.8407	0.8081
		5-KNN	0.8437	0.7820	0.7612
		BN	0.8774	0.8184	0.7956
	10	1-KNN	0.8349	0.7798	0.7444
		3-KNN	0.8058	NaN	0.6388
		5-KNN	0.8239	0.7496	0.7318
		BN	0.8580	0.7834	0.7834
	20	1-KNN	0.7912	0.6932	0.6894
		3-KNN	0.8250	0.7440	0.7198
		5-KNN	0.8431	NaN	0.7293
		BN	0.8645	0.7923	0.7990

نکته‌ای که در جداول این بخش قابل ذکر اسـت، مقادیر NaN برای آزمون‌های Precision و Recall است، این مقادیر در صـورتی به وجود می‌آید که مخرج کسـر برابر صفر باشد. جدول ۲ نشان‌دهنده نتایج حاصل از آزمون‌های مختلف بر روی روش پیشنهادی بر اساس طبقه بندهای 1NN، 3NN، 5NN و BN است؛ در این آزمون‌ها ویژگی‌های گسسته مفقود با شبکه بیزین و سایر ویژگی‌ها با ELM تخمین زده شد. پارامترهای تنظیم شده برای ELM شامل حالت‌های مختلف توابع فعال و تعداد ۵، ۱۰

و ۲۰ نرون در لایه مخفی است. همان‌طور که از نتایج جدول ۲ مشاهده می‌شود، در دو حالت بیشترین دقت ۰٫۸۹۰۳ حاصل شده است. پارامترهای تنظیم شده برای این دو حالت به‌صورت زیر است:

۱. کامل کردن داده‌های مفقود گسسته با BN و سایر داده‌های مفقود با ELM با تعداد ۱۰ نرون در لایه مخفی و تابع فعال Sig و طبقه‌بندی بر اساس BN

۲. کامل کردن داده‌های مفقود گسسته با BN و سایر داده‌های مفقود با ELM با تعداد ۲۰ نرون در لایه مخفی و تابع فعال Hardlim و طبقه‌بندی بر اساس BN

دقت طبقه‌بندی ELM با توابع فعالسازی مختلف و تعداد ۵، ۱۰ و ۲۰ نرون در لایه مخفی بر روی روش پیشنهادی مورد بررسی قرار گرفته‌است. نتایج نشان داد با در نظر گرفتن تابع فعال sin و ۲۰ نرون در لایه مخفی برای ELM در جهت جبران داده‌های مفقود و طبقه‌بندی مجموعه داده با ELM با تابع فعال Sig و تعداد ۲۰ نرون در لایه مخفی، دقت ۰. حاصل شده است (جدول ۳)؛ اما دقت طبقه‌بندی ELM در مقابل BN کمتر است. از این آزمایشات می‌توان نتیجه گرفت، جبران داده‌های مفقود با ترکیب BN-ELM و طبقه‌بندی بر اساس شبکه بیزین بیشترین دقت را کسب کرده است. روش پیشنهادی با روش‌های مختلف تخمین داده مورد مقایسه قرار گرفت. نتایج حاصل از مقایسه روش پیشنهادی با روش‌های Hotdeck، KNN، WKNN، BN و Mean بر اساس معیار دقت در جدول ۴ نشان داده شده است. همان‌طور که مشاهده می‌شود، روش پیشنهادی بالاترین دقت را نسبت به سایر روش‌های تخمین داده در هر سه‌طبقه بند 5NN، ELM و BN دارد.

جدول ۲- نتایج طبقه بند ELM بر روی روش پیشنهادی (تابع فعال Sin و تعداد ۲۰ نرون در لایه مخفی)

No. of Hidden Neurons	Activation Function	Accuracy	Precision	Recall
5	Sig	0.8133	NaN	0.6217
	Sin	0.65803	0.5764	0.6250
	Hardlim	0.8002	NaN	0.5378
	Radbas	0.7963	NaN	0.5846
10	Sig	0.8468	NaN	0.6714
	Sin	0.6775	0.5511	0.5737
	Hardlim	0.8062	NaN	0.5500
	Radbas	0.8077	NaN	0.7144
20	Sig	0.8583	0.7785	0.7676
	Sin	0.7353	0.6208	0.6307
	Hardlim	0.8141	NaN	0.6221
	Radbas	0.8058	NaN	0.6551

جدول ۳- نتایج مقایسه روش پیشنهادی با سایر روش‌های تخمین داده بر اساس معیار دقت

Imputation methods	Classifier		
	KNN	BN	ELM
Hotdeck	0.7389	0.8451	0.7940
KNN	0.7438	0.8516	0.6838
Mean	0.7250	0.8258	0.7950
WKNN	0.7369	0.8129	0.7338
BN	0.7448	0.8523	0.5612
BN-ELM (Proposed Method)	0.8228	0.8903	0.8583

نتایج حاصل از مقایسه روش پیشنهادی با روش‌های Hotdeck، KNN، WKNN، BN و Mean بر اساس نرخ داده‌های مفقود (۵، ۱۰ و ۱۵ درصد)و معیار RMSE در جدول ۵ نشان داده شده است. همان‌طور که مشاهده می‌شود، روش پیشنهادی کمترین خطا RMSE را در هر سه نرخ داده‌های مفقود نسبت به سایر روش‌های تخمین داده دارد.

جدول ۴- نتایج مقایسه روش پیشنهادی با سایر روش‌های تخمین داده بر اساس معیار RMSE

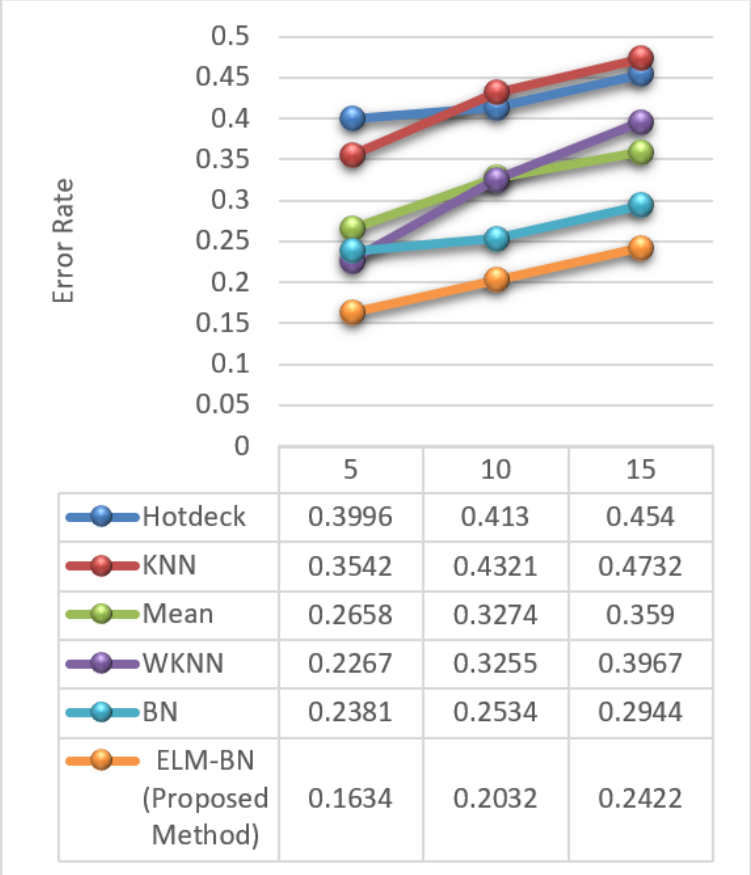
Imputation methods	Missing Rate (%)	RMSE
Hotdeck	5	۰,۳۹۹۶
	10	۰,۴۱۳۰
	15	۰,۴۵۴۰
KNN	5	۰,۳۵۴۲
	10	۰,۴۳۲۱
	15	۰,۴۷۳۲
Mean	5	۰,۲۶۵۸
	10	۰,۳۲۷۴
	15	۰,۳۵۹۰
WKNN	5	۰,۲۲۶۷
	10	۰,۳۲۵۵
	15	۰,۳۹۶۷

BN	5	۰,۲۳۸۱
	10	۰,۲۵۳۴
	15	۰,۲۹۴۴
BN-ELM (Proposed Method)	5	۰,۱۶۳۴
	10	۰,۲۰۳۲
	15	۰,۲۴۲۲

شکل ۳- نرخ خطای طبقه‌بندی روش پیشنهادی در مقایسه با سایر روش‌های تخمین داده

شـکل ۳ نشـان‌دهنده نتایج حاصـل از نرخ خطا طبقه‌بندی روش پیشـنهادی با روش‌های تخمین داده Hotdeck، KNN، WKNN، BN و Mean بر اسـاس طبقه بندهای BN، ELM و KNN است. همان‌طور که مشاهده می‌شود، روش پیشنهادی کمترین نرخ خطا طبقه‌بندی را نسبت به سایر روش‌های تخمین داده در هر سه طبقه بند KNN، ELM و BN دارد.

شـکل ۴ نشان‌دهنده نرخ خطا طبقه‌بندی روش پیشنهادی با سایر روش‌ها با نرخ ۵، ۱۰ و ۲۰ درصد نرخ گمشدگی نشان داده شده است، همانطور که مشاهده می‌شود روش پیشنهادی کمترین نرخ خطا را داراست.



شکل ۴- نرخ خطای طبقه‌بندی روش پیشنهادی با سایر روش‌های تخمین داده به ازای درصد مختلف مفقودی

جدول ۵- نتایج مقایسه روش پیشنهادی بر سایر روش‌ها

Method	Accuracy
KNN [23]	60.88
RF [24]	64.00
NNSel [25]	68.32
Laplacian weight RF [26]	72.77
MLP (5 × FC) for all missing values [15]	74.3
MLP (5 × FC) for Discrete values + 1NN for Continuous values	71.23
MLP (5 × FC) for Continuous values + 1NN for Discrete values	73.6
GRNN (5×FC) [16]	80.0
FS-fuzzy-AIRS (50–50%) [37]	81.8
KNN2	84.38
KNN-ELM	83.37
ELM-BN (Proposed Method)	**89.03**

برای جمع‌بندی بهتر روش پیشنهادی، روش مذکور با برخی از روش‌های گذشته که از دادگان مورد استفاده در این پژوهش استفاده کرده‌اند نیز مقایسه و صحت عملکرد آن‌ها مورد بررسی قرارگرفته‌است. نتایج این ارزیابی‌ها در جدول ۶ نشان داده شده است. در این بررسی علاوه بر مقایسه با دو مورد از کارهای اخیر که در مراجع [25] و [26] به آنها اشاره شده است. صحت دسته‌بندی چارچوب پیشنهادی با تعدادی از روشهای دسته بندی معرفی شده در این حوزه از جمله مراجع [15]، [26] و [31] مقایسه شده است. نتایج این ارزیابی‌ها نیز حاکی از موفقیت چارچوب پیشنهادی در بهبود صحت طبقه‌بندی پس از پرکردن مقادیر مفقود است. در جدول ۶ شبکه عصبی MLP نیز به عنوان یکی از روش‌های مورد مقایسه گنجانده شده است. در آزمایش‌ها بر روی این روش، سه رویکرد بررسی شده است. جدا از اینکه مانند تمام روشها، MLP برای پر کردن همه مقادیر مفقود استفاده شده است، در دو آزمایش دیگر، روش MLP برای پر کردن متغیرها با مقادیر گسسته و مقادیر پیوسته نیز به صورت جداگانه بررسی شده است. در هر مورد (منظور پرکردن تنها مقادیر گسسته یا تنها مقادیر پیوسته)، برای پر کردن سایر مقادیر از روش 1NN استفاده شده است.

۵- نتیجه‌گیری

یکی از موضوعاتی که غالباً در بسیاری از مسائل مختلف آماری موردتوجه قرارگرفته، بحث داده‌های مفقود است. داده‌های مفقود، یکی از مشکلاتی است که اغلب محققان و تحلیلگران در هنگام کار با مجموعه داده‌ها، با آن روبرو هستند؛ بنابراین تخمین

داده‌های مفقود یکی از مسائل مهم در تشخیص می‌باشد، زیرا اگر این داده‌ها به‌درستی تخمین زده نشوند، باعث کاهش دقت می‌شود. بااین‌حال پیش‌بینی دقیق از نتیجه بیماری یکی از مسائل چالش‌برانگیز می‌باشد، اگر مقادیر مفقود غیر آموزنده و حاوی اطلاعات مفیدی نباشد تعمیر و اصلاح آن ساده است و این در صورتی است که مقادیر مفقود نسبتاً کم باشد، ولی اگر مقادیر مفقود زیاد باشد خطر همگرایی بالا می‌رود و راه‌حل‌های اشتباه افزایش می‌یابند؛ بنابراین ما نیاز به رویکرد قوی و مقیاس‌پذیر داریم تا به تخمین داده‌های ناقص بپردازیم.

در این پژوهش از ترکیب شبکه بیزین و ماشین یادگیری مفرط جهت تخمین داده‌های مفقود در مجموعه داده هپاتیت استفاده شد. روش پیشنهادی بر اساس معیارهای دقت، Precision، Recall و RMSE مورد ارزیابی قرار گرفت. نتایج نشان داد، جبران داده‌های مفقود با ترکیب BN-ELM و طبقه‌بندی بر اساس شبکه بیزین بیشترین دقت ۰٫۸۹۰۳ را کسب کرده است. همچنین روش پیشنهادی با سایر روش‌های تخمین داده از جمله Hotdeck، KNN، WKNN، Mean و تعدادی از روش‌های اخیر بر اساس طبقه بندهای BN، ELM و KNN مورد مقایسه قرار گرفت. نتایج برتری روش پیشنهادی را در تخمین داده‌های مفقود نشان داد.

۶- مراجع

[1] P. Meesad and G. G. Yen, "Combined numerical and linguistic knowledge representation and its application to medical diagnosis," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 33, pp. 206-222, 2003.

[2] A. Ciampi, J. Thiffault, J.-P. Nakache, and B. Asselain, "Stratification by stepwise regression ,correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates," Computational statistics & data analysis, vol. 4, pp. 185-204, 1986.

[3] R. B. Davis and J. R. Anderson, "Exponential survival trees," Statistics in Medicine, vol. 8, pp. 947-961, 1989.

[4] P. M. Rancoita, M. Zaffalon, E. Zucca, F. Bertoni, and C. P. De Campos, "Bayesian network data imputation with application to survival tree analysis," Computational Statistics & Data Analysis ,vol. 93, pp. 373-387, 2016.

[5] R.J. Little and D. B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, 2014.

[6] S. Van Buuren, Flexible Imputation of Missing Data, CRC press, 2012.

[7] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," Expert Systems with Applications, vol. 38, pp. 9014-9022, 2011.

[8] B .Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," Expert Systems with Applications, vol. 41, pp. 1476-1482, 2014.

[9] I. Maglogiannis, E. Loukis, E. Zafiropoulos, and A. Stasis, "Support vectors machine-based identification of heart valve diseases using heart sounds," Computer methods and programs in biomedicine, vol. 95, pp. 47-61, 2009.

[10] S. Zeuzem, R. Ghalib, K. R. Reddy, P. J. Pockros, Z. B. Ari ,Y. Zhao, et al., "Grazoprevir–Elbasvir Combination Therapy for Treatment-Naive Cirrhotic and Noncirrhotic Patients with Chronic Hepatitis C Virus Genotype 1, 4, or 6 Infection A Randomized TrialC-EDGE Treatment-Naive Trial of Grazoprevir–Elbasvir," Annals of internal medicine, vol. 163, pp. 1-13, 2015.

[11] M. F. Keller, M. Saad, J. Bras, F. Bettella, N. Nicolaou, J. Simón-Sánchez, et al., "Using genome-wide complex trait analysis to quantify 'missing heritability'in Parkinson's disease," Human molecular genetics, vol. 21, pp. 4996-5009, 2012.

[12] A. B. Wilson, "Predicting a Missing Baseline Value in a Rare Disease Registry Using Extrapolation of Mixed Models Estimates," Pharmacoepidemiology and Drug Safety, vol. 22, p. 230, 2013.

[13] D. Sovilj, E. Eirola, Y. Miche, K.-M. Björk, R. Nian, A. Akusok, et al., "Extreme learning machine for missing data using multiple imputations," Neurocomputing, vol. 174, pp. 220-231, 2016.

[14] M. S. B. Sehgal, I. Gondal, and L. S. Dooley, "Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data," Bioinformatics, vol. 21, pp. 2417-2423, 2005.
[15] L. Ozyilmaz and T. Yildirim, "Artificial neural networks for diagnosis of hepatitis disease," in Neural Networks, 2003. Proceedings of the International Joint Conference on, pp. 586-589, 2003.
[16] W. Duch, R. Adamczak, and K. Grabczewski, "Optimization of logical rules derived by neural procedures," in Neural Networks, 1999. IJCNN'99. International Joint Conference on, pp. 669-674, 1999.
[17] L. Folguera, J. Zupan, D. Cicerone, and J. F. Magallanes, "Self-organizing maps for imputation of missing data in incomplete data matrices," Chemometrics and Intelligent Laboratory Systems, vol. 143, pp. 146-151, 2015.
[18] S. Arciniegas-Alarcón, M. García-Peña, W. J. Krzanowski, C. Rengifo, Missing value imputation in a data matrix using the regularised singular value decomposition, MethodsX, 11, 2023, 102289, https://doi.org/10.1016/j.mex.2023.102289.
[19] W.C. Lin, C.F. Tsai, J.R. Zhong, Deep learning for missing value imputation of continuous data and the effect of data discretization, Knowledge-Based Systems, 239, 2022, 108079, https://doi.org/10.1016/j.knosys.2021.108079.
[20] W. Liu, L. Luo, L. Zhou. Online missing value imputation for high-dimensional mixed-type data via generalized factor models, Computational Statistics & Data Analysis, 187, 2023, 107822. https://doi.org/10.1016/j.csda.2023.107822.
[21] A. Liguori, R. Markovic, M. Ferrando, J. Frisch, F. Causone, C. van Treeck. Augmenting energy time-series for data-efficient imputation of missing values, Applied Energy, 334, 2023, 120701, https://doi.org/10.1016/j.apenergy.2023.120701.
[22] S. Jeong, C. Joo, J. Lim, H. Cho, S. Lim, J. Kim, A novel graph-based missing values imputation method for industrial lubricant data, Computers in Industry, 150, 2023, 103937, https://doi.org/10.1016/j.compind.2023.103937.
[23] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics, vol. 17, no 6, pp. 520–525, 2001.
[24] D. J. Stekhoven, P. Bühlmann, Missforest—non-parametric missing value imputation for mixed-type data, Bioinformatics, vol. 28, no. 1, pp. 112–118, 2012.
[25] S. Faisal, G. Tutz, Imputation methods for high-dimensional mixed-type datasets by nearest neighbors, Comput. Biol. Med., 104577, 2021.
[26] L. Ren, A.S. Seklouli, H. Zhang, T. Wang, A. Bouras, An adaptive Laplacian weight random forest imputation for imbalance and mixed-type data, Information Systems, vol. 111, 102122, 2023.
[27] A. Purwar and S. K .Singh, "Hybrid prediction model with missing value imputation for medical data," Expert Systems with Applications, vol. 42, pp. 5621-5631, 2015.
[28] A. Surya Alianso; L. Syafaah; A. Faruq, K-nearest neighbor imputation for missing value in hepatitis data, AIP Conf. Proc. 2453, 020057,2022.
[29] C. J. Wu, "On the convergence properties of the EM algorithm," The Annals of statistics, pp. 95-103, 1983.
[30] G. McLachlan and T. Krishnan, The EM algorithm and extensions vol. 382: John Wiley & Sons, 2007.
[31] S. K. Ng, T. Krishnan, and G. J .McLachlan, "The EM algorithm," in Handbook of computational statistics, ed: Springer, pp. 139-172, 2012.
[32] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," Neurocomputing, vol. 70, pp. 489-501, 2006.
[33] G-.B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, pp. 513-529, 2012.
[34] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," Neurocomputing, vol. 101, pp. 229-242, 2013.
[35] J. Moody and Christian J. Darken. Fast learning in networks of locally-tuned processing units. Neural Comput. 1, 2, 281–294, 1989. https://doi.org/10.1162/neco.1989.1.2.281
[36] hepatitis Data set: https://archive.ics.uci.edu/ml/datasets/hepatitis.
[37] K. Polat and S. Güneş, "A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS," Computer methods and programs in biomedicine, vol. 88, pp. 164-174, 2007.