مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره دوم شماره اول. ص۳۲-۲٤، ۱۳۹٤

جانشینی مقادیر مفقود و تأثیر آن بر دقت کلاسه بندی در داده کاوی پزشکی

حمیدرضا طهماسبی[1]*، ملیحه آموزگار[1]، هادی آدینه[1]

• دریافت مقاله: ۹۴/۳/۱ • پذیرش مقاله: ۹۴/۳/۲۵

مقدمه: وجود مقادیر مفقود در داده‌های پزشکی می‌تواند تمام فرآیند داده کاوی و تفسیرهای حاصل را تحت تأثیر قرار دهد. بنابراین برخورد با این مقادیر ضروری می‌باشد. در این پژوهش تأثیر روش‌های مختلف برخورد با مقادیر مفقود بر روی دقت کلاسه‌بندی داده‌های پزشکی مورد ارزیابی قرار گرفت.

روش: در این مطالعه، تأثیر روش‌های معروف جانشینی مقادیر مفقود شامل Mean/mode، Hot Deck، K-Nearest Neighbor، Maximum Possible Value، All Possible Value، Case Deletion و Regression بر روی دقت کلاسه‌بندی مجموعه داده‌های پزشکی سرطان سینه، ناراحتی قلبی، بیماری‌های پوستی، هپاتیت، تیروئید، دیابت، تومور اولیه، بیماران کبدی، سرطان ریه و بعد از جراحی، به ازای شش نرخ مختلف مقادیر مفقود، ارزیابی شد. در آزمایش‌ها از دو کلاسه‌بند شبکه‌های عصبی و نزدیکترین k همسایه در نرم افزار داده کاوی Weka استفاده شد. برای تخمین دقت، از روش 10-Fold cross validation استفاده شد.

نتایج: نتایج نشان داد برای کلاسه‌بند شبکه‌های عصبی، همه روش‌های جانشینی در برابر نرخ‌های مختلف مقادیر مفقود، تأثیرات متفاوتی در دقت کلاسه‌بندی داشتند. برای کلاسه‌بند نزدیکترین k همسایه، روش جانشینی Mean/mode در مقایسه با سایر روش‌ها تقریباً با افزایش نرخ مقادیر مفقود، باعث افزایش دقت کلاسه‌بندی گردید. در مجموع، هیچ یک از روش‌های جانشینی به ازای همه نرخ-های مختلف مقادیر مفقود، همواره بیشترین دقت را نتیجه نداده و برتری نداشت.

نتیجه‌گیری: تحلیل نتایج نشان می‌دهد روش‌های جانشینی بررسی شده به ازای همه نرخ‌های مختلف از مقادیر مفقود شده لزوماً باعث بهبود دقت کلاسه‌بندی نگردیده و هیچ کدام از روش‌های جانشینی بررسی شده بهترین روش نیستند.

کلید واژه‌ها: مقادیر مفقود، روش‌های جانشینی، داده کاوی پزشکی، کلاسه‌بندی

• **ارجاع:** طهماسبی حمیدرضا، آموزگار ملیحه، آدینه هادی. جانشینی مقادیر مفقود و تأثیر آن بر دقت کلاسه بندی در داده کاوی پزشکی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۴؛ ۲(۱):۳۲-۲۴.

۱. کارشناسی ارشد مهندسی کامپیوتر، مربی، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد کاشمر، کاشمر، ایران.

* **نویسنده مسؤول:** خراسان رضوی، کاشمر، دانشگاه آزاد اسلامی واحد کاشمر

• **شماره تماس:** ۰۹۱۵۱۰۴۶۱۱۷

• **Email:** htahmasebi2002@yahoo.com

مقدمه

پزشکی مدرن حجم انبوهی از اطلاعات ذخیره شده در پایگاه داده‌های پزشکی را تولید می‌کند. رشد سریع این پایگاه‌های داده، انگیزه‌ای شده است تا محققین پزشکی از تکنیک‌های داده کاوی برای استخراج دانش از این پایگاه داده‌ها استفاده کنند. داده کاوی فرآیند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌ها، برای کشف الگوهای موجود و روابط ناشناخته میان داده‌ها می‌باشد. وجود درصد زیادی از مقادیر مفقود شده در داده‌های پزشکی، معمول می‌باشد [۱]. این مقادیر به دلایل مختلفی از جمله سهل انگاری و خطای انسانی در ورود داده‌ها، خطاهای دستگاه‌ها و تجهیزات اندازه‌گیری، اندازه‌گیری‌های نادرست، امتناع از پاسخ یا تکمیل برخی فیلدهای پرسشنامه در یک مجموعه داده ایجاد می‌شوند. اجتناب از داده‌های مفقود شده در مجموعه داده‌های پزشکی، حتی اگر نهایت دقت هم در جمع آوری داده‌ها شود، باز هم غیر ممکن است. این مقادیر ممکن است اطلاعات با اهمیتی درباره مجموعه داده‌ها را مخفی نگه دارند و باعث ایجاد مشکلات مختلفی در کشف دانش و به کارگیری الگوریتم‌های داده کاوی شوند [۲،۳].

اغلب الگوریتم‌های داده کاوی با این فرض طراحی شده‌اند که هیچ مقدار مفقودی در مجموعه داده‌ها وجود ندارد. بنابراین برخورد با این مقادیر در مرحله پیش پردازش داده‌ها بسیار ضروری می‌باشد که به جانشینی مقادیر مفقود مشهور می‌باشد [۴،۵]. به خصوص در حالتی که مجموعه داده شامل حجم زیادی از مقادیر مفقود باشد، برخورد مناسب با این مقادیر می‌تواند به طور قابل توجهی کیفیت داده کاوی را بالا ببرد [۶]. جانشینی مقادیر مفقود، همچنان یک موضوع چالش انگیز در یادگیری ماشین و داده کاوی به شمار می‌رود [۷]. حذف نمونه-های حاوی مقادیر مفقود در یک مجموعه داده، ممکن است باعث شود ویژگی‌ها و خصوصیات مجموعه داده اصلی حفظ نشود. هچنین منجر به از دست رفتن نمونه‌های زیادی از مجموعه داده شده و در نتیجه اندازه مجموعه داده کاهش یابد که باعث کاهش کارایی داده کاوی و تحلیل‌ها می‌گردد [۸].

در این مقاله، تأثیر روش‌های معروف جانشینی مقادیر مفقود شامل Mean/mode، Hot Deck، KNN (K-Nearest Neighbor)، Maximum Possible Value، All Possible Value، Case Deletion و Regression بر روی دقت کلاسه بندی ده مجموعه داده پزشکی سرطان سینه، ناراحتی قلبی، بیماری‌های پوستی، هپاتیت، تیروئید، دیابت، تومور اولیه، بیماران کبدی، سرطان ریه و اطلاعات بعد از عمل جراحی بیماران مورد ارزیابی قرار می‌گیرند. در آزمایش‌ها از دو کلاسه‌بند شبکه‌های عصبی (Artificak Neural Network) و نزدیکترین k همسایه (K-Nearest Neighbor) که جزء مشهورترین و پرکاربردترین الگوریتم‌های کلاسه‌بندی در داده کاوی پزشکی محسوب می‌شوند، استفاده شده است. این دو کلاسه‌بند در بین کلاسه‌بندهای موجود، کمترین مقاومت را در برابر داده‌های مفقود شده دارند [۹]. آزمایش‌ها بر روی شش نرخ متفاوت از مقادیر مفقود انجام گرفت. در مجموعه داده‌های انتخابی، تنوع اندازه مجموعه داده-ها بین ۳۲ تا ۷۲۰۰ نمونه، نرخ‌های مختلف مقادیر مفقود، و وجود هر دو نوع داده‌های عددی و اسمی مدنظر بوده است.

در روش جانشینی Case Deletion، نمونه‌هایی که حاوی حداقل یک ویژگی با مقدار مفقود شده هستند، حذف شده و از آن‌ها در تحلیل‌ها استفاده نمی‌گردد [۳].

در روش Mean/Mode در صورتی که یک ویژگی حاوی مقدار مفقود شده از نوع عددی باشد با میانگین مقدار آن ویژگی در سایر نمونه‌ها جانشین می‌گردد. در صورتی که ویژگی حاوی مقدار مفقود شده از نوع اسمی باشد با مد آن ویژگی در سایر نمونه‌ها جانشین می‌گردد. به طوری که مقداری که بیشترین تکرار را در بین مقادیر آن ویژگی دارا است برای مقادیر مفقود شده آن ویژگی منظور می‌گردد [۳].

در روش Hot Deck برای هر نمونه دارای مقدار مفقود، شبیه ترین نمونه به آن پیدا شده و مقادیر مفقود با مقادیر متناظر با آن در شبیه‌ترین نمونه جایگزین می‌شود [۱۰]. اگر مقدار متناظر در شبیه‌ترین نمونه نیز مفقود شده بود، از دومین شبیه‌ترین نمونه استفاده می‌گردد. به همین ترتیب این کار آنقدر تکرار می‌شود تا بالاخره مقادیر مفقود شده جایگزین گردند.

در روش جانشینی K-Nearest Neighbor، مقادیر مفقود شده در یک نمونه با مقادیر در k نمونه از شبیه‌ترین نمونه‌ها به آن نمونه جایگزین می‌شوند. برای ویژگی‌های کیفی، مقداری که بیشترین تکرار در میان k نزدیکترین نمونه را داشته است به عنوان مقدار جانشینی انتخاب می‌شود. برای ویژگی‌های کمی، میانگین مقادیر نزدیکترین k همسایه به عنوان مقدار جانشینی انتخاب می‌شود [۱۱]. شباهت بین دو نمونه با استفاده از توابع فاصله محاسبه می‌گردد که انتخاب یک تابع فاصله مناسب و همچنین مقدار مناسب k از چالش‌های این روش می‌باشند.

در روش Maximum Possible Value در بین مقادیر قابل پذیرش برای یک ویژگی خاص بیشترین مقدار آن برای

جانشینی انتخاب می‌گردد. به عبارتی دیگر، یک ویژگی، یکی از مقادیر در یک بازه خاص یا در مجموعه خاص از مقادیر، را می تواند داشته باشد. بیشترین مقدار در این بازه یا مجموعه، برای جانشینی انتخاب می‌گردد [۱۰].

در روش All Possible Values تمام مقادیر ممکن یک ویژگی دارای مقدار مفقود، جانشین مقدار مفقود شده می‌گردند. به این صورت که یک نمونه حاوی مقدار مفقود شده با مجموعه‌ای از نمونه‌های جدید جایگزین می‌شود که در هر نمونه جدید، یک مقدار ممکن از ویژگی دارای مقدار مفقود، جایگزین مقدار مفقود شده می‌گردد. سایر مقادیر ویژگی‌های نمونه‌های جدید همان مقادیر نمونه اولیه می‌باشند. اگر در یک نمونه، چندین ویژگی وجود داشته باشد که مقدار آن‌ها مفقود شده است، ابتدا جانشینی برای اولین ویژگی دارای مقدار مفقود انجام می‌گردد، سپس برای ویژگی دوم و همین طور تا وقتی که جانشینی روی همه مقدار مفقود صورت پذیرد [۱۰].

در روش Regression یک تابع رگرسیون بر مبنای داده‌های موجود در مجموعه داده‌ها می‌سازد. در این روش با استفاده از مقادیر ویژگی‌های معلوم، مقدار ویژگی مفقود شده در یک نمونه مشخص می‌شود. به عبارت دیگر، روش رگرسیون مقدار مفقود شده یک ویژگی را بر مبنای رابطه آن ویژگی با سایر ویژگی‌ها در مجموعه داده‌ها پیش بینی می‌کند [۳].

Acuna و Rodriguez تأثیر چهار روش جانشینی پایه‌ای و ساده Deletion Case، KNN، Mean و Median در مواجه با مقادیر مفقود شده با استفاده از کلاسه‌بندهای KNN و LDA(Linear Discriminant Analysis) بر روی ۱۲ مجموعه داده مختلف شامل چهار مجموعه داده پزشکی بیماری قلبی، سرطان سینه، دیابت و هپاتیت با حداکثر ۱۲درصد مقادیر مفقود، بررسی کرده و نشان دادند که هیچ یک از روش‌های جانشینی تست شده، تأثیر قابل توجهی بر دقت کلاسه بندی ندارد [۱۲].

Batista و Monard دقت کلاسه‌بندهای درخت تصمیم C4.5 و استنتاج CN2 با به کارگیری سه روش جانشینی KNN، Mean و Median بر روی چهار مجموعه داده نسبتاً کوچک آنزیم کبدی، سرطان سینه، دیابت و جلوگیری از بارداری را تست کرده که مقادیر مفقود به صورت مصنوعی و کاملاً تصادفی به تعدادی از ویژگی‌های مجموعه داده‌ها تزریق شده‌اند [۱۳]. نتایج نشان می‌دهند که روش جانشینی KNN بهترین دقت را داشته است.

Silva و همکاران با استفاده از شبکه‌های عصبی پرسپترون چند لایه روشی برای جانشینی خودکار مقادیر مفقود ارائه و با روش‌های جانشینی میانه/مد، رگرسیون و Hot Deck مورد مقایسه و ارزیابی قرار داده‌اند [۱۴]. نتایج بررسی‌ها بر روی ۱۵ مجموعه داده مختلف شامل مجموعه داده‌های پزشکی بیماری‌های قلبی و دیابت، نشان می‌دهند که برای مجموعه داده‌های فقط شامل مقادیر عددی، تمام روش‌ها نتایج خوبی دارند. برای مجموعه داده‌های با مقادیر دسته‌ای، مدل پیشنهاد شده بهترین نتیجه را داشته است اگر چه هزینه محاسباتی آن بالا می‌باشد.

Munirah از دو روش جانشینی ساده میانگین مقادیر ویژگی کلیه نمونه‌ها و میانگین مقادیر ویژگی نمونه‌های هم کلاس با نمونه دارای مقدار مفقود، در مرحله پیش پردازش دو الگوریتم داده کاوی رگرسیون لجستیک و شبکه‌های عصبی پرسپترون چندلایه استفاده کرده و با تست بر روی مجموعه داده‌های پزشکی سرطان سینه، دیابت و تیروئید مشخص شد که برای مدل رگرسیون لجستیک، بیشترین میانگین دقت در استفاده از روش جانشینی میانگین مقادیر ویژگی کلیه نمونه‌ها و برای مدل شبکه عصبی هیچ کدام از دو روش جانشینی تأثیر بهتری نداشته اند [۱۵].

هدف کلی از انجام تحقیق حاضر، تعیین روش یا روش‌های مناسب برای جانشینی مقادیر مفقود است که باعث افزایش دقت کلاسه بندی در داده کاوی پزشکی گردد.

روش

مجموعه داده‌های مورد استفاده از مخزن داده بخش یادگیری ماشین و سیستم‌های هوشمند دانشگاه کالیفرنیا (UCI) [۱۶]، گرفته شده‌اند. مشخصات این ده مجموعه داده پزشکی استفاده شده در جدول ۱ مشاهده می‌شود. در هفت مجموعه داده، مقادیر مفقود شده وجود دارند که از هر یک از مجموعه داده‌ها حذف می‌گردند. در نتیجه می‌توان بر روی تولید مقادیر مفقود و ارزیابی کارایی روش‌های جانشینی کنترل کامل داشت [۱۷]. سپس مقادیر مفقود به صورت کاملاً تصادفی و با نرخ‌های ۵، ۱۰ ، ۲۰ ، ۳۰ ، ۴۰ و ۵۰ درصد به تمام ویژگی‌های هر یک از مجموعه داده‌های کامل به صورت مصنوعی اعمال شدند. به طوری که برای هر مجموعه داده، شش مجموعه داده حاوی مقادیر مفقود با نرخ‌های مختلف تولید می‌گردد. اغلب پژوهش‌ها و مطالعات مرتبط تولید مقادیر مفقود به صورت کاملاً تصادفی صورت گرفته است [۱۰،۱۳،۱۴،۱۸].

جدول ۱: مشخصات مجموعه داده‌ها

مجموعه داده	نوع بیماری	تعداد نمونه‌ها	تعداد ویژگی‌ها	تعداد کلاس‌ها	مقدار مفقود شده
Breast Cancer Wisconsin	سرطان سینه	۶۹۹	۱۰	۲	دارد
Pima Indians Diabetes	دیابت	۷۶۸	۸	۲	ندارد
Hepatitis	هپاتیت	۱۵۵	۱۹	۲	دارد
Heart Disease	بیماری قلبی	۲۷۰	۱۳	۲	ندارد
Thyroid	تیروئید	۷۲۰۰	۲۱	۳	ندارد
Dermatology	بیماری‌های پوستی	۳۶۶	۳۴	۶	دارد
Primary Tumor	تومور اولیه	۳۳۹	۱۷	۲۲	دارد
Indian Liver Patient	کبد	۵۸۳	۱۰	۲	دارد
Post-Operative Patient	بعد از عمل جراحی	۹۰	۸	۳	دارد
Lung Cancer	سرطان ریه	۳۲	۵۶	۲	دارد

روش‌های جانشینی Mean/mode، HD (Hot Deck)، KNN، MPV(Maximum Possible Value)، All Possible Value، APV، CD (Case Deletion) و Regression با استفاده از زبان برنامه نویسی C# در محیط Net Framework. پیاده سازی شد، سپس بر روی هر یک از مجموعه داده‌های حاوی مقادیر مفقود اعمال شده‌اند.

در روش جانشینی All Possible Value، از نسخه‌ای از آن استفاده شده است که برای یک ویژگی دارای مقدار مفقود شده در یک نمونه، فقط همه مقادیر ممکن برای آن ویژگی در همان کلاس جانشین می‌گردند [۱۷]. در روش جانشینی Hot Deck از تابع فاصله (Heterogeneous HEOM(Euclidean Overlap Metric [۱۸] استفاده می‌گردد. این تابع به صورت زیر است:

فرض کنید نمونه‌ای با n ویژگی به صورت $X=(x_1,x_2,...,x_n)$ نمایش داده شود. همچنین متغیر باینری m به این صورت تعریف گردد که $m_j=1$ اگر مقدار ویژگی x_j مفقود شده و $m_j=0$ اگر مقدار ویژگی x_j مشخص باشد. برای دو نمونه ی X_a و X_b فاصله HEOM بین آن‌ها به صورت زیر تعریف می‌گردد:

$$d(X_a,X_b)=\sqrt{\sum_{j=1}^{n} d_j(x_{aj},x_{bj})^2} \quad (۱)$$

که $d_j(x_{aj},x_{bj})$ فاصله ی بین j امین ویژگی دو نمونه X_a و X_b به صورت زیر محاسبه می‌شود:

$$d_j(x_{aj},x_{bj})=\begin{cases} 1 & if\ (1-m_{aj})(1-m_{bj})=0 \\ d_O(x_{aj},x_{bj}) & if\ x_j\ is\ a\ categorical\ attribute \\ d_N(x_{aj},x_{bj}) & if\ x_j\ is\ a\ quantitative\ attribute \end{cases} \quad (۲)$$

اگر مقدار ویژگی j در یکی از دو نمونه مفقود شده باشد $d_j(x_{aj},x_{bj})$ برابر یک (بیشترین فاصله) و در غیر این صورت $d_j(x_{aj},x_{bj})$ برابر صفر می‌گردد. اگر ویژگی j از نوع دسته‌ای باشد و این مقدار در هر دو نمونه مثل هم باشد مقدار d_O برابر صفر و در غیر این صورت برابر یک می‌گردد. اگر ویژگی j از نوع کمی باشد مقدار d_N از رابطه زیر به دست می‌آید:

$$(۳)\ d_N(x_{aj},x_{bj})=\frac{|x_{aj}-x_{bj}|}{\max(x_j)-\min(x_j)}$$

که $\max(x_j)$ و $\min(x_j)$ به ترتیب بیشترین مقدار و کمترین مقدار ویژگی j در بین تمام نمونه‌ها می‌باشند. برای محاسبه $d_N(x_{aj},x_{bj})$ مقادیر داده‌ها ابتدا به بازه بین صفر تا یک نرمال می‌شوند.

در روش جانشینی KNN نیز از تابع فاصله HEOM استفاده می‌گردد. پس از محاسبه فاصله بین نمونه‌ها در بین k نزدیکترین نمونه به نمونه دارای مقدار مفقود شده، در صورتی که ویژگی jام از نوع کمی باشد، میانه مقدار ویژگی j در k نمونه به عنوان جانشینی برای مقدار مفقود شده انتخاب می‌گردد و در صورتی که ویژگی j از نوع دسته‌ای باشد، مد مقدار ویژگی j در k نمونه به عنوان جانشینی برای مقدار مفقود شده انتخاب می‌گردد. در محاسبه فاصله به منظور جانشینی مقدار مفقود ویژگی j در یک نمونه خاص، نمونه-هایی که مقدار ویژگی j آن‌ها مفقود شده است در نظر گرفته نشدند.

پس از اعمال روش های جانشینی بر روی مجموعه داده های با نرخ های مختلف مقادیر مفقود شده، دو کلاسه بند نزدیکترین k همسایه و شبکه های عصبی پرسپترون چندلایه بر روی مجموعه داده های زیر اعمال می گردند:

- مجموعه داده‌های کامل اولیه
- مجموعه داده‌های حاوی ۶ نرخ متفاوت از مقادیر مفقود شده
- مجموعه داده‌های کامل حاصل شده از اعمال ۷ روش جانشینی بر روی هر یک از مجموعه داده حاوی یک نرخ مشخص از مقادیر مفقود

برای اعمال دو کلاسه بند مذکور از نرم افزارWeka نسخه ۳.۷.۸ استفاده شده است که پارامترهای استفاده شده برای هر الگوریتم همان مقادیر پیش فرض تعیین شده در این نرم افزار می‌باشد. مقدار k در الگوریتم KNN برابر ۵ در نظر گرفته شده است. برای تخمین کارایی کلاسه بندها از روش ارزیابی متقاطع ۱۰ تکه برابر (10-Fold cross validation) استفاده می‌شود.

نتایج

میانگین تفاضل دقت کلاسه بندی روی مجموعه داده‌های کامل حاصل از هر روش جانشینی و مجموعه داده‌های حاوی مقادیر مفقود را برای هر نرخ از مقادیر مفقود در جدول ۲ مشاهده می‌شود. این میانگین به این صورت به دست می‌آید: الگوریتم کلاسه‌بندی بر روی هر یک از ده مجموعه داده کامل حاصل از اعمال یک روش جانشینی خاص در یک نرخ مقادیر مفقود مشخص اعمال می‌گردد. از طرفی الگوریتم کلاسه‌بندی بر روی هر یک از ده مجموعه داده حاوی مقادیر مفقود با همان نرخ مقادیر مفقود مشخص، اعمال گردیده و تفاضل دقت کلاسه‌بندی برای هر مجموعه داده محاسبه می‌گردد و بالاخره میانگین ده تفاضل حاصل، محاسبه و در جدول قرار می‌گیرد. مقادیر منفی در این جدول بیانگر این هستند که استفاده از روش جانشینی نسبت به عدم استفاده از آن باعث کاهش و بدتر شدن دقت کلاسه‌بندی شده است.

در این جدول کلاسه بند نزدیکترین k همسایه به اختصار، KNN و شبکه های عصبی پرسپترون چندلایه به اختصار، ANN نامیده شده است.

در جدول ۲، روش Complete data، حالتی است که روش جانشینی خاصی روی داده ها اعمال نشده است و داده ها از همان ابتدا کامل و بدون مقادیر مفقود می باشند. برای این حالت در جدول ۲، میانگین تفاضل دقت هر کلاسه‌بند روی مجموعه داده‌های کامل اولیه و مجموعه داده‌های حاوی مقادیر مفقود برای هر نرخ از مقادیر مفقود نشان داده شده است.

از جدول ۲ مشاهده می‌شود که بیشترین بهبود برابر ۱/۸۲ درصد برای جانشینی در مجموعه داده حاوی ۴۰ درصد مقادیر مفقود در روش جانشینی Case Deletion در کلاسه‌بندی KNN می‌باشد. در این حالت کلاسه بندی روی مجموعه داده‌های کامل ۰/۵۵ درصد نسبت به مجموعه داده‌های حاوی مقادیر مفقود بهبود دقت دارد.

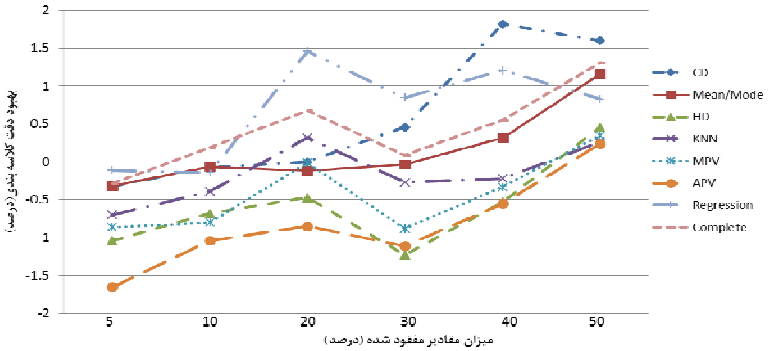
نمودار ۱ و ۲ نیز به ترتیب بهبود دقت کلاسه‌بندی دو کلاسه بند KNN و ANN را به ازای نرخ‌های مختلف مقادیر مفقود نشان می‌دهد. این دو نمودار، نشان می‌دهند که روش‌های جانشینی بررسی شده در برابر نرخ‌های مختلف مقادیر مفقود، تأثیرات متفاوتی در دقت کلاسه‌بندی داشته‌اند.

بحث و نتیجه گیری

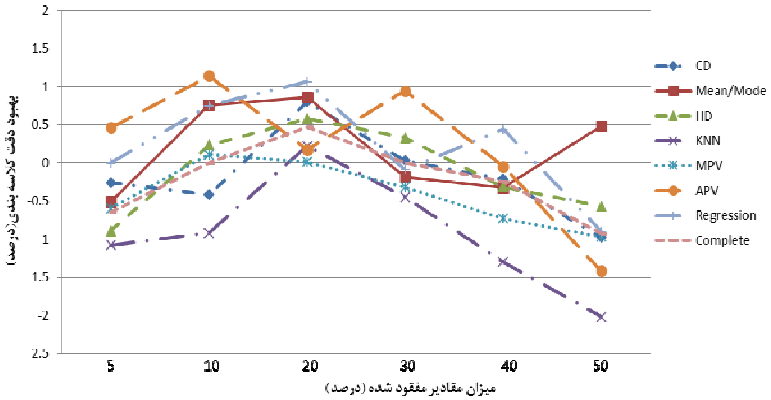
نتایج نشان می‌دهند که هیچ یک از روش‌های جانشینی بررسی شده به ازای همه نرخ‌های مختلف مقادیر مفقود، همواره بیشترین دقت کلاسه‌بندی را نسبت به سایرین نتیجه نمی‌دهد و نمی‌توان برتری روش جانشینی خاصی را در برخورد با مقادیر مفقود شده در یک مجموعه داده برای کلاسه بندی KNN را به تمام نرخ‌های مقادیر مفقود تعمیم داد. مطالعه انجام شده توسط Acuna و Rodriguez [۱۲] که بر روی فقط چهار مجموعه داده پزشکی بیماری قلبی، سرطان سینه، دیابت و هپاتیت با حداکثر ۱۲درصد مقادیر مفقود انجام شده است، برای کلاسه‌بندی KNN نیز نتیجه مشابهی داشته است. اما با افزایش نرخ مقادیر مفقود، روش جانشینی Mean/mode در مقایسه با سایر روش‌ها تقریباً باعث افزایش دقت کلاسه‌بندی گردیده است. در حالی که سایر روش‌ها کارایی منظمی نسبت به نرخ مقادیر مفقود نداشته‌اند.

جدول ۲: میانگین تفاضل دقت کلاسه بندی

روش جانشینی	نرخ مقادیر مفقود(درصد)	کلاسه‌بند KNN	کلاسه‌بند ANN	روش جانشینی	نرخ مقادیر مفقود(درصد)	کلاسه‌بند KNN	کلاسه‌بند ANN
CD	۵	-۰/۳۲	-۰/۲۶	MPV	۵	-۰/۸۶	-۰/۶
	۱۰	-۰/۰۷	-۰/۴۲		۱۰	-۰/۸	۰/۱۱
	۲۰	۰	۰/۸		۲۰	-۰/۰۱	۰/۰۱
	۳۰	۰/۴۶	۰/۰۳		۳۰	-۰/۸۸	-۰/۳۲
	۴۰	**۱/۸۲**	-۰/۲۲		۴۰	-۰/۳۳	-۰/۷۳
	۵۰	۱/۶	-۰/۹۸		۵۰	۰/۳۵	-۰/۹۷
Mean/mode	۵	-۰/۳۲	-۰/۵۱	APV	۵	-۱/۶۵	۰/۴۶
	۱۰	-۰/۰۶	۰/۷۶		۱۰	-۱/۰۴	۱/۱۴
	۲۰	-۰/۱۲	۰/۸۶		۲۰	-۰/۸۵	۰/۱۷
	۳۰	-۰/۰۳	-۰/۱۹		۳۰	-۱/۱۱	۰/۹۴
	۴۰	۰/۳۲	-۰/۳۳		۴۰	-۰/۵۵	-۰/۰۵
	۵۰	۱/۱۶	۰/۴۸		۵۰	۰/۲۴	-۱/۴۲
HD	۵	-۱/۰۴	-۰/۸۹	Regression	۵	-۰/۱۱	۰
	۱۰	-۰/۶۸	۰/۲۳		۱۰	-۰/۱۴	۰/۷۵
	۲۰	-۰/۴۷	۰/۵۸		۲۰	۱/۴۶	۱/۰۶
	۳۰	-۱/۲۳	۰/۳۳		۳۰	۰/۸۵	-۰/۰۹
	۴۰	-۰/۵۳	-۰/۳۲		۴۰	۱/۲۱	۰/۴۴
	۵۰	۰/۴۶	-۰/۵۷		۵۰	۰/۸۳	-۰/۹
KNN	۵	-۰/۷	-۱/۰۸	Complete data	۵	-۰/۲۸	-۰/۶۵
	۱۰	-۰/۳۹	-۰/۹۲		۱۰	۰/۱۹	۰
	۲۰	۰/۳۲	۰/۲۲		۲۰	۰/۶۸	۰/۴۷
	۳۰	-۰/۲۷	-۰/۴۵		۳۰	۰/۰۹	۰
	۴۰	-۰/۲۲	-۱/۳		۴۰	۰/۵۵	-۰/۲۵
	۵۰	۰/۲۷	-۲/۰۲		۵۰	۱/۳۱	-۰/۹۳



نمودار ۱: بهبود دقت کلاسه بندی KNN



نمودار ۲ : بهبود دقت کلاسه بندی ANN

با توجه به نمودار ۲ مشاهده می‌شود که همه روش‌های جانشینی بررسی شده در برابر نرخ‌های مختلف مقادیر مفقود، تأثیرات متفاوتی در دقت کلاسه‌بندی داشته‌اند. به جزء روش جانشینی Mean/Mode، در سایر روش‌ها به ازای ۵۰ درصد مقادیر مفقود، دقت کلاسه‌بندی به شدت کاهش می‌یابد. همانند کلاسه بند KNN، در این نمودار نیز کاملاً مشهود است که هیچ کدام از روش‌های جانشینی بررسی شده به ازای تمام مجموعه داده-های حاوی نرخ‌های مختلف مقادیر مفقود برتری ندارد.

مطالعه انجام شده توسط Acuna و Rodriguez [۱۲] و همچنین Munirah [۱۵] نیز نتایج مشابهی دارند. هر چند روش‌های جانشینی بررسی شده در این دو مطالعه کمتر بوده و مجموعه داده‌های پزشکی نیز فقط چهار مجموعه داده بوده‌اند. بنابراین در مجموع می‌توان نتیجه گرفت که به کارگیری روش-های جانشینی مذکور به ازای همه نرخ‌های مختلف از مقادیر مفقود شده لزوماً باعث بهبود دقت کلاسه بندی نمی‌گردد. اگر چه یک روش جانشینی خاص به ازای مقادیر مشخصی از مقادیر مفقود باعث بهبود دقت کلاسه‌بندی می‌گردد. به علاوه بهبود دقت کلاسه‌بندی رابطه‌ای با میزان مقادیر مفقود شده نداشته است. این نتایج همچنین نشان می‌دهند که هیچ کدام از روش‌های جانشینی بررسی شده بهترین روش جانشینی نیستند. مجموعه داده‌ها در کاربردهای مختلف نیازمند روش‌های جانشینی متفاوتی هستند که می‌بایست با توجه به ویژگی‌ها و خصوصیات مجموعه داده‌ها انتخاب شوند. در کارهای آتی تحلیل و بررسی رفتار روش های جانشینی بر روی مجموعه داده‌های حاوی مقادیر مفقود که کاملاً تصادفی نیستند مورد توجه خواهد بود. همچنین بررسی تأثیر روش‌های جانشینی بر روی دقت کلاسه‌بندی سایر الگوریتم‌های کلاسه بندی و بررسی سایر روش‌های جانشینی مقادیر مفقود شده نیز می‌تواند مد نظر قرار گیرد.

تشکر و قدردانی

این مقاله حاصل به شماره مجوز ۵۶۸ با عنوان " ارزیابی تأثیر روش‌های جانشینی مقادیر مفقود بر دقت کلاسه‌بندی در داده کاوی پزشکی" بوده که از دانشگاه آزاد اسلامی واحد کاشمر به عنوان حامی طرح مذکور تشکر و قدردانی می‌گردد.

References

1. Tahmasbi HR. Data mining application in medical, opportunities and challenges. National Conference Applied research in science and Engineering; 2013 Apr 24-26; Islamic Azad University, Takestan Branch; 2013.

2. Vinod NC, Punithavalli M. Classification of incomplete data handling techniques an overview. International Journal on Computer Science and Engineering. 2011;3(1):340–4.

3. Suthar B, Patel H, Goswami A. A survey: classification of imputation methods in data mining. International Journal of Emerging Technology and Advanced Engineering. 2012;2(1):309–12.

4. Doove LL, Van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. Computational Statistics & Data Analysis. 2014;72:92–104.

5. Liu Y, Brown SD. Comparison of five iterative imputation methods for multivariate classification. Chemometr Intell Lab. 2013;120:106–15.

6. Olamiti AO, Osofisan AO. Experimental comparison of missing value treatment methods in students' enrolment data. European Journal of Scientific Research. 2009;33(4):546–74.

7. Somasundaram RS, Nedunchezhian R. Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. Int J Comput Appl. 2011;21(10):14–9.

8. Rahman MG, Islam MZ. FIMUS: A framework for imputing missing values using co-appearance, correlation and similarity analysis. Knowledge-Based Systems. 2014;56:311–27.

9. Naderi S, Moghaddam N, Kabir EA. Analysis of supervised learners to extract knowledge about the lighting angels in frontal face images. Iranian Journal of Electrical and Computer Engineering. 2011;9(1):21-8. Persian.

10. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition. 2008;41(12):3692–705.

11. Zhang S. Shell-neighbor method and its application in missing data imputation. Appl Intell. 2011;35(1):123–33.

12. Acuna E, Rodriguez C. The treatment of missing values and its effect on classifier accuracy. Classification, Clustering, and Data Mining Applications. 2004; 639–47.

13. Batista GE , Monard MC. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence. 2003;17(5-6):519–33.

14. Silva-Ramírez EL, Pino-Mejías R, López-Coello M, Cubiles-de-la-Vega MD. Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Netw. 2011;24(1):121-9.

15. Munirah Y. The Impact of missing value methods and normalization techniques on the performance of data mining models [dissertation]. Universiti Utara Malaysia; 2011.

16. UCI Machine Learning Repository. [cited 2015 Mar 20]. Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html

17. Zhu B, He C, Liatsis P. A robust missing value imputation method for noisy data. Appl Intell. 2012;36(1):61–74.

18. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010;50(2):105-15.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2015; 2(1):24-32

Replacement of Missing Values and its Effect on the Classification Accuracy in Medical Data Mining

Hamid Reza Tahmasbi[1]*, Malihe Amoozgar[1], Hadi Adine[1]

• **Received:** 22 May, 2015 • **Accepted:** 15 June, 2015

Introduction: The missing values in medical data may impact the data mining process and any kind of interpretation. Thus the treatment of these missing values is a necessary task. In this research, the effect of various methods of dealing with missing values on medical data classification accuracy is evaluated.

Method: This paper studied the effect of missing data replacement methods including Mean/Mode, Hot Deck, K-Nearest Neighbor, Maximum Possible Value, All Possible Value, Case Deletion, and Regression on classification accuracy for two popular classifiers namely K-nearest-neighbor and Neural Networks from Weka Data mining tool on 10 medical datasets including Breast Cancer, Cardiac Problems, Dermatology, Hepatitis, Thyroid, Diabetes, Primary Tumor, Liver Patient, Lung Cancer and Post-Operative Patient. These were selected from the six amounts of missing values. For classification accuracy estimation, the 10-fold cross validation method is used.

Results: The results show that although the mean/mode method almost had better classification improvement that, none of the replacement methods for all amounts of missing values, is not always the most accurate classification with increasing amounts of missing values for the K-nearest-neighbor classifier. There was no supremacy for all the replacement methods against the various amounts of missing values for any of the replacement methods for all data sets with different amounts of missing values.

Conclusion: The current study shows that the replacement methods that have been evaluated for all the different rates of missing values do not necessarily improve the accuracy of classification and none of the investigated replacement methods is not absolutely the best one.

Key words: Missing values, Replacement methods, Medical Data Mining, Classification

• **Citation:** Tahmasbi HR, Amoozgar M, Adine H. Replacement of Missing Values and its Effect on the Classification Accuracy in Medical Data Mining. Journal of Health and Biomedical Informatics 2015; 2(1): 24-32.

1. M.Sc. of Computer Engineering, Lecturer of Computer Engineering Dept., Islamic Azad University Kashmar Branch, Kashmar, Iran.

***Correspondence**: Islamic Azad University Kashmar Branch, Kashmar, Razavi Korasan, Iran.

• **Tel:** 09151046117 •**Email:** htahmsebi2002@yahoo.com