بررسی الگوریتم‌های رده‌بندی در پیش‌بینی داده‌های سلامت: یک مطالعه مروری

حجت‌اله حمیدی[1]، عاطفه دارایی[2]

مقاله مروری

چکیده

داده‌کاوی ابزاری جهت استخراج اطلاعات مفید از مجموعه داده‌های عظیم، از جمله زمینه‌های مورد علاقه محققان در حوزه سلامت محسوب می‌شود. رده‌بندی، یک تابع یادگیری می‌باشد که هر داده را به یکی از دسته‌های از قبل تعریف شده، نگاشت می‌کند. بر اساس گزارش‌های سازمان بهداشت جهانی، بیماری‌های قلبی، کلیوی، دیابت و سرطان‌ها در سال ۲۰۱۲ عامل ۶۸ درصد از مرگ‌ها بوده‌اند. پژوهش حاضر، با هدف مطالعه و بررسی انواع الگوریتم‌های رده‌بندی و نتایج آن‌ها درون حوزه سلامت در مطالعات پیشین انجام شد. این مطالعه از نوع مروری- نقلی بود که در آن، مطالعات مرتبط برای بیماری‌های قلبی، سرطان سینه و دیابت از سال ۲۰۰۳ تا ۲۰۱۵ بررسی گردید. کلمات کلیدی «Data mining، Classification، Health، Heart disease، Diabetes و Breast cancer» در پایگاه‌های اطلاعاتی ScienceDirect، Elsevier، Springer و IEEE جستجو و منابع هر مقاله و مقالات استناد شده به آن نیز جمع‌آوری شد. پس از حذف مطالعات نامتناسب، ۳۴ مقاله انتخاب گردید. جمع‌بندی مطالعات نشان داد که تکرار استفاده از الگوریتم شبکه عصبی، برای هر سه بیماری بیشتر بود. الگوریتم‌های شبکه عصبی و بیز ساده برای بیماری قلبی، نزدیک‌ترین همسایگان برای سرطان سینه و شبکه عصبی برای دیابت بالاترین دقت را داشت. به طور کلی می‌توان دریافت، با وجود این که نمی‌توان با قطعیت یک الگوریتم را بهترین الگوریتم برای هر بیماری دانست، اما تعیین بهترین الگوریتم‌ها برای هر بیماری، می‌تواند برای مطالعات آینده مفید باشد.

واژه‌های کلیدی: رده‌بندی؛ داده‌کاوی؛ شبکه عصبی؛ بیماری قلبی

دریافت مقاله: ۱۳۹۴/۶/۱۱ اصلاح نهایی: ۱۳۹۵/۴/۱۲ پذیرش مقاله: ۱۳۹۵/۴/۱۴

ارجاع: حمیدی حجت‌اله، دارایی عاطفه. **بررسی الگوریتم‌های رده‌بندی در پیش‌بینی داده‌های سلامت: یک مطالعه مروری.** مدیریت اطلاعات سلامت ۱۳۹۵؛ ۱۳ (۳): ۲۴۲-۲۳۵

مقدمه

به گزارش سازمان بهداشت جهانی WHO (World Health Organization)، بیماری‌های غیر واگیر شامل بیماری‌های قلبی، کلیوی، دیابت و سرطان‌ها، در سال ۲۰۱۲ عامل ۶۸ درصد مرگ‌ها بوده‌اند (۱). بیماری قلبی، اولین عامل مرگ و میر جهانی در سال ۲۰۱۲ بود؛ به طوری که موجب مرگ ۱۷/۵ میلیون نفر شده است (۲). سکته قلبی، از مهم‌ترین دلایل مرگ و میر در ایران می‌باشد (۳). همچنین، سرطان سینه با ۵۲۱ هزار مورد منجر به مرگ، پنجمین سرطان رایج است (۴). دیابت نیز در سال ۲۰۱۲ عامل مرگ ۱/۵ میلیون نفر در سراسر جهان بوده است (۵). آمار ابتلا به این بیماری‌ها، روز به روز در حال افزایش است. داده‌های بیماران، منبع عظیمی است که نیاز به پردازش و تحلیل دارد تا بتواند موجب صرفه‌جویی‌های مالی و همچنین، کمک به پزشکان در تصمیم‌گیری شود (۷، ۶). بنابراین، افزایش پیوسته تعداد بیماران و همچنین، اهمیت جلوگیری از عواقب تشخیص نادرست یا دیرهنگام بیماری، موجب افزایش تمایل به استفاده از روش‌های با دقت و سرعت بالا، جهت تحلیل سریع و دقیق حجم عظیم داده‌های بیماران شده است. به همین دلیل، استخراج دانش و الگوهای پنهان از داده‌های حوزه سلامت می‌تواند در تصمیم‌گیری‌های سریع‌تر و دقیق‌تر به پزشکان کمک کند (۸).

ذهن انسان ممکن است که در مواجهه با حجم عظیمی از داده‌ها، دچار خطا شود یا بر اساس تجربیات پیشین استنتاج کند. از این‌رو، استفاده از روش‌هایی در استنتاج‌ها که داده‌ها را بدون فرضیه قبلی تحلیل می‌کند، می‌تواند مفید باشد (۹). داده‌کاوی، فرایند کشف اطلاعات مجموعه عظیمی از داده‌ها می‌باشد (۱۰). یکی از مهم‌ترین روش‌های داده‌کاوی، رده‌بندی (Classification) است که با توجه به مقادیر موجود، به پیش‌بینی مقادیر ناشناخته می‌پردازد (۱۱). دقت (Accuracy)، از مهم‌ترین روش‌های ارزیابی رده‌بندی است که به معنی داده‌های موجود در دسته درست نسبت به تمام داده‌ها می‌باشد (۱۲). مسأله اصلی در این مطالعه، یافتن الگوریتم‌های مناسب برای هر بیماری، جهت تشخیص صحیح و به موقع بود. از این‌رو، هدف از انجام مطالعه حاضر، بررسی الگوریتم‌های رده‌بندی مورد استفاده در پژوهش‌های پیشین از نظر تکرار و عملکرد الگوریتم‌ها، برای این سه بیماری بود تا با جمع‌بندی مطالعات، جهت‌گیری مناسبی برای مطالعات آینده فراهم شود.

روش بررسی

در این مطالعه مروری، پژوهش‌های انجام شده طی سال‌های ۲۰۰۳ تا ۲۰۱۵ مورد بررسی قرار گرفت. مقالات با جستجو در پایگاه‌های علمی ScienceDirect، Elsevier، Springer و IEEE جمع‌آوری شد. واژه‌های

مقاله حاصل تحقیق مستقل بدون حمایت مالی و سازمانی است.

۱- استادیار، مهندسی کامپیوتر، گروه فن‌آوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران (نویسنده مسؤول)
Email: h_hamidi@kntu.ac.ir
۲- دانشجوی کارشناسی ارشد، فن‌آوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران

کلیدی مورد استفاده در این مطالعه شامل Data mining، Classification، Health، Heart disease، Diabetes و Breast cancer بود. مطالعاتی که در عنوان آن‌ها به صراحت از روش‌های داده‌کاوی به غیر از رده‌بندی نام برده شده بود (یعنی مقالاتی که از سایر روش‌های داده‌کاوی استفاده کردند)، مورد بررسی قرار نگرفت. پس از بررسی عنوان، چکیده و کل مقاله، مقالات نامتناسب حذف شد. مطالعاتی که نتایج را بر اساس معیارهای ارزیابی کمی روش‌های رده‌بندی همچون دقت مدل، ارایه داده بود، بررسی شدند و مطالعاتی که به بررسی کیفی پرداخته بود، کنار گذاشته شدند. در نهایت، ۳۴ مقاله مورد بررسی قرار گرفت. در ابتدا، مطالعات بر اساس نوع بیماری و پس از آن مقالات هر بیماری بر اساس الگوریتم‌های مورد استفاده طبقه‌بندی شد. جهت تحلیل مقالات، از آمار توصیفی در نرم‌افزار Excel نسخه ۲۰۱۳ استفاده شد.

یافته‌ها

مطالعات بسیاری در زمینه داده‌کاوی بیماری قلبی با استفاده از شبکه عصبی ANN (Artificial neural network) صورت گرفته است. Yan و همکاران، مدلی را بر اساس شبکه عصبی پرسپترون چند لایه برای تشخیص ۵ بیماری قلبی اصلی ارایه دادند که بیماری مزمن قلبی، با ۸۲/۹ درصد بالاترین دقت را داشته است (۱۳). Das و همکاران یک روش شبکه عصبی گروهی برای تشخیص بیماری قلبی پیشنهاد کردند که به دقت ۸۹/۰۱ درصدی رسیده است (۱۴). Dangare و Apte با استفاده از شبکه عصبی، J_{48} و بیز ساده Naive bayes و با افزودن ویژگی‌های چاقی و دخانیات، مدلی برای پیش‌بینی بیماری قلبی پیشنهاد نمودند. دقت شبکه عصبی و J_{48} پس از افزودن ویژگی‌های جدید، ۱۰۰ و ۹۹/۶۲ درصد بوده است (۱۵). Santhanam و Ephzibah برای پیش‌بینی بیماری قلبی، از رگرسیون و رده‌بندی شبکه عصبی پیش‌خور استفاده کردند. نتایج آن‌ها نشان دهنده عملکرد بهتر و دقت بالاتر شبکه عصبی نسبت به رگرسیون بود (۱۶). Shao Yuehjen و همکاران از روش‌های لجستیک رگرسیون LR (Logistic regression)، MARS (Multivariate adaptive regression splines) و مجموعه Rough، هر کدام در ترکیب با شبکه عصبی، روی مجموعه داده بیماری قلبی استفاده نمودند که ترکیب شبکه عصبی با MARS، بالاترین دقت را داشت (۱۷). Abuhasel و همکاران روشی شامل AdaBoost و شبکه عصبی با تابع عضویت فازی برای رده‌بندی داده‌های بیماری قلبی، صرع، پارکینسون و هپاتیت ارایه دادند. نتایج آن‌ها نشان دهنده عملکرد بهتر مدل در حالت استفاده از AdaBoost و شبکه عصبی، نسبت به حالت استاندارد شبکه عصبی بوده است (۱۸). Abdel-Aal مدلی بر اساس اجرای مکرر الگوریتم شبکه عصبی پیشنهاد داد که دقت مدل برای بیماری قلبی ۸۵ درصد و بهتر از سرطان سینه بود (۱۹). Dash و همکاران یک مدل ترکیبی از شبکه عصبی و سیستم فازی برای رده‌بندی داده‌های بیماری‌های مختلف ارایه کردند که دقت برای بیماری قلبی ۷۶/۹۶ درصد و برای سرطان سینه ۸۱/۶۹ درصد بوده است (۲۰).

تحقیق Dennis و Muthukrishnan مدل ترکیبی از سیستم فازی و الگوریتم ژنتیک را برای رده‌بندی بیماری قلبی پیشنهاد نمود که دقت ۷۶/۶۷ درصد به دست آورد (۲۱). Lahsasna و همکاران از قواعد فازی با رده‌بندهای گروهی برای تشخیص بیماری قلبی، استفاده کرد که دارای دقت ۸۴/۴۴ درصد می‌باشد (۲۲). محمودآبادی و تبریزی یک سیستم فازی برای بیماری قلبی شامل درخت تصمیم و الگوریتم رقابت استعماری پیشنهاد دادند که به دقت ۹۴/۹۲ درصد رسیده است (۲۳). کهرمانلی و الهوردی مدلی شامل شبکه عصبی و شبکه عصبی فازی را برای دیابت و بیماری قلبی توسعه دادند. دقت مدل پیشنهادی روی داده‌های بیماری قلبی، ۸۶/۸ درصد و برای داده‌های دیابت، ۸۴/۲۴ درصد بوده است (۲۴).در پژوهش Nguyen و همکاران برای رده‌بندی داده‌های بیماری، از ترکیب مدل استاندارد افزاینده فازی و الگوریتم ژنتیک استفاده گردید که دقت برای بیماری قلبی، ۷۸/۷۸ درصد و برای سرطان سینه، ۹۷/۴ درصد به دست آمد (۲۵). آن‌ها همچنین، روشی برای رده‌بندی داده‌های بیماری پیشنهاد دادند که از سیستم فازی نوع ۲ استفاده می‌کند. دقت مدل برای بیماری قلبی و سرطان سینه به ترتیب ۸۱ و ۹۷/۹ درصد بود (۲۶).

Anbarasi و همکاران مدلی برای پیش‌بینی بیماری قلبی از J_{48}، بیز ساده و رده‌بندی با استفاده از دسته‌بندی پیشنهاد دادند. یافته‌های آن‌ها نشان از برتری درخت تصمیم، با دقت ۹۹/۲ درصد داشت (۲۷). الگوریتم Bagging، توسط Tu و همکاران برای تشخیص بیماری قلبی پیشنهاد شد. الگوریتم Bagging به دقت ۸۶/۶۴ درصد رسیده است (۲۸). Kumar و Sahoo الگوریتمی با ترکیب بیز ساده و الگوریتم ژنتیک به منظور بهبود رده‌بندی بیماری قلبی پیشنهاد داده‌اند که دقت ۱۰۰ درصدی را به دست آورد (۲۹). Fei Sheng ترکیبی از SVM (Support vector machine) و بهینه‌سازی ازدحام ذرات را برای تشخیص آریتمی قلبی استفاده نمود که منجر به نتایج بهتر نسبت به روشی مانند شبکه عصبی شد (۳۰). Subanya و Rajalaxmi مدلی از SVM و کلونی زنبور عسل را برای رده‌بندی بیماری قلبی پیشنهاد کردند که دقت ۸۶/۷۶ درصدی را نشان داد (۶). در مطالعه Polat و همکاران، تکنیک سیستم تشخیص ایمنی مصنوعی با مکانیزم تخصیص منابع فازی برای تشخیص بیماری قلبی پیشنهاد گردید که به دقت ۸۷ درصدی رسیده بود (۳۱). Jen و همکاران سیستمی با استفاده از نزدیک‌ترین همسایگان K-NN (K-Nearest neighbors) برای داده‌های بیماری پیشنهاد داده‌اند که دقت آن برای دیابت ۸۶/۴۹ درصد و بهتر از بیماری قلبی بود (۳۲). مطالعات انجام شده در زمینه بیماری قلبی، در جدول ۱ نشان داده شده است.

در زمینه سرطان سینه، شبکه عصبی از الگوریتم‌های پرکاربرد بود. Delen و همکاران برای پیش‌بینی سرطان سینه، از شبکه عصبی، درخت تصمیم و رگرسیون لجستیک استفاده نمودند. درخت تصمیم با دقت ۹۳/۶ درصد، نسبت به شبکه عصبی عملکرد بهتری داشته است (۳۳). Karabatak و Ince یک سیستم تشخیص سرطان سینه شامل شبکه عصبی و قوانین انجمنی را پیشنهاد دادند. ترکیب شبکه عصبی با دو مدل از قوانین انجمنی، به دقت‌های بالای ۹۵ درصد رسیده است (۳۴). Seera و Lim Chee مدل ترکیبی از شبکه عصبی فازی، درخت رده‌بندی، رگرسیون و جنگل تصادفی برای بیماری‌های مختلف را توسعه دادند که این مدل به دقت بالاتری برای سرطان سینه نسبت به دیابت رسیده است (۳۵). Mohapatra و همکاران یک روش شبکه عصبی پیش‌خور و جستجوی فاخته، برای رده‌بندی بیماری‌های مختلف را پیشنهاد کردند. دقت این مدل برای سرطان سینه ۹۷/۷۷ درصد و برای دیابت ۸۷/۵ درصد به دست آمد (۳۶).

Fan Chin و همکاران یک مدل ترکیبی از درخت تصمیم فازی و الگوریتم ژنتیک برای رده‌بندی داده‌های سرطان سینه و اختلالات کبدی پیشنهاد دادند که به دقت ۹۸/۴ درصد برای سرطان سینه رسیده است (۳۷). Onan از روش K-NN در ترکیب با Rough set، برای تشخیص سرطان سینه استفاده نمود که

دقت ۹۹/۷۱ درصدی داشت (۳۸). Polat و همکاران از سیستم تشخیص ایمنی مصنوعی و درخت تصمیم، برای تشخیص سرطان سینه استفاده کردند که دارای دقت ۹۸/۵۱ درصدی بود (۳۹). مطالعات بررسی شده برای سرطان سینه در جدول ۲ نشان داده شده است.

جدول ۱: مطالعات انجام شده در زمینه بیماری‌های قلبی (۳۲-۱۳، ۶)

الگوریتم رده‌بندی	دقت رده‌بندی (درصد)
شبکه عصبی پرسپترون چند لایه	بیماری مزمن قلبی (۸۲/۹۰)
شبکه عصبی	۸۹/۰۱
شبکه عصبی	۱۰۰/۰۰
بیز ساده	۹۴/۴۴
درخت تصمیم	۹۹/۶۲
شبکه عصبی پیش‌خور	۹۰/۵۴
شبکه عصبی	شبکه عصبی و LR (۷۸/۵۷)
	شبکه عصبی و MARS (۸۲/۱۴)
	شبکه عصبی و Rough set (۷۹/۵۰)
شبکه عصبی	۹۷/۴۰
شبکه عصبی	۸۵/۰۰
شبکه عصبی پرسپترون چندلایه	۷۶/۹۶
سیستم فازی	۷۶/۶۷
سیستم فازی	۸۴/۴۴
سیستم فازی	۹۴/۹۲
شبکه عصبی‌ترکیبی (شبکه عصبی فازی)	۸۶/۸۰
سیستم فازی	۷۸/۷۸
سیستم فازی	۸۱/۰۰
درخت تصمیم	۹۹/۲۰
بیز ساده	۹۶/۵۰
رده‌بندی با استفاده از دسته‌بندی	۸۸/۳۰
Bagging	۸۶/۶۴
درخت تصمیم	۷۸/۹۱
بیز ساده	۱۰۰/۰۰
ماشین بردار پشتیبان	۸۶/۷۶
ماشین بردار پشتیبان	۹۵/۶۲
سیستم تشخیص ایمنی مصنوعی	۸۷/۰۰
نزدیک‌ترین همسایگان	۷۹/۳۶

MARS: Multivariate adaptive regression splines; LR: Logistic regression

SVM، سیستم‌های فازی، K-NN و شبکه عصبی در مطالعات مرتبط با دیابت مورد توجه قرار گرفت. Calisir و Dogantekin ترکیب SVM با تحلیل افتراقی خطی و Wavelet morlet را برای تشخیص دیابت پیشنهاد دادند که دقت ۸۹/۷۴ درصدی را نشان داد (۴۰). Bekri و Govardhan یک مدل ماشین بردار پشتیبان با حداقل مربعات برای تعیین ابتلا به دیابت در گروه‌های مختلف خونی را ارایه کردند که به دقت ۹۵/۷۸ درصد رسید (۴۱).

جدول ۲: مطالعات انجام شده در زمینه سرطان سینه (۳۹-۳۳، ۲۶، ۲۵، ۲۰، ۱۹)

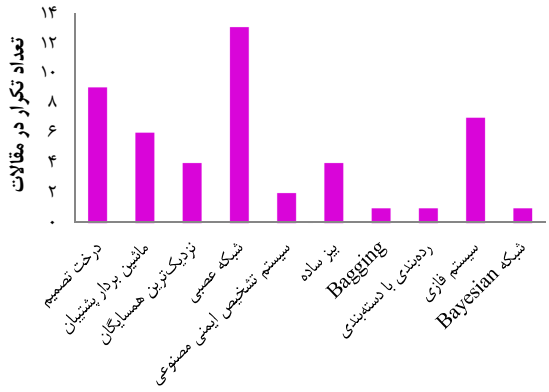
الگوریتم رده‌بندی	دقت رده‌بندی (درصد)
شبکه عصبی	۸۳/۷۵
شبکه عصبی پرسپترون چند لایه	۸۱/۶۹
سیستم فازی	۹۷/۴۰
سیستم فازی	۹۷/۹۰
شبکه عصبی	۹۳/۶۰
درخت تصمیم	۹۱/۲۰
شبکه عصبی	با قوانین انجمنی نوع ۱: ۹۷/۴۰
	با قوانین انجمنی نوع ۲: ۹۵/۶۰
شبکه عصبی	۹۸/۸۴
شبکه عصبی پیش‌خور	۹۷/۷۷
درخت تصمیم	۹۸/۴۰
نزدیک‌ترین همسایگان	۹۹/۷۱
سیستم تشخیص ایمنی مصنوعی	۹۸/۵۱

حیدری و همکاران ۵ روش رده‌بندی SVM، شبکه عصبی، درخت تصمیم، K-NN و شبکه Bayesian را روی مجموعه داده‌ای از بیماران مبتلا به دیابت نوع ۲ اعمال کردند. بیشترین دقت برابر با ۹۷/۴۴ درصد برای شبکه عصبی بود (۴۲). در پژوهشی، یک سیستم رده‌بندی فازی و کلونی مورچگان برای تشخیص دیابت ارایه گردید که دقت ۸۴/۲۴ درصدی را به دست آورد (۴۳). Beloufa و Chikh نیز روشی را برای تشخیص دیابت با استفاده از کلونی زنبور عسل و سیستم فازی پیشنهاد نمودند که استفاده از کلونی زنبور عسل مصنوعی منجر به دقت ۸۴/۲۱ درصد شد (۴۴). رمضان‌خوانی و همکاران مدلی برای شناسایی افراد با خطر پایین ابتلا به دیابت نوع ۲، با استفاده از درخت تصمیم ارایه دادند که دارای دقت ۹۰/۵ درصد می‌باشد (۴۵). مطالعات انجام شده در زمینه دیابت، در جدول ۳ نشان داده شده است.

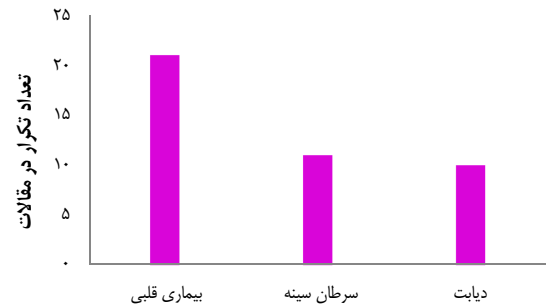
جدول ۳: مطالعات انجام شده در زمینه دیابت (۴۵-۴۰، ۳۶، ۳۵، ۳۲، ۲۴)

الگوریتم رده‌بندی	دقت رده‌بندی (درصد)
شبکه عصبی ترکیبی (شبکه عصبی فازی)	۸۴/۲۴
نزدیک‌ترین همسایگان	۸۶/۴۹
شبکه عصبی	۷۸/۳۹
شبکه عصبی پیش‌خور	۷۸/۵۰
ماشین بردار پشتبان	۸۹/۷۴
ماشین بردار پشتیبان	۹۵/۷۸
ماشین بردار پشتیبان	۸۱/۹۰
شبکه عصبی	۹۷/۴۴
درخت تصمیم	۹۵/۰۳
نزدیک‌ترین همسایگان	۹۰/۸۵
شبکه Bayesian	۹۱/۶۰
سیستم فازی	۸۴/۲۴
سیستم فازی	۸۴/۲۱
درخت تصمیم	۹۰/۵۰

توزیع الگوریتم‌های رده‌بندی، در شکل ۱ و فراوانی تکرار بیماری‌ها در شکل ۲ ارایه شده است.



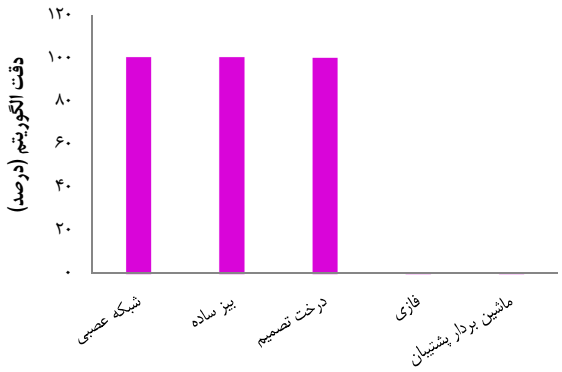
شکل ۱: توزیع الگوریتم‌های رده‌بندی در مطالعات پیشین



شکل ۲: توزیع بیماری‌های بررسی شده در مطالعات پیشین

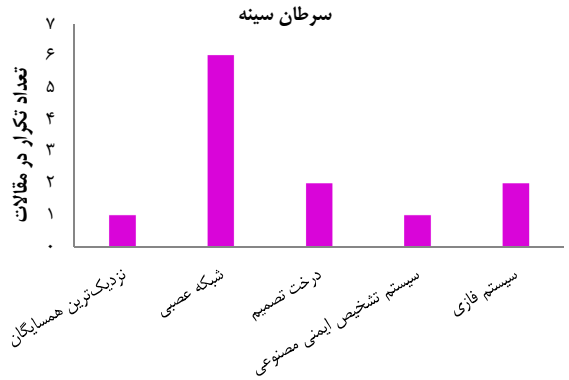
در این بخش، توزیع الگوریتم‌های رده‌بندی برای سه بیماری قلبی، سرطان سینه و دیابت ارایه شده است. در شکل‌های ۳ و ۴، به ترتیب تکرار الگوریتم‌های رده‌بندی و بالاترین دقت مهم‌ترین الگوریتم‌های رده‌بندی، برای بیماری قلبی نشان داده شده است.

شکل ۳: الگوریتم‌های مورد استفاده برای بیماری قلبی



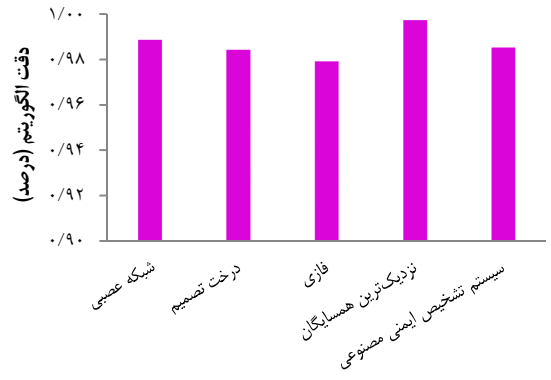
شکل ۴: بالاترین دقت مهم‌ترین الگوریتم‌های رده‌بندی برای بیماری قلبی

در شکل‌های ۵ و ۶، به ترتیب تکرار الگوریتم‌های رده‌بندی و بالاترین دقت مهم‌ترین الگوریتم‌های رده‌بندی، برای سرطان سینه ارایه شده است.



شکل ۵: الگوریتم‌های مورد استفاده برای سرطان سینه نزدیک‌ترین

در شکل‌های ۷ و ۸، به ترتیب تکرار الگوریتم‌های رده‌بندی و بالاترین دقت مهم‌ترین الگوریتم‌های رده‌بندی، برای دیابت ارایه شده است.



شکل ۶: بالاترین دقت مهم‌ترین الگوریتم‌های رده‌بندی برای سرطان سینه

شکل ۷: الگوریتم‌های مورد استفاده برای دیابت

شکل ۸: بالاترین دقت مهم‌ترین الگوریتم‌های رده‌بندی برای دیابت

بحث

از نتایج آمار توصیفی می‌توان دریافت که شبکه عصبی، درخت تصمیم و سیستم فازی پرکاربردترین الگوریتم‌ها، در مطالعات بررسی شده بود. سایر نتایج نشان داد که بیماری‌های قلبی، بیشترین تعداد تکرار را داشت. برای بیماری‌های قلبی، شبکه عصبی و سیستم‌های فازی از پرکاربردترین روش‌ها بود. در این زمینه، الگوریتم شبکه عصبی، بیز ساده و درخت تصمیم بیشترین دقت را داشت که می‌تواند برای به کارگیری در زمینه بیماری قلبی مناسب باشد. شبکه عصبی، دارای بیشترین کاربرد برای سرطان سینه بود. در زمینه سرطان سینه، الگوریتم‌های مورد استفاده در مقالات، همگی دقت مناسبی داشتند و بیشترین دقت به الگوریتم K-NN مربوط شد. در دیابت، شبکه عصبی بیشترین کاربرد و بالاترین دقت را به خود اختصاص داشت. همچنین، ماشین بردار پشتیبان و درخت تصمیم می‌تواند به عنوان الگوریتم‌های مناسب برای دیابت در نظر گرفته شود.

با توجه به بررسی‌ها، می‌توان استنباط کرد که شبکه عصبی، بیز ساده و درخت تصمیم الگوریتم‌های مناسبی برای بیماری‌های قلبی تلقی می‌شوند. در سرطان سینه، بر خلاف این که K-NN بهترین عملکرد را در میان الگوریتم‌ها داشت، اما نمی‌توان از دقت بالای شبکه عصبی، سیستم تشخیص ایمنی مصنوعی، درخت تصمیم و سیستم فازی چشم‌پوشی کرد. در دیابت، اختلاف قابل توجهی میان عملکرد الگوریتم‌ها مشاهده شد؛ به طوری که شبکه عصبی، SVM و درخت تصمیم را می‌توان به عنوان الگوریتم‌های مناسب در این بیماری در نظر گرفت. دقت حاصل از K-NN و سیستم فازی نشان داد که استفاده از این الگوریتم‌ها در دیابت می‌تواند منجر به نتایج مناسبی نگردد. به طور کلی، شبکه عصبی به دلیل قابلیت استفاده در مسایل پیچیده، کاربرد بیشتری در هر سه بیماری داشت که نشان از برتری این روش نسبت به روش‌های دیگر دارد. از محدودیت‌های مطالعه حاضر می‌توان به عدم دسترسی به برخی مقالات و در نتیجه تعداد مقالات کمتر از حد انتظار اشاره کرد.

نتیجه‌گیری

جمع‌بندی تکنیک‌های رده‌بندی متعدد استفاده شده برای استخراج الگوهای با معنی از داده‌های عظیم در حوزه سلامت، می‌تواند به محققان و پزشکان در جهت انتخاب مسیر مناسب جهت پژوهش‌های آینده یاری رساند. اگرچه، هیچ الگوریتم رده‌بندی به الزام بهترین الگوریتم برای یک بیماری مشخص نیست و در مورد هیچ الگوریتم داده‌کاوی نمی‌توان با قطعیت بیان داشت که آن الگوریتم منجر به بهترین نتایج می‌شود. به طور کلی، نتایج به دست آمده برای هر بیماری، می‌تواند منجر به دید روشنی جهت انتخاب الگوریتم‌ها در پژوهش‌های آینده شود. با توجه به این مطالعه، ممکن است که استفاده از الگوریتم‌های شبکه عصبی و بیز ساده در بیماری قلبی، K-NN برای سرطان سینه، شبکه عصبی و SVM در دیابت، برای مطالعات دیگر در زمینه این بیماری‌ها نیز باعث کسب نتایج مناسبی گردد.

پیشنهادها

با توجه به این که تشخیص بیماری‌ها از حساسیت زیادی برخوردار است و تشخیص نادرست، می‌تواند منجر به مواردی همچون مرگ شود، لازم است که از روش‌های داده‌کاوی با دقت بالا استفاده گردد. پیشنهاد می‌شود که در این زمینه، روش‌های ترکیبی داده‌کاوی مورد استفاده قرار گیرد که می‌تواند منجر به دقت بالاتری شود. همچنین، روش‌های تکاملی مانند الگوریتم ژنتیک در قالب رویکرد انتخاب ویژگی یا مقداردهی الگوریتم‌ها، می‌تواند دقت الگوریتم‌ها را افزایش دهد.

تشکر و قدردانی

بدین وسیله نویسندگان از جناب آقای دکتر عباس خسروی، به جهت ارایه نظرات سازنده و راهنمایی‌های ارزشمندشان تشکر و قدردانی به عمل می‌آورند.

References

1. World Health Organization. The top 10 causes of death [Online]. [cited 2016]; Available from: URL: http://www.who.int/mediacentre/factsheets/fs310/en/index2.html
2. World Health Organization. Cardiovascular diseases (CVDs) [Online]. [cited 2016]; Available from: URL: http://www.who.int/mediacentre/factsheets/fs317/en/
3. Daraei A, Hamidi H. Predicting myocardial infarction using data mining and a two stage feature selection method. Proceedings of the 1st International Conference on New Research Achievements in Electrical and Computer Engineering;

2016 May 13; Tehran, Iran. [In Persian].
4. World Health Organization. Cancer [Online]. [cited 2015]; Available from: URL: http://www.who.int/mediacentre/factsheets/fs297/en/
5. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 2006; 3(11): e442.
6. Subanya B, Rajalaxmi R. Feature selection using Artificial Bee Colony for cardiovascular disease classification. Proceedings of the International Conference on Electronics and Communication Systems; 2014 Feb 13-14; Coimbatore, India; 2014. p. 1-6.
7. Daraei A, Hamidi H. Review of classification algorithms for heart disease and liver disorders. Proceedings of the International Conference on Modern Research in Management and Industrial Engineering; 2015 Nov 16; Tehran, Iran. [In Persian].
8. Esfandiari N, Babavalian MR, Moghadam AME, Tabar VK. Knowledge discovery in medicine: Current issue and future trend. Expert Syst Appl 2014; 41(9): 4434-63.
9. Moghaddassi H, Hoseini A, Asadi F, Jahanbakhsh M. Application of data mining. Health Inf Manage 2012; 9(2): 297-304. [In Persian].
10. Vanaja S, Rameshkumar K. Performance analysis of classification algorithms on medical diagnoses-a survey. Journal of Computer Science 2015; 11(1): 30-52.
11. Ghazanfari M, Alizadeh S, Teymourpour B. Data mining and knowledge discovery. Tehran, Iran: Iran University of Science and Technology Publications; 2008. [In Persian].
12. Hamidi H, Daraie A. A new hybrid method for improving the performance of myocardial infarction prediction. Journal of Community Health Research 2016; 5(2): 5-6.
13. Yan H, Zheng J, Jiang Y, Peng C, Li Q. Development of a decision support system for heart disease diagnosis using multilayer perceptron. Proceedings of the International Symposium on Circuits and Systems; 2003 May 25-28; Bangkok, Thailand; 2003. p. 709-13.
14. Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl 2009; 36(4): 7675-80.
15. Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. Int J Comput Appl 2012; 47(10): 44-8.
16. Santhanam T, Ephzibah EP. Heart disease classification using PCA and feed forward neural networks. In: Prasath R, Kathirvalavakumar T, Editors. Mining Intelligence and Knowledge Exploration. Berlin, Germany: Springer International Publishing; 2013. p. 90-9.
17. Shao Yuehjen E, Hou Chia D, Chiu Chih C. Hybrid intelligent modeling schemes for heart disease classification. Appl Soft Comput 2014; 14(Part A): 47-52.
18. Abuhasel K, Iliyasu A, Fatichah C. A combined adaboost and newfm technique for medical data classification. In: Kim KJ, Joukov N, Editors. Information science and applications (ICISA). Berlin, Germany: Springer Science+Business Media; 2016. p. 801-9.
19. Abdel-Aal RE. GMDH-based feature ranking and selection for improved classification of medical data. J Biomed Inform 2005; 38(6): 456-68.
20. Dash T, Nayak S, Behera HS. Hybrid gravitational search and particle swarm based fuzzy MLP for medical data classification. In: Jain LC, Behera HS, Mandal JK, Mohapatra DP, Editors. Computational intelligence in data mining. Berlin, Germany: Springer; 2014. p. 35-43.
21. Dennis B, Muthukrishnan S. AGFS: Adaptive genetic fuzzy system for medical data classification. Appl Soft Comput 2014; 25: 242-52.
22. Lahsasna A, Ainon RN, Zainuddin R, Bulgiba A. Design of a fuzzy-based decision support system for coronary heart disease diagnosis. J Med Syst 2012; 36(5): 3293-306.
23. Jain LC, Patnaik S, Ichalkaranje N. Intelligent computing, communication and devices: proceedings of ICCD 2014. Berlin, Germany: Springer Science+Business Media; 2014. p. 415-27.
24. Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. Expert Syst Appl 2008; 35(1□2): 82-9.
25. Nguyen T, Khosravi A, Creighton D, Nahavandi S. Classification of healthcare data using genetic fuzzy logic system and wavelets. Expert Syst Appl 2015; 42(4): 2184-97.
26. Nguyen T, Khosravi A, Creighton D, Nahavandi S. Medical data classification using interval type-2 fuzzy logic system and wavelets. Appl Soft Comput 2015; 30: 812-22.
27. Anbarasi M, Anupriya E, Iyengar N. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. Int J Eng Sci Technol 2010; 2(10): 5370-6. [In Persian].
28. Tu M, Shin D, Shin D. Effective diagnosis of heart disease through bagging approach. In: Biomedical Engineering and Informatics. Proceedings of the 2nd International Conference on BioMedical Engineering and Informatics; 2009 Oct 17-19; Tianjin, China.
29. Kumar S, Sahoo G. Classification of heart disease using naive bayes and genetic algorithm. In: Jain LC, Behera HS, Mandal JK, Mohapatra DP, Editors. Computational intelligence in data mining. Berlin, Germany: Springer Science+Business Media; 2014. p. 269-82.
30. Fei Sheng W. Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine.

Expert Syst Appl 2010; 37(10): 6748-52.
31. Polat K, Sahan S, Gunes S. Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. Expert Syst Appl 2007; 32(2): 625-31.
32. Jen Chih H, Wang Chien C, Jiang Bernard C, Chu Yan H, Chen Ming S. Application of classification techniques on development an early-warning system for chronic illnesses. Expert Syst Appl 2012; 39(10): 8852-8.
33. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005; 34(2): 113-27.
34. Karabatak M, Ince MC. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl 2009; 36(2, Part 2): 3465-9.
35. Seera M, Lim Chee P. A hybrid intelligent system for medical data classification. Expert Syst Appl 2014; 41(5): 2239-49.
36. Mohapatra P, Chakravarty S, Dash PK. An improved cuckoo search based extreme learning machine for medical data classification. Swarm Evol Comput 2015; 24: 25-49.
37. Fan Chin Y, Chang Pei C, Lin Jyun J, Hsieh JC. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput 2011; 11(1): 632-44.
38. Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Syst Appl 2015; 42(20): 6844-52.
39. Polat K, Sahan S, Kodaz H, Gunes S. A new classification method for breast cancer diagnosis: feature selection artificial immune recognition system (FS-AIRS). In: Wang L, Chen K, Ong YS, Editors. Advances in natural computation. Berlin, Germany: Springer Science & Business Media, 2005. p. 830-8.
40. Calisir D, Dogantekin E. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Syst Appl 2011; 38(7): 8311-5.
41. Bekri FE, Govardhan A. OFW-ITS-LSSVM: weighted classification by LSSVM for diabetes diagnosis. International Journal of Advanced Computer Science & Application 2012; 3(3): 84-93.
42. Heydari M, Teimouri M, Heshmati Z, Alavinia SM. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int J Diabetes Dev Ctries 2016; 36(2): 167-73.
43. Ganji Mostafa F, Abadeh Mohammad S. A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Syst Appl 2011; 38(12): 14650-9.
44. Beloufa F, Chikh MA. Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Comput Methods Programs Biomed 2013; 112(1): 92-103.
45. Ramezankhani A, Pournik O, Shahrabi J, Khalili D, Azizi F, Hadaegh F. Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study. Diabetes Res Clin Pract 2014; 105(3): 391-8.

Assessment of Classification Algorithms in the Prediction of Healthcare Data: A Literature Review

Hojatollah Hamidi[1], Atefeh Daraei[2]

Review Article

Abstract

Data mining, as a tool for extracting useful information from large data sets, has been one of the areas of interest to researchers in the field of health. Classification is a learning function by which data is mapped to one of the predefined categories. According to World Health Organization (WHO), heart disease, renal disease, diabetes and cancer have been the cause of 68% of all deaths in 2012. The aim of this research was to study various types of classification algorithms and the results of previous researches in this regard in the field of health. In this narrative review, studies on heart disease, breast cancer, and diabetes, published from 2003 to 2015, were investigated. The keywords of “data mining”, “classification”, “health”, “heart disease”, “diabetes”, and “breast cancer” were searched in ScienceDirect, Elsevier, Springer, and IEEE databases. In addition, references and citations of each retrieved article were collected. After the elimination of unsuitable studies, 34 articles were selected. Literature review showed that frequency of use of neural network algorithm was the highest for all three diseases. Neural network and Naïve Bayes for heart disease, K-nearest neighbors for breast cancer, and neural network for diabetes had the highest accuracy. In general, it can be concluded that although no algorithm can be consider the best algorithm for each disease with certainty, determining the best algorithm for each disease could be useful for future studies.

Keywords: Classification; Data Mining; Neural Network; Heart Disease

Received: 2 Sep, 2015 Accepted: 4 Jul, 2016

Citation: Hamidi H, Daraei A. **Assessment of Classification Algorithms in the Prediction of Healthcare Data: A Literature Review.** Health Inf Manage 2016; 13(3): 235-42

Article resulted from an independent research without financial support.

1- Assistant Professor, Computer Engineering, Department of Information Technology, School of Industrial Engineering, Khajeh Nasir Toosi University of Technology, Tehran, Iran (Corresponding Author) Email: h_hamidi@kntu.ac.ir

2- MSc Student, Information Technology, School of Industrial Engineering, Khajeh Nasir Toosi University of Technology, Tehran, Iran