مقایسه کارایی مدل‌های پیش‌بینی عود مجدد سرطان پستان مبتنی بر تکنیک‌های داده‌کاوی

الهام میرزاکاظمی[1]، محمد غمگسار ناصری[2]

مقاله پژوهشی

چکیده

مقدمه: پس از به کارگیری روش‌های درمان سرطان پستان، احتمال عود مجدد بیماری وجود دارد. هدف از انجام پژوهش حاضر، به کارگیری تکنیک‌های داده‌کاوی به منظور ارایه مدل‌های پیش‌بینی عود مجدد سرطان پستان بود.

روش بررسی: در این مطالعه توصیفی، از ۱۸ ویژگی مربوط به ۸۰۹ بیمار مبتلا به سرطان پستان استفاده شد. برای ایجاد مدل پیش‌بینی عود مجدد سرطان پستان در مرحله پیش‌پردازش مجموعه داده، از الگوریتم‌های بیشینه‌سازی امید ریاضی EM (Expectation Maximization) و درخت تصمیم دسته‌بندی و رگرسیون C and R (Classification and Regression) استفاده گردید. سپس در مرحله یادگیری مدل، پنج الگوریتم داده‌کاوی شامل شبکه‌های عصبی، درخت تصمیم C and R، درخت تصمیم C۵، شبکه Bayes و ماشین بردار پشتیبان SVM (Support Vector Machine) به کار گرفته شد. در نهایت، جهت ارزیابی کارایی تکنیک‌های مورد استفاده، الگوریتم درخت تصمیم J۴۸ با K-Fold برابر ۱۰ و روش‌های آنالیز داده‌ها مورد استفاده قرار گرفت.

یافته‌ها: دقت الگوریتم‌های EM و C and R در مرحله پیش‌پردازش داده‌ها به ترتیب ۰/۶۴۱ و ۰/۴۲۰ بود. دقت پنج الگوریتم به کار رفته در مرحله یادگیری مدل نیز به ترتیب ۰/۸۵۸، ۰/۸۶۵، ۰/۸۷۰، ۰/۸۸۳ و ۰/۹۹۸ به دست آمد.

نتیجه‌گیری: مدلی که در مرحله پیش‌پردازش از الگوریتم EM و در مرحله یادگیری از الگوریتم SVM بهره می‌گیرد، کارایی بالاتری نسبت به سایر مدل‌های ایجاد شده دارد.

واژه‌های کلیدی: داده‌کاوی؛ عود مجدد؛ سرطان پستان؛ الگوریتم

دریافت مقاله: ۱۳۹۵/۸/۱۷ پذیرش مقاله: ۱۳۹۶/۶/۲۹

ارجاع: میرزاکاظمی الهام، غمگسار ناصری محمد. **مقایسه کارایی مدل‌های پیش‌بینی عود مجدد سرطان پستان مبتنی بر تکنیک‌های داده‌کاوی.** مدیریت اطلاعات سلامت ۱۳۹۶؛ ۱۴ (۴): ۱۴۹-۱۴۴

مقدمه

سرطان پستان، رایج‌ترین سرطان در میان زنان می‌باشد. بر اساس آمارهای سازمان جهانی بهداشت، از هر ۸ تا ۱۰ زن، یک نفر به سرطان پستان مبتلا می‌گردد (۱). این بیماری هر ساله باعث مرگ و میر فراوانی در بین زنان و مردان می‌شود. با وجود پیشرفت‌های بسیاری که در زمینه تشخیص زودهنگام و درمان مناسب این بیماری صورت گرفته است، هنوز هم سرطان سینه سردسته علل مرگ به علت سرطان در بین زنان می‌باشد (۲). الگوی اپیدمیولوژی سرطان پستان در ایران مشابه با کشورهای منطقه مدیترانه شرقی و سایر کشورهای در حال توسعه می‌باشد و روند بروز آن در سال‌های گذشته دچار تغییر شده است. در ایران بر اساس آمار مرکز مدیریت بیماری‌های وزارت بهداشت، سرطان پستان از نظر بروز بین تمام سرطان‌ها در زنان، همچنان در رتبه اول قرار دارد و با میزان بروز تعدیل شده سنی ASR (Age Standardized Rate)، ۲۷/۱۵ و ۶۹۷۶ مورد در سال ۱۳۸۶، بالاترین موارد بروز سرطان می‌باشد. متأسفانه سن بروز سرطان پستان در زنان ایرانی پایین‌تر از میانگین سن جهانی است (۳).

پیش‌بینی عود مجدد سرطان پستان، یکی از پرطرفدارترین اقدامات انجام شده برای توسعه رویکردهای داده‌کاوی می‌باشد. بررسی‌ها و مطالعات گوناگونی در زمینه مشکلات ناشی از پیش‌بینی بقای بیماران مبتلا به سرطان پستان با استفاده از روش‌های آماری و شبکه‌های عصبی مصنوعی ANNs (Artificial Neural Networks) صورت گرفته است، اما فقط تعداد کمی از مطالعات حوزه پزشکی در زمینه عود سرطان با استفاده از روش‌های داده‌کاوی می‌باشد. البته تشخیص و پیش‌بینی انواع بیماری‌ها با استفاده از تکنیک‌های داده‌کاوی امکان‌پذیر است. داده‌کاوی در پزشکی به فرایند استخراج اطلاعات معتبر از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های پزشکی و استفاده از آن جهت پیش‌بینی، تشخیص و کمک به درمان بیماری گفته می‌شود. کشف الگوهای مفید بین بیماری و علایم بالینی و آزمایشگاهی بیمار، از جمله کاربردهای داده‌کاوی در پزشکی به شمار می‌رود. تکنیک‌های داده‌کاوی می‌تواند به کاهش تعداد پاسخ‌ها و نتایج مثبت و منفی کاذب در تصمیم‌گیری پزشکان کمک نماید. به کمک داده‌کاوی، پژوهشگران می‌توانند الگوها و روابط بین تعداد زیادی از متغیرها را شناسایی کنند. از سوی دیگر، پیش‌بینی نتایج حاصل از یک بیماری با استفاده از ذخایر اطلاعاتی موجود در پایگاه‌های داده برای آن‌ها امکان‌پذیر شده است (۷-۴).

طلوعی اشلقی و همکاران در مورد توسعه مدل‌های پیش‌بینی عود مجدد

مقاله با حمایت مؤسسه آموزش عالی جهاد دانشگاهی رشت انجام شده است.

۱- مربی، مهندسی نرم‌افزار، گروه کامپیوتر و برق، مؤسسه آموزش عالی جهاد دانشگاهی رشت، رشت، ایران (نویسنده مسؤول)

Email: e_mkazemi@yahoo.com

۲- مربی، ریاضی کاربردی، گروه کامپیوتر و برق، مؤسسه آموزش عالی جهاد دانشگاهی رشت، رشت، ایران

سرطان پستان، از الگوریتم‌های درخت تصمیم‌گیری C۵، ماشین بردار پشتیبان (Support Vector Machine) SVM و ANNs استفاده نمودند که هر سه روش دارای دقت بالایی بود، اما روش SVM بالاترین دقت و کمترین خطا را نسبت به سایر روش‌ها داشت. همچنین، روش C۵ در میان این سه راهبرد، نسبت به بقیه دارای دقت پایین‌تری بود (۷).

پژوهش‌های متعددی در علوم پزشکی با استفاده از تکنیک‌های داده‌کاوی برای ارایه مدل‌های پیش‌بینی صورت گرفته است. به عنوان مثال، Ravi Kumar و همکاران مطالعه‌ای را بر روی پایگاه داده معروف ویسکانسین انجام دادند تا مدل طبقه‌بندی دقیقی را برای پیش‌بینی سرطان پستان ارایه نمایند. در تحقیق آن‌ها، SVM نسبت به پنج روش دیگر به کار رفته، دارای دقت بالاتری بود (۸). همچنین، Delen و همکاران تکنیک‌های درخت تصمیم، رگرسیون لجستیک و ANNs را بر روی ۲۰ متغیر داده‌ای موجود در پایگاه داده (SEER) Surveillance, Epidemiology, and End Results انجام دادند و بین این سه تکنیک، مدل پیش‌بینی ارایه شده توسط الگوریتم درخت تصمیم C۵ را دارای دقت بالاتری برآورد نمودند (۹). Choi و همکاران در تحقیق خود، از ۹ متغیر پایگاه داده SEER به عنوان ورودی استفاده کردند و دقت الگوریتم‌های شبکه عصبی، شبکه Bayesian و شبکه ترکیبی را ارزیابی نمودند. تکنیک شبکه عصبی در واقع از ترکیب دو الگوریتم شبکه Bayesian و ANN ایجاد شده است که نسبت به روش‌های دیگر، از دقت بالاتری در ارایه مدل پیش‌بینی برخوردار می‌باشد (۱۰). Shajahaan و همکاران نیز پژوهشی را در مورد مدل‌های پیش‌بینی بر روی داده‌های سرطان پستان انجام دادند و کارایی و دقت چندین تکنیک داده‌کاوی را ارزیابی نمودند. روش‌های به کار گرفته شده شامل درخت تصادفی، ۳ Iterative Dichotomiser (ID۳)، درخت تصمیم دسته‌بندی و رگرسیون C and R (Classification and Regression)، C4.5 و Naive Bayes بود. از بین این روش‌ها، الگوریتم درخت تصادفی نسبت به سایر روش‌ها دقت بالاتری در ارایه مدل پیش‌بینی سرطان پستان داشت (۱۱).

Subasini و همکاران مطالعاتی را در زمینه مدل‌های پیش‌بینی سرطان پستان بر روی الگوریتم‌های C۵، ID۳، Apriori، C4.5 و Naive Bayes انجام دادند که از بین این روش‌ها، الگوریتم C۵ نسبت به سایر روش‌ها دقت بالاتری در ارایه مدل مورد نظر داشت (۱۲). پژوهش Kulkarni و Bhagwat نیز در همین زمینه انجام گرفت. آن‌ها روش‌های مختلف طبقه‌بندی را برای ارزیابی دقت مدل پیش‌بینی عود مجدد سرطان پستان به کار گرفتند. این تکنیک‌ها شامل Naive Bayes، IBk، شبکه Bayes، Kstar، Naive Bayes Multinomial Text، InputMappedClassifier، One Rule (OneR)، Decision Table، Partial Decision Tree (PART)، ZeroR، Decision Stump، Logistic Model Tree (LMT)، JRip، Locally weighted Learning (LWL)، RandomForest، J۴۸، Ethernet Private Tree (EP Tree) و K-Nearest Neighbors Hamming Distance (KNN-Hamming Distance) بود که از میان آن‌ها، الگوریتم‌های JRip و KNN-Hamming Distance بالاترین دقت را نسبت به سایر روش‌ها داشت (۱۳). کیانی و آتشی مدل پیش‌آگهی مبتنی بر داده‌کاوی را برای پیش‌بینی عود مجدد سرطان پستان طراحی نمودند. آن‌ها در روش مطالعه خود، به دلیل وجود مقادیر تهی در مجموعه داده، از الگوریتم بیشینه‌سازی امید ریاضی EM (Expectation Maximization) به عنوان یکی از فازهای پیش‌پردازش داده‌ها استفاده کردند و سپس با استفاده از الگوریتم J۴۸، یک مدل پیش‌آگهی عود مجدد سرطان پستان را ارایه نمودند (۱۴).

با توجه به این که احتمال عود مجدد سرطان پستان پس از درمان این بیماری همواره وجود دارد و از طرف دیگر، سرطان پستان دومین سرطان شایع در بین زنان می‌باشد، ارایه مدل‌های پیش‌بینی عود مجدد سرطان پستان که بتواند در تعیین روند درمان به پزشک و بیمار کمک نماید، از اهمیت ویژه‌ای برخوردار است. از سوی دیگر، داده‌کاوی مجموعه روش‌های کاربردی برای ارایه مدل‌های پیش‌بینی بیماری‌ها دارد. بنابراین، در تحقیق حاضر ضمن بررسی مدل‌های پیش‌بینی مبتنی بر دسته‌بندی داده‌های سرطان پستان، مدلی با دقت بالاتر ارایه گردید.

روش بررسی

در این مطالعه توصیفی، داده‌ها از مرکز تحقیقات سرطان دانشگاه شهید بهشتی دریافت شد. داده‌ها مربوط به اطلاعات بایگانی شده ۸۰۹ بیمار مبتلا به سرطان پستان بود که طی سال‌های ۱۳۸۰ تا ۱۳۸۸ به این مرکز مراجعه کرده بودند (۱۴). پیش‌پردازش داده‌ها شامل انتخاب زیرمجموعه ویژگی‌های مرتبط و گسسته‌سازی مجموعه داده انجام شد. ابتدا برخی از ویژگی‌های غیر مرتبط حذف و ۱۸ ویژگی مرتبط برای ایجاد مدل آماده‌سازی گردید و سپس به منظور تبدیل ویژگی‌های بازه‌ای و نرخی به ویژگی‌های گسسته، مرحله گسسته‌سازی داده‌ها انجام شد. جدول ۱ عناوین مربوط به ویژگی‌های به کار رفته در پژوهش و دامنه مقادیر مربوط به هر ویژگی را نشان می‌دهد.

جدول ۱: عناوین ویژگی‌های استفاده شده در ایجاد مدل پیش‌بینی بعد از اعمال مرحله پیش‌پردازش

نام ویژگی	دامنه مقایر	نام ویژگی	دامنه مقایر
درجه پاتولوژی	{۱ و ۲ و ۳ و ۴}	تعداد غدد لنفاوی برداشته شده زیر بغل	متغیر در افراد
استروژن رسپتور	{۰ و ۱}	اندازه تومور اولیه (سانتی‌متر)	متغیر در افراد
پروژسترون رسپتور	{۰ و ۱}	سابقه جراحی	{۱ و ۲ و ۳}
P۵۳	{۰ و ۱}	میزان تحصیلات	{۰ و ۱ و ۲ و ۳ و ۴ و ۵}
HER-۹	{۰ و ۱}	سابقه فامیلی سرطان	{۰ و ۱ و ۲ و ۴}
مرحله بیماری	{۱ و ۲ و ۳ و ۴ و ۵}	وضعیت قاعدگی	{۱ و ۲ و ۳}
مرحله کلینیکی	{۱ و ۲ و ۳ و ۴ و ۵}	وضعیت تأهل	{۰۱ و ۲ و ۳ و ۴}
عود مجدد سرطان پستان	{۰ و ۱}	سن	متغیر در افراد
تعداد غدد لنفاوی مثبت	متغیر در افراد	مرحله کلینیکی	{۰ و ۱}

HER: Human Epidermal Receptor

جدول ۲: مقایسه کارایی الگوریتم‌های مورد استفاده در مرحله پیش‌پردازش داده‌ها

معیار ارزیابی / الگوریتم مرحله پیش‌پردازش	TPR	FPR	Precision	Recall	F-Measure	MCC	Roc Area	PRC Area
الگوریتم EM	۰/۸۲۷	۰/۸۳۱	۰/۶۴۱	۰/۸۴۴	۰/۸۴۷	۰/۸۴۴	۰/۲۲۶	۰/۸۴۷
الگوریتم درخت تصمیم C and R	۰/۷۴۴	۰/۷۴۹	۰/۴۲۰	۰/۷۳۸	۰/۷۵۲	۰/۷۴۵	۰/۳۷۵	۰/۷۵۲

C and R: Classification and Regression; PRC Area: Precision-Recall Area; MCC: Matthews Correlation Coefficient; TPR: True Positive Rate; FPR: False Positive Rate; EM: Expectation Maximization

ویژگی عود مجدد سرطان پستان به عنوان متغیر هدف انتخاب شد که دارای دو مقدار صفر و یک و به ترتیب نشان دهنده عدم عود مجدد و عود مجدد بیماری بود. با توجه به این که در رکوردهای اطلاعاتی مربوط به بیماران، تعدادی از فیلدها مقادیر تهی و گمشده داشت، از دو روش برای تخمین آن‌ها استفاده گردید.

روش اول با استفاده از الگوریتم EM در نرم‌افزار SPSS نسخه ۲۴ (version 24, IBM Corporation, Armonk, NY) (۱۴) و روش دوم با استفاده از الگوریتم C and R در نرم‌افزار Clementine نسخه ۱۲ اجرا شد. برای ارزیابی کارایی دو روش مذکور، از الگوریتم درخت J۴۸ و روش K-Fold با K = ۱۰ استفاده گردید. این الگوریتم در نرم‌افزار Weka 3.7 اجرا شد (۱۴).

بعد از انجام مراحل پیش‌پردازش و آماده‌سازی داده‌ها و ارزیابی کارایی دو روش مذکور، مرحله یادگیری برای ایجاد مدل پیش‌بینی عود مجدد سرطان پستان اجرا شد. بدین منظور، مجموعه داده‌ای که دارای دقت بالاتر در تخمین مقادیر تهی و گمشده بود، مورد استفاده قرار گرفت. همچنین، در این مرحله از پنج الگوریتم با ناظر داده‌کاوی در نرم‌افزار Clementine نسخه ۱۲ برای ایجاد مدل استفاده گردید. الگوریتم‌ها شامل شبکه‌های عصبی، درخت تصمیم C and R، درخت تصمیم C۵، شبکه Bayes و SVM بود. در نهایت، جهت ارزیابی کارایی تکنیک‌های مورد استفاده و تعیین مدلی با کارایی بالاتر، از روش آنالیز داده‌ها در نرم‌افزار Clementine نسخه ۱۲ و الگوریتم J۴۸ در نرم‌افزار Weka استفاده گردید. لازم به ذکر است که داده‌های به دست آمده، تنها در راستای هدف پژوهشی و پاسخ به سؤال مطالعه مورد استفاده قرار گرفت.

یافته‌ها

پس از اجرای دو الگوریتم EM و C and R در مرحله پیش‌پردازش داده‌ها در نرم‌افزار Weka، کارایی این دو الگوریتم ارزیابی گردید. نتایج مربوط به مقایسه کارایی دو الگوریتم در جدول ۲ ارایه شده است.

یافته‌ها نشان داد که کارایی الگوریتم EM نسبت به الگوریتم C and R در تخمین مقادیر تهی و گمشده برای مجموعه داده جمع‌آوری شده سرطان پستان، دقت بالاتری داشت. بنابراین، در پژوهش حاضر از مجموعه داده به دست آمده از اعمال الگوریتم EM، در مرحله بعدی (مرحله یادگیری مدل) استفاده گردید و مدل‌های پیش‌بینی عود مجدد سرطان پستان با استفاده از تکنیک‌های مختلف داده‌کاوی ایجاد شد.

جدول ۳ نتایج ارزیابی کارایی مدل‌های پیش‌بینی به دست آمده را نشان می‌دهد. دقت دسته‌بندی بیانگر آن است که مدل پیش‌بینی ارایه شده چند درصد از رکوردهای آزمایشی را به درستی دسته‌بندی کرده است. معیار نرخ خطای دسته‌بندی بر عکس معیار دقت دسته‌بندی می‌باشد. حساسیت مدل نشان دهنده تعداد افرادی است که بر اساس مدل، مستعد عود مجدد سرطان پستان شناخته شده‌اند نسبت به کل افرادی که واقعاً دچار عود مجدد شده‌اند. معیار ویژگی بیان کننده نسبت تعداد افرادی است که به درستی توسط مدل تشخیص داده شده‌اند که دچار عود مجدد سرطان پستان نمی‌شوند نسبت به تعداد کل افرادی که واقعاً دچار عود مجدد نشده‌اند. معیار صحت نشانگر تعداد تشخیص‌های صحیح توسط مدل نسبت به کل تعداد رکوردهای موجود می‌باشد. معیارهای مثبت کاذب و منفی کاذب نیز درصد تشخیص‌های نادرست با استفاده از مدل را نشان می‌دهد. مثبت کاذب یعنی نسبت مواردی که به طور نادرست توسط مدل مستعد عود مجدد سرطان پستان تشخیص داده شده‌اند و در واقعیت فرد دچار عود مجدد نشده و منفی کاذب به معنای نسبت مواردی که به طور نادرست مستعد عدم عود مجدد سرطان پستان تشخیص داده شده‌اند و در حقیقت فرد دچار عود مجدد شده است. یافته‌ها نشان داد که کارایی الگوریتم SVM نسبت به سایر الگوریتم‌های مورد بررسی بالاتر بود. معیار مثبت کاذب برابر با صفر و منفی کاذب نیز نزدیک به صفر بود و این یافته اهمیت بالایی در نحوه ادامه روند درمان توسط پزشک خواهد داشت.

جدول ۳: مقایسه کارایی الگوریتم‌های دسته‌بندی در ایجاد مدل پیشگویی عود مجدد سرطان پستان

معیار ارزیابی / الگوریتم مرحله یادگیری مدل	دقت دسته‌بندی	نرخ خطای دسته‌بندی	حساسیت	ویژگی	صحت	مثبت کاذب	منفی کاذب
شبکه عصبی	۰/۸۵۸	۰/۱۴۲	۰/۷۶۲	۰/۹۰۳	۰/۸۵۸	۰/۲۱۱	۰/۱۱۱
درخت تصمیم C and R	۰/۸۶۵	۰/۱۳۵	۰/۸۱۵	۰/۸۸۹	۰/۸۶۵	۰/۲۲۳	۰/۰۹۰
درخت تصمیم C۵	۰/۸۷۰	۰/۱۳۰	۰/۶۶۹	۰/۹۶۵	۰/۸۷۰	۰/۰۹۸	۰/۱۴۰
شبکه Bayes	۰/۸۸۳	۰/۱۱۷	۰/۷۶۵	۰/۹۳۸	۰/۸۸۳	۰/۱۴۶	۰/۱۰۶
SVM	۰/۹۹۸	۰/۰۰۲	۰/۹۹۲	۱/۰۰۰	۰/۹۹۸	۰	۰/۰۰۴

C and R: Classification and Regression; SVM: Support Vector Machine

بحث

در این پژوهش با به کارگیری تکنیک‌های داده‌کاوی بر روی داده‌های جمع‌آوری شده مرکز تحقیقات سرطان دانشگاه شهید بهشتی، مدل پیش‌بینی عود مجدد سرطان پستان ایجاد شد. این مدل به پزشک کمک می‌کند که قبل از گسترش سرطان به سایر قسمت‌های بدن، پیش‌بینی درستی داشته باشد و اقدامات لازم را برای بیمار در زمان مناسب انجام دهد. البته نکته بسیار مهم، دقت مطلوب و قابل قبول مدل‌های پیش‌بینی می‌باشد. کارایی مدل ارایه شده در مطالعه حاضر از نظر معیارهای ارزیابی مدل‌های پیش‌بینی همچون دقت، حساسیت و ویژگی در مقایسه با نتایج تحقیق طلوعی اشلقی و همکاران (۷) دارای مقدار بالاتری بود. این مدل از نظر معیار دقت نیز در مقایسه با الگوریتم‌های به کار رفته توسط Kulkarni و Bhagwat (۱۳)، مقدار بیشتری را نشان داد. همچنین، مدل ارایه شده در بررسی حاضر از نظر معیارهای حساسیت، ویژگی، دقت، صحت، مثبت کاذب و منفی کاذب در مقایسه با مدل ارایه شده توسط کیانی و آتشی (۱۴) کارایی بالاتری داشت.

از طرف دیگر، معیارهای مثبت کاذب و منفی کاذب نیز در ارزیابی مدل‌های پیش‌بینی پزشکی از اهمیت ویژه‌ای برخوردار است. این معیارها در پژوهش کیانی و آتشی به ترتیب ۴۶ و ۱۴ درصد گزارش گردید (۱۴). معیار مثبت کاذب نشان دهنده این است که فرد به اشتباه بیمار شناسایی شده است و این امر می‌تواند برای فرد مشکلات روحی به دنبال داشته باشد. مقدار این معیار در مدل ارایه شده مطالعه حاضر، صفر بود. معیار منفی کاذب در مدل‌های پیش‌بینی حوزه پزشکی بدین معنی است که فرد به اشتباه سالم تشخیص داده می‌شود و به طور قطع ادامه روند درمان تحت تأثیر قرار می‌گیرد و می‌تواند عواقب خطرناکی را به همراه داشته باشد. مقدار این معیار در مدل مطالعه حاضر، ۰/۴ درصد به دست آمد که مقدار نزدیک به صفر این معیار، میزان اعتماد به مدل ارایه شده در پژوهش را نشان می‌دهد. بنابراین، مدل مذکور می‌تواند توسط پزشک در پیش‌بینی عود مجدد بیماری با ضریب اطمینان به نسبت بالایی مورد استفاده قرار گیرد.

تحقیق حاضر محدودیت‌هایی داشت و می‌تواند در تحقیقات آینده مورد توجه قرار گیرد که از آن جمله می‌توان به تعداد محدود ویژگی‌ها، محدودیت جغرافیایی محل جمع‌آوری داده‌ها و وجود مقادیر تهی و گمشده در مجموعه داده اشاره نمود.

نتیجه‌گیری

استفاده از ابزارها و مکانیزم‌های دقیق پیش‌بینی بیماری در کنار تجربه پزشکان، می‌تواند نقش مؤثری در تشخیص صحیح بیماری و انتخاب روند درمان داشته باشد. این مسأله میزان اعتماد به تشخیص صحیح بیماری را هم برای خود پزشک و هم برای بیمار به نحو مطلوبی افزایش می‌دهد. از طرف دیگر، نحوه تشخیص بیماری، تعیین کننده اقدامات بعدی برای بیمار خواهد بود. بنابراین، چنانچه از دقت پایینی برخوردار باشد، بقای فرد را تحت تأثیر قرار می‌دهد و به همین علت دقت مطلوب و قابل قبول مدل‌های پیش‌بینی، اهمیت ویژه‌ای دارد. در واقع، مدلی که با بررسی معیارهای مختلف ارزیابی گردد و دقت و حساسیت آن در پیش‌بینی بالاتر باشد، به طور قطع قابل اعتمادتر خواهد بود.

پیشنهادها

استفاده از سایر ویژگی‌های مؤثر در عود مجدد سرطان پستان در مجموعه داده، می‌تواند در افزایش کارایی و دقت مدل‌های پیش‌بینی تأثیرگذار باشد بنابراین، پیشنهاد می‌شود این موضوع در مطالعات آینده مد نظر قرار گیرد. از طرف دیگر، چنانچه مجموعه داده‌های جمع‌آوری شده در صورت امکان بیشترین داده‌های واقعی و حداقل داده‌های از دست رفته را داشته باشد، دقت مدل ارایه شده مورد اعتماد بیشتری خواهد بود. بنابراین، توصیه می‌شود در مطالعات آینده تا حد امکان از داده‌های واقعی استفاده گردد و الگوریتم‌های پیش‌پردازش جهت تکمیل داده‌های از دست رفته استفاده نشود. نکته دیگر این که می‌توان از سایر تکنیک‌های داده‌کاوی برای طراحی مدل‌های پیش‌بینی استفاده نمود و کارایی آن‌ها را بررسی کرد. به منظور رفع محدودیت‌های جغرافیایی در جمع‌آوری داده‌ها، بهتر است داده‌های توزیع شده مربوط به مناطق مختلف جغرافیایی را متمرکز نمود تا مجموعه داده‌های مورد استفاده در پژوهش، شامل دامنه جغرافیایی وسیع‌تر و تعداد رکوردهای بیشتری باشد.

تشکر و قدردانی

بدین وسیله از آقای دکتر علیرضا آتشی، عضو گروه پژوهشی انفورماتیک سرطان سپاسگزاری می‌گردد. همچنین، از مرکز تحقیقات سرطان پستان جهاد دانشگاهی که در انجام این تحقیق همکاری نمودند، تشکر و قدردانی به عمل می‌آید.

References

1. Noori Daloii MR, Tabarestani S. Molecular genetics, diagnosis and treatment of breast cancer: Review article. J Sabzevar Univ Med Sci 2010; 17(2): 74-87. [In Persian].
2. Mirmalek SA, Elham Kani F. Clinical application of breast cancer biology review of literature. Iran J Surg 2009; (17): 1-6. [In Persian].
3. Roohparvarzade N, Ghadery M, Parsa A, Allahyary A. Prevalence of risk factors for breast cancer in women (20 to 69 Years old) in Isfahan 2012-2013. Iran J Breast Dis 2014; 1(1): 52-61. [In Persian].
4. Latif AM, Momeny M, Sarram R, Agha Sarram M, Pour Ahmadi A, Haj Ebrahimi Z. Using data mining and genetic algorithm for diagnosis of breast cancer. Iran J Breast Dis 2016; 9(1): 45-56. [In Persian].
5. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory; 1992 Jul 27-29; Pittsburgh, Pennsylvania, USA. p. 144-52.
6. American Cancer Society. Breast cancer facts and figures 2015-2016. Atlanta, GA: American Cancer Society, Inc; 2015.
7. Toluei Ashlaghi A, Poorebrahimi A, Ebrahimi M, Ghasem Ahmad L. Using data mining techniques for prediction breast cancer recurrence. Iran J Breast Dis 2013; 5(4): 23-34. [In Persian].
8. Ravi Kumar G, Ramachandra GA, Nagamani K. An efficient prediction of breast cancer data using data mining techniques.

International Journal of Innovations in Engineering and Technology 2013; 2(4): 139-44.
9. 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.
10. Choi JP, Han TH, Park RW. A hybrid bayesian network model for predicting breast cancer prognosis. J Korean Soc Med Inform 2009; 15(1): 49-57.
11. Shajahaan S, Shanthi S, Mano Chitra V. Application of data mining techniques to model breast cancer data. International Journal of Emerging Technology and Advanced Engineering 2013; 3(11): 362-9.
12. Subasini A, Abubacker NF, Rekha C. Analysis of classifier to improve medical diagnosis for breast cancer detection using data mining techniques. Int J Advanced Networking and Applications 2014; 5(6): 2117-22.
13. Kulkarni S, Bhagwat M. Predicting breast cancer recurrence using data mining techniques. Int J Comput Appl 2015; 122(23): 26-31.
14. Kiani B, Atashi A. A prognostic model based on data mining techniques to predict breast cancer recurrence. Journal of Health and Biomedical Informatics 2014; 1(1): 26-31. [In Persian].

Comparing the Functionality of Predicting Models for Breast Cancer Recurrence Based on Data Mining Techniques

Elham Mirzakazemi[1], Mohammad Ghamgosar-Naseri[2]

Original Article

Abstract

Introduction: After applying breast cancer treatment methods, there is a possibility of recurrence of the disease. The aim of the present study was using data mining techniques in order to provide predicting models for breast cancer recurrence.

Methods: 18 features of 809 patients were used in the current descriptive study. The study consisted of two phases, preprocessing phase and model learning. Expectation Maximization (EM) and Classification and Regression (C and R) were used for the analysis of the first phase. In order to analyze the second phase, the five algorithm model including Neural Network, C and R, the decision tree algorithm C5.0, Bayes Net, and Support Vector Machine (SVM) was used.

Results: The accuracy of the EM and C and R algorithms was 0.641 and 0.420, respectively, in the preprocessing phase. The accuracy of Neural Network, C and R, the decision tree algorithm C5.0, Bayes Net, and SVM algorithms was 0.858, 0.865, 0.870, 0.883, and 0.998, respectively, for the model learning phase.

Conclusion: According to the findings, the model with the application of EM algorithm in the first phase and SVM algorithm in the second phase had the highest functionality. It was also important in determining the treatment process.

Keywords: Data Mining; Recurrence; Breast Cancer; Algorithm

Received: 07 Nov., 2016 Accepted: 20 Sep., 2017

Citation: Mirzakazemi E, Ghamgosar-Naseri M. **Comparing the Functionality of Predicting Models for Breast Cancer Recurrence Based on Data Mining Techniques.** Health Inf Manage 2017; 14(4): 144-9

Article funded by Institute of Higher Education, Rasht Academic Center for Education, Culture and Research (ACECR).

1- Lecturer, Computer Software Engineering, Department of Computer and Electrical, Institute of Higher Education, Rasht Academic Center for Education, Culture and Research (ACECR), Rasht, Iran (Corresponding Author) Email: e_mkazemi@yahoo.com

2- Lecturer, Applied Mathematics, Department of Computer and Electrical, Institute of Higher Education, Rasht Academic Center for Education, Culture and Research (ACECR), Rasht, Iran