Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

مجله تخصصی اپیدمیولوژی ایران؛ بهار ۱۳۹۷، دوره ۱۴، شماره ۱: صفحات ۲۹-۱۹.

مقاله پژوهشی

پیش‌بینی بقاء بیماران مبتلا به سرطان روده بزرگ در بخش پرتودرمانی بیمارستان نمازی شیراز با استفاده از روش‌های داده‌کاوی ماشین بردار پشتیبان و بگینگ

سوگند ستاره[۱]، میثاق ظهیری اصفهانی[۲]، محمد زارع بندامیری[۳]، احمد رئیسی[۴]، رضا عباسی[۵،۶]

[۱] دانشجوی دکتری تخصصی انفورماتیک پزشکی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

[۲] دانشجوی دکتری تخصصی مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع‌رسانی پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران

[۳] بخش آنکولوژی رادیولوژی، بیمارستان نمازی، دانشگاه علوم پزشکی شیراز، شیراز، ایران

[۴] دانشجوی کارشناسی ارشد انفورماتیک پزشکی، مرکز تحقیقات انفورماتیک پزشکی، موسسه آینده پژوهی، دانشگاه علوم پزشکی کرمان، کرمان، ایران

[۵] دانشجوی دکتری تخصصی مدیریت اطلاعات سلامت، مرکز تحقیقات مدیریت اطلاعات سلامت، دانشگاه علوم پزشکی کاشان، کاشان، ایران

[۶] پژوهشگر مرکز تحقیقات انفورماتیک پزشکی، موسسه آینده پژوهی، دانشگاه علوم پزشکی کرمان، کرمان، ایران

نویسنده رابط: رضا عباسی، نشانی: کاشان، گروه فن آوری و مدیریت اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی کاشان، تلفن: ۵۵۵۸۳۱۸۸ ، پست الکترونیک: Rezaabbasi2001@gmail.com

تاریخ دریافت: ۹۶/۰۴/۶؛ پذیرش: ۹۶/۸/۰۶

مقدمه و اهداف: با توجه به روند رو به رشد سرطان روده‌بزرگ در ایران در سال‌های اخیر، پیش‌بینی پیامد سرطان و اطلاعات بالینی پایه مربوط به آن بااهمیت است. روش‌های داده‌کاوی در پیش‌بینی و تشخیص سرطان‌ها می‌تواند مورداستفاده قرار گیرند. هدف از انجام این مطالعه تعیین عملکرد دو الگوریتم پیش‌بینی کننده ماشین بردار پشتیبان و بگینگ در پیش بینی بقاء بیماران مبتلا به سرطان روده بزرگ است.

روش کار: جمعیت موردمطالعه ۵۷۰ بیمار مبتلا به سرطان روده‌بزرگ با مرحله تومور ۱ تا ۴، مراجعه‌کننده به بخش پرتودرمانی بیمارستان نمازی شیراز شامل ۳۳۸ بیمار زنده و ۲۳۲ بیمار فوت شده از سال ۱۳۸۵ تا ۱۳۹۰ می‌باشند. برای پیش بینی بقاء بیماران مبتلا به سرطان روده‌بزرگ از روش ماشین بردار پشتیبان و روش بگینگ استفاده شد. برای تحلیل داده‌ها نیز از نرم‌افزار Weka نسخه ۳.۶.۱۰ استفاده گردید.

یافته‌ها: بیشترین و کمترین محل قرارگیری تومورها مربوط به رکتوم و کولون چپ و به میزان ۵۱ و ۹ درصد بود. روش درمانی در بیش از ۸۰٪ از بیماران نیز ابتدا عمل جراحی و سپس شیمی‌درمانی و یا رادیوتراپی بود. در عملکرد دو الگوریتم بر اساس صحت، ویژگی و حساسیت محاسبه‌شده از ماتریس درهم‌ریختگی تعیین، مورداستفاده قرار گرفت. به ترتیب میزان صحت، ویژگی و حساسیت در الگوریتم ماشین بردار پشتیبان ۸۴/۴، ۸۰ و ۸۷/۵ درصد و در الگوریتم بگینگ ۸۳/۲، ۷۵ و ۸۸ درصد به دست آمد.

نتیجه گیری: نتایج حاصل از این مطالعه نشان داد که هر دو روش حساسیت و ویژگی قابل قبولی در پیش بینی بقاء بیماران مبتلا به سرطان روده‌بزرگ دارند اما ماشین بردار پشتیبان از میزان صحت بیشتری برخوردار بود

واژگان کلیدی: سرطان روده بزرگ، پیش‌بینی بقاء، داده کاوی، ماشین‌بردار پشتیبان، بگینگ

Expert PDF Trial

مقدمه

سرطان روده بزرگ سومین سرطان شایع در هر دو جنس در جهان محسوب می‌شود به‌گونه‌ای که در مردان پس از سرطان ریه و پروستات در رتبه سوم و در زنان پس از سرطان پستان، در رتبه دوم سرطان‌های شایع قرار می‌گیرد (۱). در ایران بر اساس آمار وزارت بهداشت سرطان روده بزرگ چهارمین سرطان شایع کشور در هر دو جنس است که ۸/۱۲ درصد کل سرطان‌ها را در برمی‌گیرد (۲). مطالعاتی که در زمینه سرطان روده بزرگ در ایران انجام‌گرفته نشان‌دهنده آن است که اگرچه میزان بروز این سرطان کمتر از میزان بروز جهانی است اما الگوی رخداد آن و همچنین سال‌های ازدست‌رفته ناشی از آن در کشور رو به افزایش است (۳ ,۴). تشخیص به‌موقع و پیش‌بینی پیامد این سرطان بسیار بااهمیت بوده و می‌تواند به درمان و بقاء بیماران کمک کند (۵ ,۶).

داده‌کاوی از روش‌هایی است که به‌منظور یافتن الگو از درون پایگاه‌های بزرگ داده و نیز پیش‌بینی نتایج سرطان‌ها با استفاده از الگوریتم‌های پیش‌بینی کننده بکار می‌رود (۷ ,۸). با کمک

Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

داده‌کاوی می‌توان الگوهایی که به‌سختی قابل‌تشخیص هستند را از پایگاه‌های داده استخراج نمود (۱۰ ,۹). امروزه رویکردهای داده‌کاوی مانند کشف دانش جدید از پایگاه‌های داده به‌عنوان ابزار تحقیقاتی مناسبی برای پژوهشگران حوزه پزشکی تبدیل‌شده است و محققین از تکنیک‌های مختلف برای پیش‌بینی بقا یا عود مجدد سرطان‌ها استفاده می‌کنند (۹). هدف از داده‌کاوی پیش‌بینی کننده در حوزه پزشکی و بالینی، رسیدن به مدلی است که بتوان با استفاده از اطلاعات مخصوص بیمار، پیامدها را پیش‌بینی نموده و بر اساس آن تصمیم‌گیری کرد (۱۱).

تاکنون مطالعات زیادی در ایران به بررسی احتمال بقاء سرطان روده بزرگ پرداخته‌اند اما در هیچ‌کدام از آن‌ها از روش‌های داده‌کاوی استفاده‌نشده است. همچنین در این مطالعات، اغلب عوامل دموگرافیک و عوامل مربوط به تومور (اندازه تومور، مرحله و درجه تمایز تومور) موردبررسی قرارگرفته‌اند (۱۹- ۱۲). در این مطالعه علاوه بر عوامل ذکرشده، عوامل دیگری از قبیل رویکردهای درمانی متفاوت و نیز تعداد گره‌های لنفاوی جداشده در حین عمل جراحی موردبررسی قرار گرفتند. علاوه بر آن علیرغم مطالعات متعددی که در جهان و با استفاده از روش‌های داده‌کاوی و مقایسه آن‌ها در پیش‌بینی بقاء بیماران مبتلا به سرطان روده بزرگ در سراسر جهان انجام گردیده، بازهم این بیماری به‌اندازه دیگر سرطان‌ها، نظیر سرطان ریه و پستان پوشش داده نشده است. در ادامه، نتایج بررسی‌ها نشان داد تاکنون مطالعه‌ای مبنی بر مقایسه دو الگوریتم پیش‌بینی کننده، یعنی روش تجمیعی بگینگ با ماشین بردار پشتیبان صورت نگرفته است (۲۵-۲۰،۵). بنابراین هدف از انجام این مطالعه، تعیین و مقایسه عملکرد این دو الگوریتم با یکدیگر در پیش‌بینی بقاء بیماران مبتلا به سرطان روده بزرگ بود.

روش کار

مجموعه داده مورداستفاده در پژوهش و توصیف آن

جمعیت موردمطالعه در این پژوهش تمامی بیماران (۵۷۰ بیمار) مبتلا به سرطان روده بزرگ که در طی سال‌های ۱۳۸۵ تا ۱۳۹۰ به مرکز رادیوتراپی آنکولوژی بیمارستان نمازی شیراز مراجعه کرده بودند و زمان‌های بقای متفاوت داشتند، انجام شد که از این تعداد، ۳۳۸ بیمار زنده و ۲۳۲ بیمار فوت‌شده بودند.

ابزار جمع‌آوری داده در این مطالعه چک‌لیستی پژوهشگر ساخته که مورد تأیید متخصصین داخلی و انکولوژی قرار گرفت و شامل سه قسمت اطلاعات دموگرافیک بیماران، اطلاعات مربوط به تشخیص و ویژگی‌های تومور و نیز اطلاعات مربوط به عود و محاسبه مدت بقاء بود. در این مطالعه اطلاعات تمامی بیماران به‌صورت گذشته‌نگر (از زمان مطالعه به سمت زمان قبل) از طریق اطلاعات پرونده پزشکی آن‌ها جمع‌آوری گردید. معیار ورود بیماران به این مطالعه، ابتلا به سرطان اولیه روده بزرگ است و بیماران دارای عود و یا متاستاز به سایر ارگان‌ها از این مطالعه کنار گذاشته شدند. بنابراین تمامی بیماران در این مطالعه دارای مرحله تومور یک تا سه می‌باشند و بیماران دارای مرحله چهار تومور وارد مطالعه نشدند. متغیرهای استخراج‌شده از پرونده بیماران و همچنین فراوانی، میانگین و انحراف معیار آن‌ها در قسمت یافته‌ها (جدول شماره ۱)، نشان داده‌شده است.

در این مطالعه سه متغیر مرحله تومور، عمق نفوذ تومور در روده بزرگ و تعداد گره‌های لنفاوی درگیر، بر اساس معیار مرحله‌بندی تومور در مورد سرطان روده بزرگ از سیستم مرحله‌بندی TNM که توسط کمیته مشترک سرطان آمریکا (AJCC[1]) (۲۶) پیشنهادشده مورداستفاده قرار گرفت و برای تومور چهار مرحله (stage) در نظر گرفته شد. T نشان‌دهنده نفوذ و گسترش تومور، N نشان‌دهنده انتشار تومور به غدد لنفاوی مجاور روده بزرگ و M نشان‌دهنده متاستاز تومور به دیگر بافت‌های بدن است.

آماده‌سازی داده‌ها

جهت آماده‌سازی داده‌ها از روش همسان‌سازی به‌وسیله نرمال‌سازی استفاده‌شده و ویژگی‌های استفاده‌شده در این پژوهش شامل دو نوع ویژگی‌های کیفی و ویژگی‌های کمی بودند. برای همسان‌سازی اثر هر یک از ویژگی‌ها، ویژگی‌های کمی نرمال‌سازی شد. روش نرمال‌سازی در این پژوهش Maximum–Minimum بود که تمامی متغیرهای کمی پیوسته با استفاده از این عملیات به بازه ۱ – نگاشته و برای این عمل از رابطه زیر استفاده شد.

$$X_{normalized} = \frac{X_j - X_{minimum}}{X_{maximum} - X_{minimum}}$$

تمامی داده‌ها در قالب فایل Excel جمع‌آوری‌شده و عملیات اصلی اجرای روش‌های یادگیری و نتایج به‌دست‌آمده در دسته‌بندی این طیف از بیماری‌ها با استفاده از نرم‌افزار Weka نسخه ۳.۶.۱۰ انجام گردید. این نرم‌افزار با زبان جاوا نوشته‌شده و به‌صورت منبع باز ارائه می‌شود و دائم در حال به‌روزرسانی است.

[1] American Joint Committee on Cancer

Expert PDF Trial

Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

همچنین این نرم‌افزار درصد صحت بالاتری را در برخی از الگوریتم‌ها نتیجه می‌دهد. عدم نیاز به برنامه‌نویسی، در دسترس بودن به‌صورت رایگان و کاربرپسند این نرم‌افزار، ازجمله دلایل انتخاب آن بوده است (۲۸, ۲۷). در این مطالعه ۱۶ ویژگی و دو روش طبقه‌بندی تجمیعی بگینگ و ماشین بردار پشتیبان جهت دسته‌بندی مورداستفاده قرارگرفته و نتایج این دو دسته بند با یکدیگر مقایسه شدند.

روش تجمیعی بگینگ

یکی از ساده‌ترین و درعین‌حال موفق‌ترین روش‌های تجمیعی برای بهبود مسئله دسته‌بندی است. این روش معمولاً در مورد درخت تصمیم به کار می‌رود، اما در مورد سایر الگوریتم‌های دسته‌بندی مانند بیز ساده، k نزدیک‌ترین همسایه و ... نیز می‌تواند به‌کاربرده شود. این روش برای داده‌های با حجم و ابعاد بالا بسیار مفید است، چراکه در این موارد پیدا کردن یک مدل یا دسته بند در یک مرحله به دلیل پیچیدگی بالا امکان‌پذیر نیست (۲۹). به دلیل این‌که مدل ترکیبی باعث کاهش واریانس هر یک از کلاسه بندهای مجزا می‌شود. معمولاً صحت حالت ترکیبی از یک کلاسه بند و یا پیش‌بینی در حالتی که به‌تنهایی بر روی کل داده پیاده‌سازی شده‌اند بهتر است (در حالت پیش‌بینی نیز این امر به‌صورت نظریه قابل‌بیان و عرضه است) و همچنین می‌توان نشان داد که در برابر اثرات داده‌های نویزی نیز قوی‌تر عمل می‌کند(۲۹).

ماشین بردار پشتیبان (SVM)

یکی از روش‌های طبقه‌بندی است که هدف آن ایجاد یک ابر صفحه با بیشترین حاشیه از نمونه‌های موجود در مرز بین دو کلاس به‌منظور جداسازی نمونه‌های دو کلاس طبقه‌بندی است که به‌صورت خطی از یکدیگر قابل‌تفکیک هستند، درصورتی‌که نمونه‌ها به‌صورت خطی قابل جداسازی نباشند، داده‌ها به فضایی با ابعاد بیشتر نگاشت پیدا می‌کنند، تا بتوان آن‌ها را در این فضای جدید به‌صورت خطی جدا کرد. این روش عموماً دقت بالایی در طبقه‌بندی از خود نشان می‌دهد. اما استفاده از آن دارای محدودیت‌هایی بوده که باعث شده است این روش محبوبیت لازم را کسب نکند. یکی از محدودیت‌های بزرگ SVM، کند بودن فرایند آموزش آن برای مسائل بزرگ است، ازاین‌رو استفاده از این روش معمولاً زمان‌بر خواهد بود. از سوی دیگر الگوریتم‌های یادگیری SVM پیچیده هستند و گاهی اوقات پیاده‌سازی آن‌ها دشوار است. روش بهینه‌سازی حداقلی متوالی[1] یک الگوریتم جدید یادگیری از ماشین بردار پشتیبان است که ازلحاظ مفهومی ساده و پیاده‌سازی و آموزش آن آسان و سریع‌تر از SVM است. SMO یک الگوریتم تکراری برای حل مسئله بهینه‌سازی در حل SVM است که این مسئله را به یک سری از کوچک‌ترین زیر مسئله‌های ممکن می‌شکند و سپس آن‌ها را به‌صورت تحلیلی حل می‌کند (۳۱, ۳۰).

ارزیابی مدل‌ها

در این مطالعه، بر اساس مطالعات مشابه (۲۱, ۳۲, ۳۳) برای انجام ارزیابی دو روش SVM و Bagging، از روش K-Fold Cross-Validation در اجرای آن‌ها استفاده شد. بر اساس این روش، داده‌ها به‌صورت تصادفی به k زیرمجموعه مجزا تقسیم می‌شوند. آموزش و آزمودن k بار انجام می‌شود، به این صورت که هر بار یکی از زیرمجموعه‌ها برای آزمودن مدل نگه‌داشته شده و بقیه برای آموزش مدل استفاده می‌شوند. این فرایند k بار تکرار می‌شود، به‌طوری‌که هر یک از زیرمجموعه‌ها دقیقاً یک بار برای آزمودن مدل به‌کاربرده می‌شوند. درنهایت نتیجه kتکرار برای دستیابی به یک برآورد نهایی میانگین‌گیری می‌شود (۳۴). در این مطالعه بر اساس مطالعات مشابه مقدار k برابر ۱۰ در نظر گرفته شد (۳۵, ۲۰).

معیارهای سنجش کارایی

به‌منظور سنجش کارایی الگوریتم ها، معیارهایی وجود دارد که صحت، حساسیت و ویژگی از متداول ترین آن ها به شمار می روند. برای محاسبه این معیارها از ماتریس جدول درهم‌ریختگی استفاده شده که کارایی الگوریتم‌ها را به تصویر می کشد (جدول شماره ۲).

حساسیت و ویژگی: توانایی هر آزمون در تشخیص درست موارد غیر بیمار و توانایی یک آزمون در تشخیص درست موارد بیماری به ترتیب مفاهیم حساسیت و ویژگی می‌باشند. این دو مفهوم با استفاده از ماتریس درهم‌ریختگی به‌صورت زیر تعریف می‌شوند:

$$\text{Specificity} = \frac{TN}{(TN+FP)} \qquad \text{Sensitivity} = \frac{TP}{(TP+FN)}$$

صحت: نرخ نمونه‌های درست طبقه‌بندی‌شده را نسبت به‌کل نمونه‌ها نشان می‌دهد و به‌صورت زیر تعریف می‌شود:

[1] Sequential Minimum Optimization (SMO)

Expert PDF Trial

Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

$$\text{Accuracy} = \frac{(TN+TP)}{(TN+TP+FN+FP)}$$

منحنی ROC (Receiver Operating Characteristic): نموداری است که از تقسیم نسبت بر میزان مثبت کاذب حساسیت (میزان مثبت واقعی) به دست می‌آید. در این حالت هر چه منحنی به گوشه چپ نمودار بیشتر باشد، صحت آن بیشتر است، زیرا در آنجا میزان مثبت واقعی «یک» و مثبت کاذب «صفر» است (۴۸).

یافته‌ها

همان‌طور که در جدول شماره ۱ نشان داده‌شده است، در حدود ۵۷٪ از بیماران مرد بودند. بیش از ۴۵٪ از بیماران در بازه سنی ۷۰-۵۰ سال قرار داشتند. بیشترین و کمترین محل قرارگیری تومورها مربوط به رکتوم و کولون چپ و به میزان ۵۱ و ۹ درصد بود. بیش از ۴۰٪ از بیماران در مرحله ۲ بیماری قرار داشتند. اندازه تومورها در بیش از ۷۰٪ از بیماران، کمتر از ۵ سانتی‌متر بود. همچنین درجه تمایز تومور در میان ۶۵٪ از بیماران خوب بود. در حدود ۷٪ از بیماران دچار متاستاز تومورهای مربوط به این سرطان به دیگر بافت‌های بدن شده بودند. روش درمانی در بیش از ۸۰٪ از بیماران نیز ابتدا عمل جراحی و سپس شیمی‌درمانی و یا رادیوتراپی بود.

صحت عملکرد الگوریتم‌ها به این صورت است که الگوریتم بگینگ از ۵۷۰ نمونه، ۴۷۴ نمونه را درست و ۹۶ نمونه را نادرست و الگوریتم SMO از ۵۷۰ نمونه، ۴۸۱ نمونه را درست و ۸۹ نمونه را نادرست دسته‌بندی کرده است. صحت دسته‌بندی در ماشین بردار پشتیبان و بگینگ به ترتیب ۸۴/۴ و ۸۳/۲ درصد به‌دست آمد.

دسته‌بندی بگینگ در مورد پیش‌بینی افراد زنده از ۳۳۸ نمونه، ۲۹۹ نمونه را درست پیش‌بینی کرده و در مورد پیش‌بینی افراد فوت‌شده نیز از ۲۳۲ نمونه فوت‌شده، ۱۷۵ نمونه را در دسته درست پیش‌بینی کرده است (جدول شماره ۳).

دسته‌بندی SMO نیز از ۳۳۸ نمونه زنده، ۲۹۶ نمونه را در دسته درست پیش‌بینی کرده و از ۲۳۲ نمونه فوت‌شده، ۱۸۵ نمونه را در دسته درست قرار داده است (جدول شماره ۴).

درنهایت معیارهای ارزیابی به‌دست‌آمده صحت، حساسیت و ویژگی برای الگوریتم SMO به ترتیب ۸۴/۴، ۸۷/۵ و ۸۰ درصد و برای الگوریتم بگینگ به ترتیب ۸۳/۲، ۸۸ و ۷۵ درصد محاسبه شد (نمودار شماره ۱).

در نمودار شماره ۲، منحنی ROC دو الگوریتم به تفکیک نشان داده‌شده است. همان‌طور که در این نمودار مشخص است، تمایل هر دو منحنی به گوشه سمت چپ نمودار بیشتر است و شیب نمودار راک در هر دو منحنی، نشان می‌دهد هر کدام از مدل‌ها عملکرد خوبی داشته و نتایج قابل قبولی را ارائه داده‌اند.

جدول شماره ۱ – فراوانی و درصد فراوانی متغیرهای مورداستفاده در مدل‌های پیش‌بینی

نام متغیر	نوع متغیر	مقادیر	فراوانی (درصد)
سن	طبقه‌ای	کمتر از ۵۰ سال	۱۸۸(۳۳/۲)
		۷۰-۵۰ سال	۲۵۹(۴۵/۷)
		بیشتر از ۷۰ سال	۱۲۰(۲۱/۱)
جنسیت	دو دویی	مرد	۳۲۵(۵۷/۳)
		زن	۲۴۲(۴۲/۷)
محل درگیری تومور	طبقه‌ای	رکتوم	۲۹۰(۵۱/۱)
		سکوم و کولون راست و کولون افقی	۱۰۸(۱۹)
		کولون چپ	۵۳(۹/۹)
		سیگموئید	۱۱۶(۲۰)
اندازه تومور	دو دویی	کمتر از 5cm	۴۰۵(۷۱/۴)
		بیشتر از 5cm	۱۶۲(۲۸/۶)
تعداد گره‌های لنفاوی برداشته‌شده در جراحی	دو دویی	کمتر از ۱۲ گره	۰
		بیشتر از ۱۲ گره	۵۰۶(۸۹/۲)
		Missing	۶۱(۱۰/۸)
تعداد گره‌های لنفاوی درگیر	طبقه‌ای بر اساس معیارهای AJCC	۰	۳۳۶(۵۹/۲)
		۱	۱۲۷(۲۲/۳)
		۲	۸۰(۱۴/۱)
		Missing	۲۴(۴/۶)

Expert PDF Trial

Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

نسبت گره‌های لنفاوی درگیر به تعداد گره‌های لنفاوی جداشده	دو دویی		کمتر از ۰/۱۶ بیشتر از ۰/۱۶ Missing	۳۶۸(۶۵) ۱۴۶(۲۵/۷) ۵۳(۹/۳)
عمق نفوذ تومور در روده بزرگ	طبقه‌ای	بر اساس معیارهای AJCC	۱ ۲ ۳	۱۱۷(۲۰/۸) ۴۲۲(۷۴/۳) ۲۸(۴/۹)
متاستاز تومور به دیگر بافت‌ها	دو دویی		دارد ندارد	۵۳۰(۹۳/۴) ۳۷(۶/۶)
مرحله تومور	طبقه‌ای	بر اساس معیارهای AJCC	۱ ۲ ۳ ۴ Missing	۹۵(۱۶/۷) ۲۳۶(۴۱/۶) ۱۸۷(۳۳) ۳۸(۶/۷) ۱۱(۲)
درجه تمایز تومور	طبقه‌ای		خوب متوسط ضعیف	۳۷۵(۶۶) ۱۵۶(۲۷/۵) ۳۶(۶/۵)
تهاجم تومور به عروق	دو دویی		دارد ندارد Missing	۳۵۴(۶۲/۴) ۲۰۸(۳۶/۶) ۵(۱)
تهاجم عصبی تومور	دو دویی		دارد ندارد Missing	۴۰۳(۷۱) ۱۶۰(۲۸/۲) ۴(۰/۸)
نوع درمان	دو دویی		جراحی و سپس شیمی‌درمانی/رادیوتراپی شیمی‌درمانی/رادیوتراپی و سپس جراحی	۴۶۱(۸۱/۳) ۱۰۶(۱۸/۷)
نوع بیمارستان	دو دویی		دولتی غیردولتی	۲۴۵(۴۳/۳) ۳۲۷(۵۶/۷)
زمان	عددی		مدت‌زمان بین تشخیص تا زمان فوت یا آخرین پیگیری (ماه)	۴۹/۴۳ ± ۲۴/۷۹

جدول شماره ۲ – ماتریس درهم‌ریختگی

	Predicted class C1	Predicted class C2
Actual class C1	TP (True Positive)	FN (False Negative)
Actual class C2	FP (False Positive)	TN (True Negative)

جدول شماره ۳ – ماتریس درهم‌ریختگی حاصل از الگوریتم بگینگ

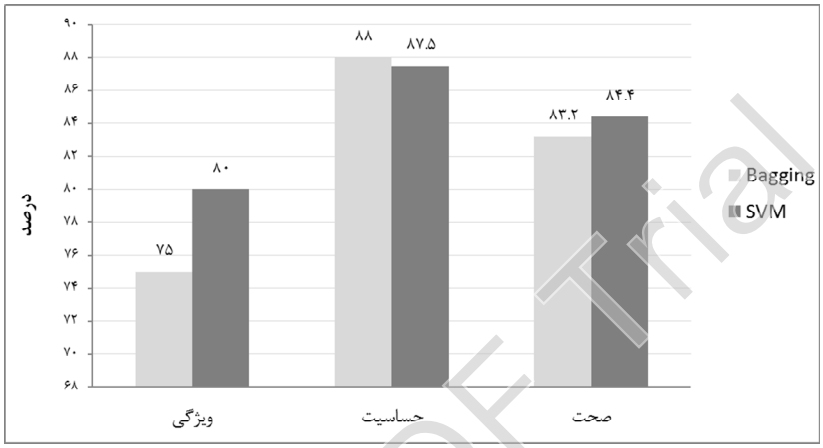
		نتیجه پیش‌بینی	
		پیش‌بینی‌شده زنده	پیش‌بینی‌شده فوت‌شده
نتیجه واقعی	واقعی زنده	۲۹۹	۳۹
	واقعی فوت‌شده	۵۷	۱۷۵

Expert PDF Trial

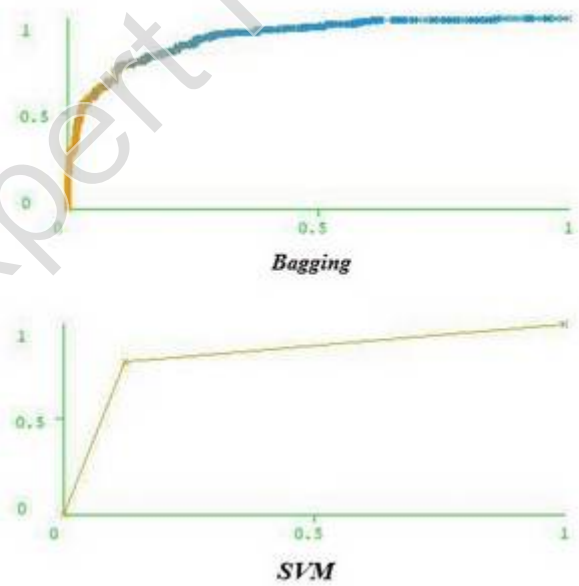
Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

جدول شماره ۴ – ماتریس درهم‌ریختگی حاصل از الگوریتم ماشین بردار پشتیبان

		نتیجه پیش‌بینی	
		پیش‌بینی‌شده زنده	پیش‌بینی‌شده فوت‌شده
نتیجه واقعی	واقعی زنده	۲۹۶	۴۲
	واقعی فوت‌شده	۴۷	۱۸۵



نمودار شماره ۱ – مقایسه معیارهای ارزیابی کارایی دو مدل پیش‌بینی



نمودار شماره ۲ – منحنی ROC مربوط به دو الگوریتم SVM و Bagging

Expert PDF Trial

Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

بحث

در این مطالعه با به‌کارگیری دو الگوریتم پرکاربرد در داده‌کاوی شامل SMO و Bagging روی داده‌های بیماران مبتلا به سرطان روده بزرگ بیمارستان نمازی شیراز دو مدل پیش‌بینی بقاء برای بیماران مورداستفاده قرار گرفت. نتایج این مطالعه نشان داد که هر دو روش از صحت خوبی برخوردار بودند. اما باوجود عملکرد مشابه این دو روش در پیش‌بینی بقاء سرطان روده بزرگ، تفاوت‌هایی نیز بین آن‌ها وجود دارد. با در نظر گرفتن نتایج به‌دست‌آمده، اگرچه از دسته‌بندی ماشین بردار پشتیبان صحت نسبتاً بالاتری نسبت به روش تجمیعی بگینگ حاصل شد، اما دسته‌بندی ماشین بردار پشتیبان در حجم بالایی از داده‌ها کندتر از روش تجمیعی بگینگ عمل می‌کند. بزرگ‌ترین محبوبیت ماشین بردار پشتیبان در بین پژوهشگران این است که داده‌های نویزی و پرت بر صحت عملکرد آن تأثیر ندارد (۳۶).

پیش‌بینی بقاء با استفاده از داده‌کاوی در مطالعات متعددی نظیر پیش‌بینی بقاء بیماران مبتلا به سرطان ریه (۳۹-۳۷) و همچنین سرطان پستان (۴۲-۴۰) انجام‌شده است. مطالعات کمتری به پیش‌بینی بقا سرطان روده بزرگ پرداخته‌اند (۲۴, ۲۱) و تاکنون هیچ مطالعه‌ای به‌طور مستقیم با استفاده از این دو روش پیرامون پیش‌بینی بقاء بیماران مبتلا به سرطان روده بزرگ انجام‌نشده است. مطالعات مشابه ازلحاظ تعداد رکورد، همچنین نوع مدل، صحت و متغیرهای استفاده‌شده موردبحث قرارگرفته است.

تعداد رکوردها

به‌منظور ارزیابی‌های دقیق‌تر، در مطالعات داده‌کاوی، بهتر است تا حد امکان از حجم زیادی از داده‌ها استفاده گردد. اغلب مطالعات انجام‌شده در این حوزه با استفاده از داده‌های موجود در پایگاه‌های داده بزرگ نظیر SEER، (نظارت اپیدمیولوژی و نتایج نهایی) بوده است (۲۱, ۴۳) که دارای حجم بسیار زیادی از داده است، اما این نوع پایگاه‌های داده، مقادیر مفقودشده بسیار زیادی دارند (۴۴). باوجوداینکه در مطالعه حاضر، از داده‌های تعداد کمتری از بیماران بومی و در یک منطقه استفاده شد، اما این اطلاعات دارای مقادیر مفقودشده بسیار کمی بودند.

نوع و صحت مدل

نوع مدل

هیچ‌یک از مطالعات مشابه از دو روش بگینگ و ماشین بردار به شکل هم‌زمان در مطالعه‌شان استفاده نکرده‌اند. مطالعات انجام‌شده مشابه یا با استفاده از یکی از این روش‌ها به پیش‌بینی بقا سرطان روده بزرگ پرداخته‌اند، مانند مطالعه (Fathy 43) که تنها با استفاده از شبکه‌های هوش مصنوعی انجام‌گرفته است و یا مانند دیگر مطالعات (۴۵, ۲۱, ۲۰, ۵) با استفاده از الگوریتم‌ها و روش‌های متعدد و به شکل ترکیبی به پیش‌بینی بقاء سرطان روده بزرگ پرداخته‌اند.

در رابطه با انواع روش‌های استفاده‌شده، بررسی منابع نشان داد که همانند مطالعه ما، دو مطالعه (۴۵, ۵) از روش SVM و یک مطالعه (۲۱) از روش بگینگ در کنار سایر روش‌های متعدد به‌منظور پیش‌بینی بقاء سرطان روده بزرگ استفاده نموده‌اند. باوجود کارایی روش بگینگ، این روش تنها در یک مطالعه (۲۱) استفاده‌شده است، بنابراین پیشنهاد می‌شود مطالعات بیشتری با استفاده از این روش انجام گیرد.

صحت مدل

نتایج این مطالعه نشان داد الگوریتم SMO نسبت به Bagging از دقت بیشتری برخوردار بود. بااین‌وجود همان‌گونه که از منحنی ROC این دو مدل مشخص است، اختلاف دقت در آن‌ها، چشمگیر نبوده و در مقایسه با نتایج مطالعه (Amasyali 46) که به مقایسه دقت انواع الگوریتم‌ها پرداخته است همخوانی دارد. نتایج حاصل از مطالعه (Pala 45) بر روی بیماران مبتلا به سرطان روده بزرگ نشان داد الگوریتم SVM در میان دیگر الگوریتم‌های مورداستفاده دارای بیشترین میزان صحت بوده که با مطالعه ما همخوان است. در مطالعه (Gao 5) نیز، میزان صحت الگوریتم SVM در حدود ۸۱ درصد اندازه‌گیری شد که همانند مطالعه ما، از مقدار خوبی برخوردار است. نزدیک بودن دقت گزارش‌شده در این مطالعه به‌دقت مطالعه ما، ممکن است به دلیل شباهت تعداد و ماهیت متغیرهای استفاده‌شده در این دو مطالعه باشد. همچنین بر اساس نتایج این مطالعه و دیگر مطالعه انجام‌شده در این زمینه (۲۰) میزان صحت به‌دست‌آمده در دو الگوریتم SVM و Bagging از الگوریتم‌های TRF، AD، RBFN و MLP بیشتر بود و بنابراین کارایی این دو الگوریتم از الگوریتم‌های فوق بیشتر به نظر می‌رسد.

Expert PDF Trial

Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

متغیرهای استفاده‌شده به‌منظور پیش‌بینی بقاء

نتایج حاصل از مطالعه Setareh و همکاران (۴۷) بر روی سرطان روده بزرگ نشان داد علاوه بر روش‌های یادگیری ماشین، انتخاب ویژگی‌ها و میزان ارتباط آن‌ها با هدف مطالعه، بر میزان صحت به‌دست‌آمده مؤثر خواهد بود. بنابراین صحت به‌دست‌آمده به ماهیت داده‌ها و نوع ویژگی‌های انتخابی برای مطالعه بستگی داشته و بهتر است در انجام این مطالعات، همواره از ویژگی‌های مرتبط‌تری استفاده شود. در مطالعه حاضر از ۱۶ متغیر مرتبط استفاده‌شده است که در مقایسه با مطالعات مشابه (۴۳, ۵) که هر دو از ۲۰ متغیر استفاده کرده‌اند، کمتر است. بااین‌وجود ازلحاظ نوع متغیرهای مورداستفاده، برخی متغیرهای بررسی‌شده در این مطالعه نظیر نسبت گره‌های لنفاوی درگیر به تعداد گره‌های لنفاوی جداشده، عمق نفوذ تومور در روده بزرگ و نیز تهاجم عصبی تومور، در مطالعات مشابه (۴۳, ۲۵, ۵) موردبررسی قرار نگرفته‌اند و در برخی از آن‌ها متغیرهای دیگری نظیر نژاد، وضعیت تأهل، نوع بافت بر اساس طبقه‌بندی بین‌المللی بیماری‌های سرطان ICD-O3 در نظر گرفته‌شده‌اند. همچنین در مطالعه ما برای متغیرهای عمق نفوذ تومور در روده بزرگ، تعداد گره‌های لنفاوی درگیر و مرحله تومور از معیار (AJCC 26) استفاده‌شده است که این موضوع تنها در یک مطالعه پیش‌بینی بقاء (۵) به کار گرفته‌شده است. مطالعه دیگری توسط (Al-Bahrani 21) انجام شد که با استفاده از ۱۳ متغیر، احتمال مدت‌زمان بقاء بیماران مبتلا به سرطان روده بزرگ، پیش‌بینی گردید. در برخی دیگر از مطالعات مشابه (۳۷) فقط متغیرهای مربوط به ویژگی‌های تومور و عوامل خطرساز موردبررسی قرارگرفته است، درحالی‌که در مطالعه ما متغیر مربوط به نوع درمان نیز برای پیش‌بینی پیامد استفاده‌شده است.

در بررسی متغیرهای استخراج‌شده از پرونده بیماران، محدودیت‌هایی وجود داشت. بیماران مراجعه‌کننده به بیمارستان دارای زمان‌های بقای متفاوتی بودند، به همین دلیل محاسبه بقاء آن‌ها در سال‌های مختلف امکان‌پذیر نبود. همچنین ازآنجاکه پرونده بیماران تنها به مقاصد درمانی تکمیل و نگهداری می‌شود، با استفاده از اطلاعات موجود در پرونده بیماران اطلاعات اپیدمیولوژیکی مفیدی ازجمله سوابق خانوادگی بیمار قابل‌دسترسی نبود. پیشنهاد می‌شود در مطالعات آینده با انجام آنالیز حساسیت، عوامل خطر مختلف در بقاء بیماران، شناسایی و میزان اهمیت آن‌ها نیز تعیین گردد. آنالیز حساسیت می‌تواند کاربرد زیادی برای متخصصین بالینی داشته باشد به این صورت که متخصصین بالینی درگیر در درمان بیماری سرطان روده بزرگ می‌توانند با استفاده از این مدل‌ها و نیز سیستم‌های پشتیبان تصمیم، عوامل خطر مختلف را شناسایی نموده و بر اساس آن‌ها تصمیم‌گیری نمایند.

AJCC از سیستمی به نام TNM (مرحله‌بندی تومور) به‌منظور پیش‌بینی بقا سرطان روده بزرگ استفاده می‌نماید. پیشنهاد می‌شود در مطالعات آتی عملکرد این سیستم با روش‌های داده‌کاوی مقایسه گردد. همچنین ازآنجاکه این دو الگوریتم دارای میزان صحت بالایی می‌باشند پیشنهاد می‌شود در پژوهش‌های آتی، عملکرد این الگوریتم‌ها نیز بر داده‌های موجود در پایگاه داده SEER نیز تعیین و یا برای محاسبه پیش‌بینی بقاء در سایر سرطان‌ها به کار گرفته شود.

نتیجه‌گیری

پیش‌بینی بقاء منجر به استفاده بهینه از منابع موجود در درمان بیماران خواهد شد. به این منظور استفاده از الگوریتم‌های مناسب اهمیت زیادی دارد. نتایج پژوهش نشان داد که دو روش داده‌کاوی شامل SMO و Bagging صحت بالایی در پیش‌بینی بقاء بیماران مبتلا به سرطان روده بزرگ دارد که این میزان در الگوریتم SMO بیشتر بود. نتایج این مطالعه می‌تواند کاربرد زیادی برای متخصصین بالینی درگیر در درمان بیماران مبتلا به سرطان روده بزرگ داشته باشد. مدل‌های پیشنهادی می‌توانند با دقت بالایی، پیامد بیماری در بیمار مبتلا به سرطان روده بزرگ را پیش‌بینی کنند.

تشکر و قدردانی

این پژوهش حاصل طرح تحقیقاتی مصوب مرکز تحقیقات انفورماتیک پزشکی دانشگاه علوم پزشکی کرمان به شماره ۹۴/۲۳۱ است. بدین‌وسیله نویسندگان مقاله، مراتب سپاس و قدردانی خود را از ریاست بیمارستان و نیز ریاست بخش رادیوتراپی بیمارستان نمازی شیراز به خاطر همکاری ایشان، اعلام می‌دارند.

Expert PDF Trial

Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

منابع

1. Garcia M, Jemal A, Ward E, Center M, Hao Y, Siegel R, et al. Global cancer facts & figures 2007. Atlanta, GA: American cancer society. 2007; 1 :52.
2. Ministry of Health and Medical Education. cancer office of management of noncontiguous disease of Health Assistance. Tehran. 2007.
3. Ostenfeld EB, Erichsen R, Iversen LH, Gandrup P, Norgaard M, Jacobsen J. Survival of patients with colon and rectal cancer in central and northern Denmark, 1998–2009. Clinical epidemiology. 2011; 3: 27.
4. SaberiFiroozi M, Kamali D, Yousefi M, Mehrabani D, Khademolhosseini F, Heydari ST, et al. Clinical characteristics of colorectal cancer in Southern Iran, 2005. Iranian Red Crescent Medical Journal. 2007; 2007: 209-11.
5. Gao P, Zhou X, Wang Z-n, Song Y-x, Tong L-l, Xu Y-y, et al. Which is a more accurate predictor in colorectal survival analysis? Nine data mining algorithms vs. the TNM staging system. PLoS One. 2012; 7: e42015.
6. Parkin DM, Muir CS. Cancer Incidence in Five Continents. Comparability and quality of data. IARC scientific publications. 1992: 45-173.
7. Han J, Pei J, Kamber M. Data mining: concepts and techniques: Elsevier; 2011.
8. Mladenic D. Data mining and decision support: integration and collaboration: Springer Science & Business Media; 2003.
9. Gupta S, Kumar D, Sharma A. Performance analysis of various data mining classification techniques on healthcare data. International journal of computer science & Information Technology (IJCSIT). 2011; 3.
10. Mathur R, Schaffer JD, LandJr WH, Heine JJ, Eschrich S, Yeatman T. Evolutionary computation with noise perturbation and cluster analysis to discover biomarker sets. Procedia Computer Science. 2011; 6: 153-8.
11. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics. 2008; 77: 81-97.
12. Karimi Zarchi, Saadat A, Jalalian HR, Esmaeili M. Epidemiology and survival analysis of colorectal cancer and its related factors. Trauma Monthly. 2011; 239-43.
13. Asghari Jafarabadi M, Mohammadi SM, Hajizadeh E, Fatemi SR. An evulation of 5-year survival of metastatic colon and rectal cancer patients using cumulative incidence models. Koomesh. 2013; 14: 207-14.
14. Emami S, Fatemi A, Farajzadegan Z, Movahed-Abtahi S. Epidemiology of colorectal cancer in Isfahan province. Govaresh. 2005; 10: 134-9.
15. Heidarnia MA, Monfared ED, Akbari ME, Yavari P, Amanpour F, Mohseni M. Social determinants of health and 5-year survival of colorectal cancer. Asian Pacific Journal of Cancer Prevention. 2013; 14: 5111-6.
16. Moghimi-Dehkordi B, Safaee A, Zali MR. Prognostic factors in 1,138 Iranian colorectal cancer patients. International journal of colorectal disease. 2008; 23: 683-8.
17. Montazer Haghighi M, Vahedi M, Mohebbi SR, Pourhoseingholi MA, Fatemi SR, Zali MR. Evaluation of 4-year survival in familial and non familial colorectal cancer. Medical Science Journal of Islamic Azad Univesity-Tehran Medical Branch. 2010; 20: 40-4.
18. Roshanaei G, Komijani A, Sadighi A, Faradmal J. Prediction of survival in patients with colorectal cancer referred to the Hamadan MRI center using of Weibull parameter model and determination of its risk factors during 2005-2013. 2014.
19. Saedi H, Ghavamnasiri M, Toosi M, Homaei F, Roodbari S. The assessment of prognostic factors in patients with nonmetastatic rectal Adenocarcinoma referred to Omid Hospital (Mashhad), Iran. Journal of Gorgan University of Medical Sciences. 2009; 11.
20. Abbasi R, Montazeri M, Zare M, editors. A Rule Based Classification Model to Predict Colon Cancer Survival. 1th Afzalipour International Medical congress on Pathology; 2015: Kerman University of Medical Sciences, Kerman, Iran.
21. Al-Bahrani R, Agrawal A, Choudhary A, editors. Colon cancer survival prediction using ensemble data mining on SEER data. Big Data, 2013 IEEE International Conference on; 2013: IEEE.
22. Bottaci L, Drew PJ, Hartley JE, Hadfield MB, Farouk R, Lee PW, et al. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. The Lancet. 1997;350.469-472.
23. Grumett S, Snow P, Kerr D. Neural networks in the prediction of survival in patients with colorectal cancer. Clinical colorectal cancer. 2003; 2: 239-44.
24. Snow PB, Kerr DJ, Brandt JM, Rodvold DM. Neural network and regression predictions of 5-year survival after colon carcinoma treatment. Cancer. 2001; 91: 1673-8.
25. Valera VA, Walter BA, Yokoyama N, Koyama Y, Iiai T, Okamoto H, et al. Prognostic groups in colorectal carcinoma patients based on tumor cell proliferation and classification and regression tree (CART) survival analysis. Annals of surgical oncology. 2007;14(1):34-40.
26. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Annals of surgical oncology. 2010; 17: 1471-4.
27. Eshwari Girish Kulkarni, Raj B. Kulkarni. WEKA Powerful Tool in Data Mining. International Journal of Computer Applications. 2016:10-5.
28. Markov Z, Russell I. An introduction to the WEKA data mining system. ACM SIGCSE Bulletin. 2006; 38: 367-8.
29. Setareh S, Safaei AA, Najafi F. Using machine learning techniques to differentiate acute coronary syndrome. Journal of Kermanshah University of Medical Sciences (J Kermanshah Univ Med Sci). 2015; 18: 631-9.
30. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011; 2: 27.
31. Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.
32. Olson DL, Delen D. Advanced data mining techniques: Springer Science & Business Media; 2008.
33. Witten IH, Mining EFD. Practical machine learning tools and techniques Morgan Kaufmann. San Francisco; 2005.
34. Alpaydin E. Introduction to machine learning: MIT press; 2004.
35. Nabovati E, Azizi A, Abbasi E, Vakili-Arki H, Zarei J, Razavi A. Using data mining to predict outcome in burn patients: a comparison between several algorithms. Health Inf Manage. 2014; 10: 799.
36. Koturwar P, Girase S, Mukhopadhyay D. A survey of classification techniques in the area of big data. arXiv preprint arXiv:150307477. 2015.
37. Agrawal A, Misra S, Narayanan R, Polepeddi L, Choudhary A. Lung cancer survival prediction using ensemble data mining on SEER data. Scientific Programming. 2012; 20: 29-42.
38. Murty NR, Babu MP. A Critical Study of Classification Algorithms for LungCancer Disease Detection and Diagnosis.

International Journal of Computational Intelligence Research. 2017; 13: 1041-8.
39. Christopher T, J.J. b. Study of Classification Algorithm for Lung Cancer Prediction. International Journal of Innovative Science, Engineering & Technology. 2016; 3: 42-9.
40. Ahmad L, Eshlaghy A, Poorebrahimi A, Ebrahimi M, Razavi A. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013; 4: 3.
41. Lakshmi K, Krishna MV, Kumar SP. Performance comparison of data mining techniques for prediction and diagnosis of breast cancer disease survivability. ASIAN JOURNAL OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY. 2013; 3.
42. Shukla A, Tiwari R, Kaur P, editors. Knowledge based approach for Diagnosis of Breast Cancer. Advance Computing Conference, 2009 IACC 2009 IEEE International; 2009: IEEE.
43. Fathy SK, A predication survival model for colorectal cancer. Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications; 2011: World Scientific and Engineering Academy and Society (WSEAS).
44. Sanders CM, Saltzstein SL, Schultzel MM, Nguyen DH, Stafford HS, Sadler GR. Understanding the limits of large datasets. Journal of Cancer Education. 2012; 27: 664-9.
45. Pala T, Camurcu A. Design of Decision Support System in the Metastatic Colorectal Cancer Data Set and Its Application. 2016.
46. Amasyali M, Ersoy O, editors. Comparison of single and ensemble classifiers in terms of accuracy and execution time. Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on; 2011: IEEE.
47. Sogand Setareh, Reza Abbasi, Misagh Zahiri Esfahani, Bandamiri MZ, editors. Using machine learning techniques to predict colon cancer survival. Medical Informatics congress; 2017; Mashhad.
48. Gönen M. Receiver operating characteristic (ROC) curves. SAS Users Group International (SUGI). 2006; 31: 210-231.

Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

Expert PDF Trial

Downloaded from irje.tums.ac.ir at 15:47 IRDT on Monday July 23rd 2018

Iranian Journal of Epidemiology Spring 2018; 14 (1): 19-29.
Original Article

Using Data Mining for Survival Prediction in Patients with Colon Cancer

Setareh S[1], Zahiri Esfahani M[2], Zare Bandamiri M[3], Raeesi A[4], Abbasi R[5,6]

[1]PhD Student of Medical Informatics, Faculty of Paramedicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2] PhD Student of Health Information Management, Faculty of Management and Medical Information Sciences, Iran University of Medical Sciences, Tehran, Iran
[3] Department of Radiation Oncology, Namazi hospital, Shiraz University Medical Sciences, Shiraz, Iran
[4] MSc Student of Medical Informatics, Medical Informatics Research Center, Institute of Future Study , Kerman University of Medical Sciences , Kerman, Iran
[5] PhD Student of Health Information Management, Health Informatin Management Research Center, Kashan University of Medical Sciencess, Kashan, Iran
[6] Researcher of Medical Informatics Research Center (MIRC), Institute of Future Study, Kerman University of Medical Sciences , Kerman, Iran
Corresponding author: Abbasi R, Rezaabbasi2001@gmail.com
(Received 27 June 2017**; Accepted** 28 September 2017**)**

Background and Objectives: Colon cancer is the third most common cancer in the world and the fourth most common cancer in Iran. It is very important to predict the cancer outcome and its basic clinical data. Due to to the high rate of colon cancer and the benefits of data mining to predict survival, the aim of this study was to survey two widely used machine learning algorithms, Bagging and Support Vector Machines (SVM), to predict the outcome of colon cancer patients.

Methods: The population of this study was 567 patients with stage 1-4 of colon cancer in Namazi Radiotherapy Center, Shiraz in 2006-2011. Three hundred and thirty eight patients were alive and 229 patients were dead. We used the Support Vector Machines (SVM) and Bagging methods in order to predict the survival of patients with colon cancer. The Weka software ver 3.6.10 was used for data analysis.

Results: The performance of two algorithms was determined using the confusion matrix. The accuracy, specificity, and sensitivity of the SVM was 84.48%, 81%, and 87%, and the accuracy, specificity, and sensitivity of Bagging was 83.95%, 78%, and 88%, respectively.

Conclusion: The results showed both algorithms have a high performance in survival prediction of patients with colon cancer but the Support Vector Machines has a higher accuracy.

Keywords: Colon cancer, Survival prediction, Data mining, Support vector machines, Bagging

Expert PDF Trial