

خون

فصلنامه علمی تحقیقاتی

دوره ۱۲ شماره ۴ زمستان ۹۴ (۳۵۷-۳۴۷)

مقاله پژوهشی

تشخیص لوسومی لنفوسیتی و میلوئیدی حاد با استفاده از انتخاب ژن داده‌های ریزآرایه و الگوریتم‌های داده کاوی

راضیه شیخ‌پور^۱، مهدی آفاصرام^۲، رباب شیخ‌پور^۳

چکیده سابقه و هدف

تکنولوژی ریزآرایه، یک تصویر کلی از میزان بیان هزاران ژن به طور هم زمان ارایه می‌دهد. تفسیر داده‌های ریزآرایه بدون آنالیز آماری و روش‌های هوش مصنوعی ممکن نیست. هدف این مقاله، تشخیص انواع لوسومی حاد با استفاده از مجموعه داده‌های ریزآرایه و الگوریتم‌های داده کاوی بود.

مواد و روش‌ها

در این مطالعه توصیفی از داده‌های بیان ۷۱۲۹ ژن مربوط به ۷۲ بیمار مبتلا به لوسومی استفاده شد. سپس با انتخاب ژن‌های مهم بر اساس روش‌های ضریب همبستگی، بهره اطلاعاتی، نسبت بهره و امتیاز Fisher و با استفاده از روش‌های جداکننده خطی، ماشین بردار پشتیبان، k نزدیک‌ترین همسایه، بیزین ساده، شبکه بیزین، نزدیک‌ترین میانگین، رگرسیون لجستیک، شبکه عصبی پرسپترون چند لایه و درخت تصمیم J48 برروی ژن‌های انتخاب شده به تشخیص لوسومی میلوژنیک و لنفوسیتیک حاد پرداخته شد.

یافته‌ها

روش‌های نزدیک‌ترین میانگین، ماشین بردار پشتیبان، k نزدیک‌ترین همسایه، بیزین ساده و شبکه عصبی پرسپترون چند لایه با استفاده از ۳۹ ژن انتخاب شده توسط نسبت بهره با دقت ۱۰۰٪، قادر به تشخیص لوسومی میلوژنیک و لنفوسیتیک حاد هستند. هم چنین روش ماشین بردار پشتیبان با استفاده از ۸۷ ژن انتخاب شده توسط بهره اطلاعاتی و روش شبکه عصبی پرسپترون چند لایه با استفاده از ۱۳۳ ژن انتخاب شده توسط بهره اطلاعاتی با دقت ۱۰۰٪، قادر به تشخیص آن می‌باشدند.

نتیجه گیری

نتایج این مطالعه نشان داد که انتخاب ژن‌ها و الگوریتم‌های داده کاوی قادر به تشخیص انواع لوسومی با دقت بسیار بالایی هستند، بنابراین با استفاده از این روش‌ها، می‌توان تصمیمات مناسبی در مورد نحوه تشخیص و درمان بیماران گرفت.

کلمات کلیدی: لوسومی لنفوسیتیک حاد، لوسومی میلوژنیک حاد، آنالیز ریزآرایه، داده کاوی

تاریخ دریافت: ۹۳/۱۰/۱۷
تاریخ پذیرش: ۹۴/۴/۲۱

۱- دانشجوی دکترا کامپیوتر - دانشکده مهندسی برق و کامپیوتر - دانشگاه یزد - یزد - ایران

۲- دکترای تخصصی کنترل سیستم‌ها - دانشیار دانشکده مهندسی برق و کامپیوتر - دانشگاه یزد - یزد - ایران

۳- مؤلف مسئول: PhD بیوشیمی - دانشکده پزشکی، واحد یزد، دانشگاه آزاد اسلامی و مرکز تحقیقات خون و انکولوژی، دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران، صندوق پستی: ۸۹۱۵۶-۵۶۹۶۵

۴۵۶

بسیاری از مطالعه ها روند بد خیمی لوسمی را به ناهنجاری های ژنتیکی نسبت می دهند و مطالعه های زیادی پیرامون کشف عوامل مولکولی در گیر در این بیماری صورت گرفته است (۷، ۸). یکی از حوزه های جدید دانش در کشف بیان ژن ها در حالت بیماری، استفاده از تکنولوژی ریز آرایه (میکرواری) است که یک تصویر کلی از میزان بیان ژن را ارایه می دهد (۸). تکنولوژی ریز آرایه که روشنی بسیار قدرتمند است، امکان بررسی هم زمان بسیاری از فعل و انفعالات زیستی را فراهم می کند و انتظار می رود با تحلیل آماری تغییرات بیان هزاران ژن به طور هم زمان، بتوان ژن های مؤثر در سرطان را شناسایی و در زمینه درمان این بیماری گام های مهمی برداشت (۹-۱۵). این تکنولوژی در دو زمینه ژنومیکس (مطالعه مجموعه ژن های موجود زنده) و پروتئومیکس (مطالعه مجموعه پروتئین های موجود زنده) کاربردهای وسیعی دارد (۸). در روش ریز آرایه هر توالی ژنی شناخته شده مورد نظر به عنوان یک پروب (Probe) روی یک آرایه (Array) شیشه ای یا نایلونی چاپ می شود. mRNA از بافت یا نمونه خون با رنگ های فلورسنت علامت گذاری می شود و پروب ها بر روی یک آرایه هیبرید می شود. به طور کلی برای تهیه آرایه DNA باید طبق مراحل زیر عمل کرد: نمونه گیری، خالص سازی نمونه، جداسازی mRNA ها، انجام رونویسی معکوس و تهیه cDNA، متصل کردن cDNA به رنگ های فلورسنت، ریختن محلول بر روی سطح ریز آرایه که از قبل توسط توالی های ژن مورد نظر پوشیده شده است، انجام هیبریداسیون میان DNA ها و توالی های سطح ریز آرایه، شستشو، بررسی و پردازش نتایج (۱۶). مهم ترین کاربردهای ریز آرایه عبارتند از؛ بررسی بیان ژن و تغییرات آن در اثر عواملی مانند درمان، عوامل بیماری زا، آسیب سلول، هیبرید سازی مقایسه ای ژنوم، تعیین محتوا ی ژنوم موجودات زنده، مقایسه آنها با یکدیگر، شناسایی چند شکلی های تک نوکلئوتیدی، تشخیص بیماری و طبقه بندی سرطان (۱۷). ابعاد بالا، تعداد نسبتاً کم نمونه ها و تغییر پذیری ذاتی در فرآیندهای آزمایشگاهی و بیولوژیکی باعث ایجاد مشکلاتی در آنالیز داده های ریز آرایه شده است، از این رو، اولین گام مهم در آنالیز داده های ریز آرایه، کاهش

سرطان بعد از بیماری های قلبی - عروقی، دو میں علت اصلی مرگ و میر در جهان می باشد. سرطان یک بیماری ژنتیکی است که در نهایت زاییده اثرات عوامل محیطی است (۲، ۱). اگر سرطان ها در مراحل اول تشخیص داده شوند، قابل معالجه هستند (۳). سرطان خون یا لوسمی؛ بیماری پیشرونده و بد خیم اعضای خون ساز بدن است و یکی از مهم ترین سرطان هایی است که جامعه بشری با آن در گیر می باشد (۴). این بیماری در اثر تکثیر و تکامل ناقص گویچه های سفید خون و پیش سازهای آن در خون و مغز استخوان ایجاد می شود. در بیماری لوسمی، مغز استخوان به صورت غیر عادی، مقدار بسیار زیادی سلول خونی تولید می کند. این سلول ها با سلول های خون طبیعی متفاوت هستند و درست عمل نمی کنند. در نتیجه، تولید سلول های سفید خون طبیعی را متوقف کرده و توانایی فرد را در مقابله با بیماری ها از بین می برند. سلول های لوسمی بر تولید سایر انواع سلول های خونی که از مغز استخوان تولید می شوند مانند گلbulول های قرمز خون و پلاکت ها نیز تاثیر می گذارند (۴).

لوسمی نیز خود بر اساس طیف، شدت و سرعت پیشرفت روند بیماری به حاد و مزمن و نیز بر اساس نوع گلbulول سفید در گیر، به لنفوئیدی و میلوئیدی تقسیم می شود (۴، ۵). ۱- لوسمی میلوژنیک حاد = AML میلوسیت ها را تحت تاثیر قرار می دهد و روندی حاد دارد. در این بیماری، مغز استخوان، میلوبلاست ها، گلbulول های قرمز یا پلاکت های غیر طبیعی می سازد. ۲- لوسمی میلوژنیک مزمن (Chronic Myeloid Leukemia = CML) سلول های مغز استخوان یا میلوسیت ها را تحت تاثیر قرار می دارد. ۳- لوسمی لنفوسيتیک حاد = ALL سلول های لنفاوی یا لنفوسيت ها را تحت تاثیر قرار می دهد و روندی حاد دارد (۶). ۴- لوسمی لنفوسيتیک مزمن (CLL) سلول های لنفاوی (Chronic Lymphocytic Leukemia) لنفوسيت ها را تحت تاثیر قرار می دهد و روندی مزمن دارد.

می‌سازد. یکی از دقیق‌ترین روش‌ها برای کشف و پیش‌بینی بیماری لوسومی، استفاده از DNA افراد و اطلاعات ژنتیکی آن‌ها می‌باشد. تکنولوژی ریز آرایه، ابزاری برای بررسی بیان هزاران ژن در حداقل زمان ممکن است. هدف سیستم پیشنهادی، تشخیص هوشمند انواع لوسومی حاد با استفاده از مجموعه داده‌های ریز آرایه و روش‌های داده کاوی است.

شناخت داده‌ها:

مرحله شناخت داده‌ها شامل جمع‌آوری داده‌های اولیه، توصیف داده‌ها و بازرسی و بررسی داده‌ها است. در این مطالعه از داده‌های بیان ۷۱۲۹ ژن مربوط به ۷۲ بیمار مبتلا به لوسومی میلوژنیک و لنفوسيتیک حاد استفاده شد که با کمک فناوری ریز آرایه به دست آمد و توسط گلوب و همکاران ارایه گردیده است^(۷). هر بیمار با بر چسب لوسومی میلوژنیک حاد یا لوسومی لنفوسيتیک حاد (ALL) مشخص گردید که نشان‌دهنده نوع لوسومی در روی بود. از ۷۲ بیمار مذکور، ۴۷ بیمار مبتلا به لوسومی لنفوسيتیک حاد و ۲۵ بیمار مبتلا به لوسومی میلوژنیک حاد بودند. برای ارزیابی کارآیی و مقایسه الگوریتم‌های داده کاوی باید داده‌ها به دو دسته آزمون و آزمایشی تقسیم شوند و تمام الگوریتم‌ها با مجموعه آموزشی یکسانی آموزش داده شده و با مجموعه آزمون یکسانی مورد آزمایش قرار گیرند. داده‌های بیان ژن مورد استفاده در این مطالعه قبلاً به دو دسته داده‌های آموزشی و داده‌های آزمون تقسیم شدند. داده‌های آموزشی، بیان ژن ۳۸ بیمار (شامل ۲۷ بیمار مبتلا به لوسومی لنفوسيتیک حاد و ۱۱ بیمار مبتلا به لوسومی میلوژنیک حاد) و داده‌های آزمون بیان ژن ۳۴ بیمار (شامل ۲۰ بیمار مبتلا به لوسومی لنفوسيتیک حاد و ۱۴ بیمار مبتلا به لوسومی میلوژنیک حاد) را مشخص نمودند.

آماده‌سازی داده‌ها:

مرحله آماده‌سازی داده‌ها جهت بهبود کیفیت داده‌های واقعی برای داده کاوی لازم است و شامل انتخاب، پاک‌سازی، تبدیل داده‌ها و نرمال‌سازی داده‌ها است. انتخاب، تبدیل و تغییر شکل ویژگی‌ها، مهم‌ترین

تعداد ژن‌ها یا به عبارتی انتخاب ژن‌های متمایزکننده است و انجام این فرآیندها بدون کمک آنالیز آماری و روش‌های هوشمند تحلیل اطلاعات ممکن نیست^(۱۸). الگوریتم‌های مختلف داده کاوی و یادگیری ماشین (Machine learning) می‌توانند در خوشبندی و دسته‌بندی ژن‌ها مورد استفاده قرار گیرند و این روش‌ها کمک مؤثری در تصمیم‌گیری در مورد تشخیص بیماری‌ها و شیوه درمان، ارایه می‌دهند^(۴). به کمک پیشرفت‌های فناوری در بیوانفورماتیک و روش‌های مولکولی، داده‌های زیادی به دست آمده که در شناخت زودرس بیماری سرطان کمک خواهد کرد. هم چنین غربالگری به موقع برای بعضی از سرطان‌ها، کمک مؤثری در تشخیص زودرس آن می‌نماید^(۲). مطالعه‌های متعددی توسط محققان بر روی مجموعه داده‌های بیان ژن لوسومی با روش‌های مختلف انجام گرفته است^(۱۹-۲۱). با توجه به این که گرفتن تصمیم مناسب برای درمان انواع لوسومی از مهم‌ترین فعالیت‌ها بعد از تشخیص نوع سرطان است، هدف از انجام این مقاله، تشخیص لوسومی میلوژنیک و لنفوسيتیک حاد با استفاده از انتخاب ژن داده‌های ریز آرایه و الگوریتم‌های داده کاوی بود.

مواد و روش‌ها

مطالعه حاضر توصیفی و داده محور است و پایه اصلی آن داده کاوی و بررسی داده‌های بیان ژن لوسومی میلوژنیک و لنفوسيتیک حاد می‌باشد که با استفاده از فناوری ریز آرایه به وجود آمده است. روش‌های مختلفی برای پیاده‌سازی و اجرای پروژه‌های داده کاوی وجود دارد. در این مطالعه، مدلی جهت تشخیص لوسومی میلوژنیک و لنفوسيتیک حاد بر اساس متداول‌ترین CRISP ارائه شده که شامل فازهای شناخت سیستم، شناخت داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی و توسعه می‌باشد. در ادامه، مراحل مدل پیشنهادی شرح داده می‌شوند.

شناخت سیستم:

در مرحله شناخت سیستم، اهداف سیستم مورد نظر بررسی و مشخص می‌گردند. رشد گسترده لوسومی در جهان نیاز به سیستمی برای تشخیص آن را ضروری

پرداخته شد. برای مدل‌سازی از نرم‌افزار Matlab R2013a و ابزار داده‌کاوی Weka استفاده می‌شود و روش‌های جداگانه خطی، ماشین بردار پشتیبان (SVM-linear)، نزدیکترین همسایه، بیزین ساده، شبکه‌ی بیزین، نزدیکترین میانگین، رگرسیون لجستیک، شبکه‌ی عصبی پرسپترون چند لایه و درخت تصمیم J48 برای مدل‌سازی داده‌ها به کار می‌روند.

در ادامه روش‌های استفاده شده برای مدل‌سازی داده‌ها شرح داده می‌شوند:

- روش جداگانه خطی: روش جداگانه خطی فرض می‌کند که نمونه‌های یک کلاس به صورت خطی از نمونه‌های کلاس دیگر جداپذیرند. جدا بودن خطی نمونه‌های یک کلاس بدین معناست که بتوان با استفاده از یک رابطه خطی، نمونه‌های یک کلاس را از نمونه‌های کلاس دیگر جدا نمود.

- روش ماشین بردار پشتیبان (SVM): این روش با ساخت یک ابرسطح (که عبارت است از یک معادله خطی)، سعی دارد بهترین ابرسطحی را پیدا کند که با حداقل فاصله، داده‌های مربوط به دو کلاس را از هم تفکیک کند.

- روش k نزدیکترین همسایه (KNN): این روش یک روش دسته‌بندی است که تصمیم‌گیری در مورد این که یک نمونه جدید در کدام کلاس قرار گیرد با بررسی تعدادی (k) از شبیه‌ترین نمونه‌ها یا همسایه‌ها انجام می‌شود. این روش برای یافتن شباهت بین نمونه‌ها نیاز به یک معیار فاصله نظیر فاصله اقلیدسی یا فاصله منتهن دارد.

- روش بیزین ساده: این روش مبتنی بر قانون بیزین است و فرض می‌کند ویژگی‌ها از هم مستقل هستند. در روش بیزین ساده تنها نیاز است تا واریانس ویژگی‌ها به ازای هر کلاس محاسبه شود و نیازی به محاسبه ماتریس کوواریانس نیست.

- شبکه بیزین: شبکه بیزین یک گراف جهت‌دار غیر حلقوی است که از گره‌ها برای نمایش ژن‌ها و از کمان‌ها برای نمایش روابط احتمالی مابین ژن‌ها استفاده می‌کند. در این شبکه، x_i یک ژن است و گره‌های والد این ژن

موضوعاتی هستند که کیفیت یک راه حل داده کاوی را تعیین می‌کنند. در داده‌های به دست آمده توسط فناوری ریز آرایه که مربوط به بیان هزاران ژن هستند، یکی از مهم‌ترین موضوعات، کاهش و انتخاب ژن‌ها است. مسئله انتخاب ژن در واقع شناسایی و انتخاب یک زیر مجموعه مفید از ژن‌ها است که حداقل توان را در پیشگویی لوسومی میلوژنیک یا لوسومی لنفوسيتیک حاد دارا باشند. در مدل پیشنهادی این مطالعه برای انتخاب ژن‌ها از روش‌های انتخاب ویژگی زیر استفاده گردید.

۱- انتخاب ژن‌ها با استفاده از روش انتخاب ویژگی مبتنی بر ضریب همبستگی (Correlation Coefficient) داده‌ها با

داده تصمیم‌گیری (برچسب کلاس)

۲- انتخاب ژن‌ها با استفاده از روش انتخاب ویژگی بهره اطلاعاتی (Information Gain)

۳- انتخاب ژن‌ها با استفاده از روش انتخاب ویژگی نسبت بهره (Gain Ratio)

۴- انتخاب ژن‌ها با استفاده از روش انتخاب ویژگی امتیاز (Fisher Score) Fisher در مدل پیشنهادی با استفاده از روش‌های فوق به رتبه‌بندی ژن‌ها پرداخته و ژن‌های مهم با بالاترین رتبه انتخاب و ژن‌های دارای رتبه پایین حذف می‌شوند. معیار انتخاب زیر مجموعه ژن‌ها در مدل پیشنهادی مطابق رابطه زیر تعریف می‌شود:

$$(1) C = \frac{S_1 + S_2 + \dots + S_k}{S_1 + S_2 + \dots + S_k + \dots + S_d}$$

در این رابطه $[S_1, S_2, \dots, S_d]$ بردار مرتب شده رتبه‌بندی ژن‌ها به صورت نزولی است. از بردار مرتب شده رتبه ژن‌ها زیر مجموعه‌ای از ژن‌ها شامل k ژن انتخاب می‌شوند به طوری که جمع رتبه‌های آن‌ها (C)، درصد جمع رتبه‌های تمام ژن‌ها باشد.

در این مطالعه، سه مقدار $0/01$ ، $0/02$ و $0/03$ برای پارامتر C در نظر گرفته شدند.

مدل‌سازی:

در مرحله مدل‌سازی با استفاده از الگوریتم‌های مختلف داده کاوی، به مدل‌سازی داده‌ها و پیدا کردن مدل بهینه

با (Parent(xi) نشان داده می‌شوند و توزیع احتمال توأم مجموعه‌ای از ژن‌ها محاسبه می‌گردد.

- روش نزدیک‌ترین میانگین: این روش بر اساس قانون بیزین است و فرض می‌کند ویژگی‌ها از هم مستقل هستند. روش نزدیک‌ترین میانگین فرض می‌کند که واریانس همه کلاس‌ها و هم چنین احتمال‌های پیشین تمام کلاس‌ها مساوی هستند و نمونه جدید را به کلاسی با نزدیک‌ترین میانگین اختصاص می‌دهد.

- رگرسیون لجستیک: رگرسیون لجستیک یکی از مدل‌های خطی تعییم یافته است که برای تحلیل رابطه یک یا چند متغیر اسمی بر متغیر پاسخ رسته‌ای به کار می‌رود. رگرسیون لجستیک، شبیه رگرسیون خطی است با این تفاوت که نحوه محاسبه ضرایب در این دو روش یکسان نمی‌باشد. رگرسیون لجستیک، به جای حداقل کردن مجدد خطاها، احتمال وقوع یک واقعه را حداقل کردن می‌کند. رگرسیون لجستیک از آماره‌های کای اسکوئر (χ^2) و والد استفاده می‌کند.

- شبکه‌های عصبی پرسپترون چند لایه: شبکه‌های عصبی مصنوعی از یک سری لایه‌ها شامل اجزای ساده‌ای به نام نرون تشکیل شده‌اند که همانگ با هم برای حل مسائل به کار می‌روند. شبکه‌های عصبی پرسپترون از چند لایه شامل لایه ورودی، لایه‌های پنهان و لایه خروجی تشکیل شده است. در شبکه عصبی پرسپترون چند لایه، هر نرون در هر لایه به تمام نرون‌های لایه قبل متصل است. لایه ورودی، یک لایه انتقال دهنده و لایه خروجی شامل مقادیر پیش‌بینی شده به وسیله شبکه است و لایه‌های پنهان که از نرون‌های پردازش‌گر تشکیل شده‌اند و محل پردازش داده‌ها هستند.

- درخت تصمیم J48: درخت تصمیم، ساختاری شبیه به فلوچارت دارد که بالاترین گره، ریشه درخت است و گره‌های برگ، دسته‌ها یا توزیع دسته‌ها را نشان می‌دهند. درخت تصمیم با مرتب کردن نمونه‌ها در درخت از گره ریشه به سمت گره‌های برگ آن‌ها را دسته‌بندی می‌کند. الگوریتم J48، درخت تصمیم C4.5 است که توسط نرم‌افزار Weka ارایه می‌شود و از مفهوم آنتروپی اطلاعات استفاده می‌کند.

ارزیابی:
در این مرحله به ارزیابی نتایج حاصل از مدل‌سازی با استفاده از شاخص‌های دقت، حساسیت و اختصاصیت پرداخته می‌شود. میزان دقت یک روش دسته‌بندی بر روی مجموعه داده‌های آزمون، درصد مشاهداتی از مجموعه آزمون است که به درستی توسط مدل مورد استفاده دسته‌بندی شده است. حساسیت عبارت است از میزانی برای مشخص کردن توانایی سیستم در تشخیص و دسته‌بندی بیماران مبتلا به لوسمی میلوژنیک حاد که سیستم آن‌ها را به صورت صحیح دسته‌بندی می‌نماید. اختصاصیت عبارت است از میزانی برای مشخص کردن توانایی سیستم در تشخیص و دسته‌بندی بیماران مبتلا به لوسمی لنفوستیک حاد که سیستم آن‌ها را به صورت صحیح لوسمی لنفوستیک حاد تشخیص می‌دهد.

توسعه:

در مرحله توسعه، با توجه به نتایج به دست آمده در مرحله ارزیابی، مدلی که دارای عملکرد مناسبی است برای دسته‌بندی داده‌ها به کار می‌رود.

یافته‌ها

در این مطالعه، سه مقدار ۰/۰۱، ۰/۰۲ و ۰/۰۳، برای پارامتر C در نظر گرفته می‌شوند. تعداد ژن‌های انتخاب شده توسط روش‌های انتخاب ژن با مقادیر مختلف پارامتر C در جدول ۱ نشان داده شده است.

همان گونه که در جدول ۱ مشخص شده است، تعداد ژن‌ها با استفاده از تمام روش‌های انتخاب ژن به طور قابل توجهی کاهش یافته است. پس از انتخاب ژن‌ها، روش‌های جداکننده خطی، نزدیک‌ترین میانگین، ماشین‌بردار پشتیبان (SVM-Linear)، روش k نزدیک‌ترین همسایه، شبکه عصبی بیزین، بیزین ساده، رگرسیون لجستیک، شبکه عصبی پرسپترون چند لایه و درخت تصمیم J48 بر روی این داده‌ها اجرا می‌گردند. در روش‌های دسته‌بندی ذکر شده، بهترین ژن‌ها در هریک از روش‌های انتخاب ژن (مقدار بهینه پارامتر C) با استفاده از روش اعتبارسنجی متقاطع با ده تکرار بر روی مجموعه آموزشی به دست آمده است.

جدول ۱: تعداد ژن‌های انتخاب شده توسط روش‌های انتخاب ژن

تعداد ژن‌های انتخاب شده	مقدار پارامتر C	روش انتخاب ژن
۳۸	۰/۰۱	ضریب همبستگی
۴۵	۰/۰۱	بهره اطلاعاتی
۳۹	۰/۰۱	نسبت بهره
۱۰	۰/۰۱	Fisher امتیاز
۷۷	۰/۰۲	ضریب همبستگی
۸۷	۰/۰۲	بهره اطلاعاتی
۸۶	۰/۰۲	نسبت بهره
۲۴	۰/۰۲	Fisher امتیاز
۱۱۶	۰/۰۳	ضریب همبستگی
۱۳۳	۰/۰۳	بهره اطلاعاتی
۱۲۳	۰/۰۳	نسبت بهره
۴۱	۰/۰۳	Fisher امتیاز

مقدار بهینه پارامتر k در روش k نزدیک‌ترین همسایه نیز با استفاده از روش اعتبارسنجی متقطع با ده تکرار بر روی مجموعه آموزشی به دست آمد. سپس آزمایش‌ها را با استفاده از ژن‌های انتخاب شده بر روی مجموعه داده‌های آزمون انجام دادیم. نتایج بررسی روش‌های گوناگون دسته‌بندی با استفاده از شاخص‌های دقت، حساسیت و اختصاصیت با استفاده از روش انتخاب ژن ضریب همبستگی بر روی مجموعه داده‌های آزمون نشان داده شد (جدول ۲).

همان گونه که در جدول ۲ مشاهده می‌شود، روش ماشین‌بردار پشتیبان با استفاده از ۷۷ ژن انتخاب شده توسط ضریب همبستگی، دارای عملکرد بهتری در مقایسه با سایر روش‌ها است و با دقت بالایی قادر به تشخیص لوسومی میلورئیک و لغوسیتیک حاد است. روش درخت تصمیم J48 و رگرسیون لجستیک، دارای عملکرد نسبتاً ضعیفی در دسته‌بندی انواع لوسومی حاد می‌باشد.

جدول ۲: نتایج عملکرد روش‌های دسته‌بندی بر روی ژن‌های انتخاب شده توسط ضریب همبستگی

نام روش	تعداد ژن	دقت	حساسیت	اختلاف
جاداکنده خطی	۷۷	٪۶۴/۷۱	٪۲۰	٪۸۳/۲۲
نزدیک‌ترین میانگین	۳۸	٪۳۸/۲۴	٪۱۰۰	٪۱۲/۵۰
ماشین‌بردار پشتیبان	۷۷	٪۹۷/۰۶	٪۱۰۰	٪۹۵/۸۳
(k=1) نزدیک‌ترین همسایه	۳۸	٪۹۴/۱۲	٪۱۰۰	٪۹۱/۶۷
شبکیه بیزین	۳۸	٪۳۸/۲۴	٪۱۰۰	٪۱۲/۵۰
بیزین ساده	۳۸	٪۹۴/۱۲	٪۱۰۰	٪۹۱/۶۷
رگرسیون لجستیک	۱۱۶	٪۲۶/۴۷	٪۵۰	٪۱۶/۶۷
شبکه عصبی پرسپترون	۳۸	٪۹۱/۱۸	٪۱۰۰	٪۸۷/۵۰
درخت تصمیم J ۴۸	۱۱۶	٪۲۶/۴۷	٪۹۰	٪۰

جدول ۳: نتایج عملکرد روش‌های دسته‌بندی بر روی ژن‌های انتخاب شده توسط بهره اطلاعاتی

نام روش	تعداد ژن	دقت	حساسیت	اختلاف
جاداکنده خطی	۸۷	٪۷۳/۵۳	٪۲۰	٪۹۵/۸۳
نزدیک‌ترین میانگین	۴۵	٪۹۷/۰۶	٪۱۰۰	٪۹۵/۸۳
ماشین‌بردار پشتیبان	۸۷	٪۱۰۰	٪۱۰۰	٪۱۰۰
(k=۵) نزدیک‌ترین همسایه	۴۵	٪۹۷/۰۶	٪۱۰۰	٪۹۵/۸۳
شبکه بیزین	۴۵	٪۳۸/۲۴	٪۱۰۰	٪۱۲/۵۰
بیزین ساده	۴۵	٪۹۷/۰۶	٪۱۰۰	٪۹۵/۸۳
رگرسیون لجستیک	۴۵	٪۴۴/۱۲	٪۱۰۰	٪۲۰/۸۳
شبکه عصبی پرسپترون	۱۲۳	٪۱۰۰	٪۱۰۰	٪۱۰۰
درخت تصمیم J ۴۸	۴۵	٪۳۲/۳۵	٪۱۰۰	٪۴/۱۷

جدول ۴: نتایج عملکرد روش‌های دسته‌بندی بر روی ژن‌های انتخاب شده توسط نسبت بهره

نام روش	تعداد ژن	دقت	حساسیت	اختصاصیت
جاداکننده خطی	۳۹	%۲۰/۵۹	%۴۰	%۱۲/۵۰
نzdیکترین میانگین	۳۹	%۱۰۰	%۱۰۰	%۱۰۰
ماشینبردار پشتیبان	۳۹	%۱۰۰	%۱۰۰	%۱۰۰
(k = ۵) نزدیکترین همسایه	۳۹	%۱۰۰	%۱۰۰	%۱۰۰
شبکه بیزین	۸۶	%۹۷/۰۶	%۱۰۰	%۹۵/۸۳
بیزین ساده	۳۹	%۱۰۰	%۱۰۰	%۱۰۰
رگرسیون لجستیک	۳۹	%۹۷/۰۶	%۱۰۰	%۹۵/۸۳
شبکه عصبی پرسپترون	۳۹	%۱۰۰	%۱۰۰	%۱۰۰
درخت تصمیم J ۴۸	۳۹	%۳۲/۳۵	%۱۰۰	%۴/۱۷

جدول ۵: نتایج عملکرد روش‌های دسته‌بندی بر روی ژن‌های انتخاب شده توسط نسبت امتیاز Fisher

نام روش	تعداد ژن	دقت	حساسیت	اختصاصیت
جاداکننده خطی	۱۰	%۴۴/۱۲	%۹۰	%۲۵
نzdیکترین میانگین	۲۴	%۹۴/۱۲	%۱۰۰	%۹۱/۶۷
ماشینبردار پشتیبان	۴۱	%۹۷/۰۶	%۱۰۰	%۹۵/۸۳
(k = ۱) نزدیکترین همسایه	۴۱	%۹۴/۱۲	%۱۰۰	%۹۱/۶۷
شبکه بیزین	۴۱	%۴۱/۱۸	%۱۰۰	%۱۶/۶۷
بیزین ساده	۲۴	%۹۷/۰۶	%۱۰۰	%۹۵/۸۳
رگرسیون لجستیک	۲۴	%۴۱/۱۸	%۱۰۰	%۱۶/۶۷
شبکه عصبی پرسپترون	۱۰	%۴۱/۱۸	%۱۰۰	%۱۶/۶۷
درخت تصمیم J ۴۸	۱۰	%۳۲/۳۵	%۱۰۰	%۴/۱۷

بر روی مجموعه داده‌های آزمون را نشان می‌دهد. همان گونه که در جدول ۴ نشان داده شده است، روش‌های نزدیکترین میانگین، ماشینبردار پشتیبان، k نزدیکترین همسایه، بیزین ساده و شبکه عصبی پرسپترون چند لایه با استفاده از ژن‌های انتخاب شده توسط نسبت بهره با دقتهای %۱۰۰ قادر به تشخیص لوسمی میلوژنیک و لنفوسیتیک حاد هستند.

روش‌های شبکه بیزین و رگرسیون لجستیک نیز دارای عملکرد خوبی در تشخیص انواع لوسمی حاد هستند. نتایج عملکرد روش‌های دسته‌بندی گوناگون با استفاده از روش انتخاب ژن بر روی مجموعه داده‌های آزمون نشان داده شد (جدول ۵).

نتایج عملکرد روش‌های دسته‌بندی گوناگون بر روی ژن‌های انتخاب شده، توسط روش بهره اطلاعاتی بر روی مجموعه داده‌های آزمون به دست آمد (جدول ۳).

نتایج جدول ۳ نشان می‌دهد که روش‌های ماشینبردار پشتیبان و شبکه عصبی پرسپترون چند لایه با استفاده از ژن‌های مناسب انتخاب شده، توسط بهره اطلاعاتی با دقتهای ۱۰۰ قادر به تشخیص انواع لوسمی حاد هستند. روش‌های نزدیکترین میانگین، k نزدیکترین همسایه و بیزین ساده نیز دارای عملکرد خوبی در تشخیص لوسمی میلوژنیک و لنفوسیتیک حاد هستند.

جدول ۴، نتایج عملکرد روش‌های دسته‌بندی گوناگون بر روی ژن‌های انتخاب شده توسط روش نسبت بهره

جدول ۶: ۳۹ ژن انتخاب شده توسط معیار نسبت بهره

شماره الحاق ژن	توصیف ژن	شماره الحاق ژن	توصیف ژن
D88422_at	CYSTATIN A	D14874_at	ADM Adrenomedullin
J04970_at	CPM Carboxypeptidase M	J04615_at	SNRPN Small nuclear ribonucleoprotein polypeptide N
J05243_at	SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)	J04990_at	CATHEPSIN G PRECURSOR
L47738_at	Inducible protein mRNA	L11669_at	Tetracycline transporter-like protein mRNA
M27891_at	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M19507_at	MPO Myeloperoxidase
M31166_at	PTX3 Pentraxin-related gene, rapidly induced by IL-1 beta	M29540_at	CARCINOEMBRYONIC ANTIGEN PRECURSOR
M54995_at	PPBP Connective tissue activation peptide III	M31994_at	ALDH1 Aldehyde dehydrogenase 1, soluble
M84526_at	DF D component of complement (adipsin)	M55150_at	FAH Fumarylacetoacetate
M96326_rna1_at	Azurocidin gene	M92287_at	CCND3 Cyclin D3
U50136_rna1_at	Leukotriene C4 synthase (LTC4S) gene	U02020_at	Pre-B cell enhancing factor (PBEF) mRNA
U82759_at	GB DEF = Homeodomain protein HoxA9 mRNA	U57094_at	Small GTP-binding protein mRNA
X70297_at	CHRNA7 Cholinergic receptor, nicotinic, alpha polypeptide 7	X66401_cds1_at	LMP2 gene extracted from H.sapiens genes TAP1, TAP2, LMP2, LMP7 and DOB
X95735_at	Zyxin	X90872_at	Gp25L2 protein
Y12670_at	LEPR Leptin receptor	Y00433_at	GPX1 Glutathione peroxidase 1
X68688_rna1_s_at	ZNF33B gene	J02783_at	P4HB Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), beta polypeptide (protein disulfide isomerase; thyroid hormone binding protein p55)
X05130_s_at	P4HB Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), beta polypeptide (protein disulfide isomerase; thyroid hormone binding protein p55)	X06182_s_at	KIT V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
M12959_s_at	TCRA T cell receptor alpha-chain	L09209_s_at	APLP2 Amyloid beta (A4) precursor-like protein 2
M31211_s_at	MYL1 Myosin light chain (alkali)	M27783_s_at	ELA2 Elastatse 2, neutrophil
X58431_rna2_s_at	HOX 2.2 gene extracted from Human Hox2.2 gene for a homeobox protein	X85116_rna1_s_at	Epb72 gene exon 1
		M31523_at	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
D88422_at	CYSTATIN A	D14874_at	ADM Adrenomedullin
J04970_at	CPM Carboxypeptidase M	J04615_at	SNRPN Small nuclear ribonucleoprotein polypeptide N

سرطان بودند. همین محققان در سال ۲۰۱۱ با روش MTSVSL قادر به تشخیص انواع لوسمی با دقت ۹۶/۹٪ شدند(۲۱). کای و همکاران در سال ۲۰۱۴ برای تشخیص لوسمی از روش I-RELIEF-NB استفاده کردند و با دقت ۹۱/۶٪ قادر به تشخیص انواع لوسمی شدند. این محققان در همان سال با استفاده از روش RELIEF-KNN برای تشخیص سرطان لوسمی به دقت ۹۴/۴٪ دست یافتند(۲۲). هنگ و همکاران در سال ۲۰۱۲ با استفاده از روش BMSF-NB به تشخیص انواع لوسمی پرداختند و با دقت ۹۶/۲۵٪ قادر به تشخیص لوسمی ALL از AML شدند. همین محققان از روش Gene SrF-NB استفاده نمودند و قادر به تشخیص لوسمی با دقت ۹۴/۵۸٪ شدند(۲۳). آزادی و همکاران در مطالعه با استفاده از داده‌های بیان ژن و آزمایش‌های آماری، ژن‌های مسئول لوسمی حاد را تشخیص دادند و در پایان مطالعه گزارش کردند که شناخت این ژن‌ها جهت درمان و حتی پیشگیری از آن می‌تواند بسیار مهم و حائز اهمیت باشد. هم چنین این محققان در مطالعه خود گزارش کردند با اطلاع از نحوه بیان این ژن‌ها در افراد مبتلا، پزشکان قادر خواهند بود که با تجویز داروها و روش‌های درمانی مناسب، میزان بیان آن‌ها را کنترل نمایند و باعث کاهش مرگ و میر ناشی از این نوع بیماری‌ها شود(۲۴).

نتیجه‌گیری

نتایج این مطالعه نشان داد که انتخاب ژن‌ها و الگوریتم‌های داده کاوی قادر به تشخیص انواع لوسمی با دقت بسیار بالایی هستند، بنابراین با استفاده از تکنولوژی ریزآرایه و الگوریتم‌های داده کاوی با تشخیص دقیق انواع لوسمی، می‌توان تصمیمات مناسبی در مورد نحوه تشخیص و درمان بیماران گرفت.

نتایج حاصل از ارزیابی روش‌های گوناگون دسته‌بندی نشان می‌دهد که روش ماشین‌بردار پشتیبان با استفاده از تمام روش‌های انتخاب ژن، دارای عملکرد بالایی در تشخیص انواع لوسمی حاد است.

روش‌های نزدیک‌ترین میانگین، ماشین‌بردار پشتیبان، k نزدیک‌ترین همسایه، بیزین ساده و شبکه عصبی پرسپترون چند لایه با استفاده از ۳۹ ژن انتخاب شده توسط نسبت بهره با دقت ۱۰۰٪، قادر به تشخیص لوسمی میلوژنیک و لفوسیتیک حاد هستند(جدول ۶). هم چنین روش ماشین‌بردار پشتیبان با استفاده از ۸۷ ژن انتخاب شده توسط بهره اطلاعاتی و روش شبکه عصبی پرسپترون چند لایه با استفاده از ۱۳۳ ژن انتخاب شده، توسط بهره اطلاعاتی با دقت ۱۰۰٪ قادر به تشخیص لوسمی میلوژنیک و لفوسیتیک حاد هستند. روش‌های درخت تصمیم ۴۸ و جداکننده خطی با استفاده از ژن‌های انتخاب شده توسط تمام روش‌های انتخاب ژن دارای عملکرد ضعیفی هستند.

بحث

در این مطالعه داده‌های حاصل از ریزآرایه بیماری لوسمی توسط روش‌های نزدیک‌ترین میانگین، ماشین‌بردار پشتیبان، k نزدیک‌ترین همسایه، بیزین ساده و شبکه عصبی پرسپترون چند لایه با استفاده از ۳۹ ژن انتخاب شده توسط نسبت بهره با دقت ۱۰۰٪ قادر به تشخیص لوسمی میلوژنیک و لفوسیتیک حاد بودند. هم چنین روش ماشین‌بردار پشتیبان با استفاده از ۸۷ ژن انتخاب شده توسط بهره اطلاعاتی و روش شبکه عصبی پرسپترون چند لایه با استفاده از ۱۳۳ ژن انتخاب شده توسط بهره اطلاعاتی با دقت ۱۰۰٪ قادر به تشخیص لوسمی میلوژنیک و لفوسیتیک حاد هستند. لین و چن با روش شبکه عصبی BP به بررسی مجموعه داده‌های بیان ژن لوسمی در سال ۲۰۱۱ پرداختند و با دقت ۹۵/۸۳٪ قادر به تشخیص انواع

References :

- 1- Sheikhpour R, Mohiti Ardekani J. The effect of progesterone on p53 protein in T47D in cell line. *Urmia Med J* 2014; 25(10): 954-60.
- 2- Parsa N. Environmental Factors, Genes and Human Cancers. *Sci Cultivation J* 2012; 2(1): 12-9.
- 3- Sheikhpour R, Ghasemi N, Yaghmaei P, Mohiti J. Immunohistochemical assessment of p53 protein and its correlation with clinicopathological parameters in breast cancer patients. *Indian J Sci Technol* 2014; 7(4): 472-9.
- 4- Toloie Ashlagi A, Mohsen Taheri S. Designing an expert system for suggesting the blood cancer treatment. *J Health Admin* 2010; 13(40): 41-50. [Article in Farsi]
- 5- Mahmood Abadi A. Lukemia. 1st ed. Tehran: Kerdgari Publication; 2007.p.1-56. [Persian]
- 6- Heidari M, Hajigholami A. Acute lymphocytic leukemia with severe eosinophilia (a case report). *J Shahrekh Univ Med Sci* 2013; 15(5): 111-5. [Article in Farsi]
- 7- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286(15): 530-8.
- 8- Zali H, Rezaei Tavirani M, Salimian J, Aolad GR, Basataminejad S. Gene expression networks to analysis DNA microarray data. *Scientific Journal of Ilam University of Medical Sciences* 2012; 20(4): 138-50. [Article in Farsi]
- 9- Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci USA* 2000; 97(22): 12079-84.
- 10- Hoyle DC, Rattray M, Jupp R, Brass A. Making sense of microarray data distributions. *Bioinformatics* 2002; 18(4): 576-84.
- 11- Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J Comput Biol* 2002; 7(6): 819-37.
- 12- Long AD, Mangalam HJ, Chan BY, Toller L, Hatfield G, Baldi P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J Biol Chem* 2001; 276(23): 19937-44.
- 13- Koga Y, Yamazaki N, Takizawa S, Kawauchi J, Nomura O, Yamamoto S, et al. Gene expression analysis using a highly sensitive DNA microarray for colorectal cancer screening. *Anticancer Res* 2014; 34(1): 169-74.
- 14- Vassella E, Galván JA, Zlobec I. Tissue microarray technology for molecular applications: investigation of cross-contamination between tissue samples obtained from the same punching device. *Microarrays* 2015; 4(2): 188-95.
- 15- Kohbalan M, Mohd SM, Safaai D. A review on missing value imputation algorithms for microarray gene expression data. *Current Bioinformatics* 2014; 9(1): 18-22.
- 16- Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. *Nat Genet* 1999; 21 (1 Suppl): 10-4.
- 17- Molaeenezadeh SF, Moradi MH. Selected genes containing microarray information using mutual information and genetic algorithm. *13th Conf Med Eng*; 2006; 1-5.
- 18- Joroughi M, Shamsi M, Saberkari HR, Sedaaghi MH, Momennezhad A. Gene selection and cancer classification based on microarray data using combined BPSO and BLDA algorithm. *Computational Intelligence Electrical Engineering* 2014; 5(2): 29-47. [Article in Farsi]
- 19- Wang Z. Neuro-fuzzy modeling for microarray cancer gene expression data. USA: Oxford University Computing Laboratory; 2005. p. 241-6.
- 20- Yu L, Liu H. Redundancy based feature selection for microarray data. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM 2004; 737-42.
- 21- Chen AH, Lin EJ. The prediction of cancer classification using a novel multi-task support vector sample learning technique. *AISS: Adv Inform Sci Serv Sci* 2011; 3(3): 92-9.
- 22- Cai H, Ruan P, Ng M, Akutsu T. Feature weight estimation for gene selection: a local hyperlinear learning approach. *BMC Bioinformatics* 2014; 15(1), 70.
- 23- Zhang H, Wang H, Dai Z, Chen M.S., Yuan Z. Improving accuracy for cancer classification with a new algorithm for genes selection. *BMC Bioinformatics* 2012; 13(1), 298.
- 24- Azadi NA, Nouri-Jaliani K, Taheri-Kalani M. Identifying differentially expressed genes based on their expressions in leukemia. *Koomesh* 2005; 6(4): 259-64. [Article in Farsi]

Original Article

Diagnosis of acute myeloid and lymphoblastic leukemia using gene selection of microarray data and data mining algorithm

Sheikhpour R.¹, Aghasaram M.¹, Sheikhpour R.^{2,3}

¹School of Electrical & Computer Engineering, Yazd University, Yazd, Iran

²School of Medicine, Islamic Azad University, Yazd Branch, Yazd, Iran

³Hematology & Oncology Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

Abstract

Background and Objectives

Microarray technology represents the expression of thousands of genes simultaneously. Microarray analysis may not be possible without statistical analysis and artificial intelligence methods. The aim of this paper is to diagnose acute leukemia using microarray data and data mining algorithms.

Materials and Methods

The expression of 7129 genes of 72 patients with leukemia was used in this study. Then, by the selection of important genes based on correlation coefficient, information gain, gain ratio and fisher score criteria and by the use of linear discriminant, support vector machine, k nearest neighbor, naïve Bayes, Bayes net, nearest mean, logistic regression, multilayer perceptron neural network and J48 decision tree methods on the selected genes, acute myeloid and lymphoblastic leukemia were attempted to be diagnosed.

Results

The methods of nearest mean, support vector machine, k nearest neighbor, naïve Bayes, and multilayer perceptron neural network are able to detect acute myeloid and lymphoblastic leukemia using 39 selected genes by the gain ratio with 100 percent accuracy. Moreover, support vector machine method using 87 selected genes by information gain and support vector machine method using 133 selected genes by information gain are able to detect acute myeloid and lymphoblastic leukemia with 100 percent accuracy.

Conclusions

The results of this study showed that gene selection and data mining algorithm are able to diagnose leukemia with high accuracy. Therefore, appropriate decisions can be made using these methods about the how of the diagnosis and treatment of patients.

Key words: Acute Lymphoid Leukemia, Acute Myeloid Leukemia, Microarray Analysis, Data Mining

Received: 7 Jan 2015

Accepted: 12 Jul 2015

Correspondence: Sheikhpour R., PhD of Biochemistry. School of Medicine, Islamic Azad University, Yazd Branch and Hematology & Oncology Research Center, Shahid Sadoughi University of Medical Sciences. P.O.Box: 89156-56965, Yazd, Iran. Tel: (+9835) 36282884; Fax: (+9835) 36282884 E-mail: robab.sheikhpour@iauyazd.ac.ir