خون
فصلنامه پژوهشی
دوره ۱۳ شماره ۳ پاییز ۹۵ (۲۱۴-۲۰۷)

مقاله پژوهشی

روشی جدید برای تشخیص لوسمی میلوژنیک و لنفوسیتیک حاد با استفاده از داده‌های بیان ژن و روش‌های یادگیری ماشین

رباب شیخ‌پور[۱]، راضیه شیخ‌پور[۲]، مهدی آقا صرام[۳]

چکیده

سابقه و هدف

لوسمی از سرطان‌های شایع در جهان است. یکی از مهم‌ترین روش‌ها برای کشف و پیش‌بینی لوسمی میلوژنیک و لنفوسیتیک حاد، استفاده از DNA افراد و اطلاعات ژنتیکی آن‌ها می‌باشد. تکنولوژی ریز آرایه، ابزاری برای بررسی بیان هزاران ژن در حداقل زمان است. تحلیل مجموعه داده‌های ریز آرایه بدون کمک آنالیز آماری و روش‌های یادگیری ماشین ممکن نیست. در این مطالعه با استفاده از مجموعه داده‌های ریز آرایه و روش‌های یادگیری ماشین به تشخیص انواع لوسمی پرداخته شد.

مواد و روش‌ها

داده‌های مورد استفاده در این پژوهش توصیفی، بیان ۷۱۲۹ ژن مربوط به ۷۲ بیمار مبتلا به لوسمی بود که با استفاده از فناوری ریز آرایه به دست آمد. سپس با استفاده از این داده‌ها، تشخیص لوسمی میلوژنیک حاد(AML) و لوسمی لنفوسیتیک حاد(ALL) با روش طبقه‌بندی ناپارامتری هسته، تابع پایه شعاعی ناهمسانگرد با استفاده از معیارهای نسبت بهره و بهره اطلاعاتی انجام شد.

یافته‌ها

روش پیشنهادی طبقه‌بندی ناپارامتری با استفاده از معیار بهره اطلاعاتی با انتخاب ۲۳۰ ژن مهم و با استفاده از معیار نسبت بهره با انتخاب ۸۶ ژن مهم با دقت ۹۷/۰۶٪ ، قادر به تشخیص انواع لوسمی‌میلوژنیک و لنفوسیتیک است، در حالی‌که روش طبقه‌بندی ناپارامتری هسته، تابع پایه شعاعی با ۷۱۲۹ ژن دارای دقت ۳۵/۲۹٪ است.

نتیجه گیری

نتایج این مطالعه نشان داد که استفاده از داده‌های بیان ژن و روش پیشنهادی با معیار نسبت بهره قادر به تشخیص لوسمی با دقت بالایی است. بنابراین به نظر می‌رسد این روش می‌تواند در تشخیص دقیق‌تر انواع لوسمی کمک کند تا تصمیمات مناسب‌تری در مورد نحوه تشخیص و درمان بیماران گرفته شود.

کلمات کلیدی: لوسمی، بیان ژن، آنالیز ریز آرایه، یادگیری ماشین

تاریخ دریافت: ۹۴/۱۰/۵
تاریخ پذیرش: ۹۵/۲/۲۲

۱- مؤلف مسئول: PhD بیوشیمی، گروه تربیت بدنی، واحد تفت، دانشگاه آزاد اسلامی، تفت، ایران و مرکز تحقیقات خون و انکولوژی، دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران، صندوق پستی: ۵۶۹۶۵-۸۹۱۵۶
۲- دانشجوی دکترای کامپیوتر ـ گروه مهندسی کامپیوتر ـ دانشگاه یزد ـ یزد ـ ایران
۳- دکترای تخصصی فناوری تست سیستم‌ها ـ دانشیار گروه مهندسی کامپیوتر ـ دانشگاه یزد ـ یزد ـ ایران

مقدمه

سرطان بیماری است که در نتیجه تقسیم غیر قابل کنترل سلول‌ها به وجود می‌آید(۱). امروزه بیش از ۱۰۰ نوع مختلف از سرطان‌ها در دنیا شناخته شده‌اند و لوسمی یکی از انواع شایع و مهلک این سرطان‌ها است(۲). لوسمی ۸٪ کل سرطان‌های جمعیت انسانی را شامل و به عنوان پنجمین سرطان شایع در جهان شناخته شده است(۳). علت دقیق ابتلا به لوسمی مشخص نیست و پژوهش‌های انجام شده روند بدخیمی بیماری لوسمی را به ژنتیک، قرارگرفتن در معرض پرتوهای یونیزه کننده و برخی مواد شیمیایی خاص و یا نارسایی سیستم ایمنی طبیعی بدن ارتباط می‌دهند(۶-۴). سرطان خون یا لوسمی، بیماری پیشرونده و بدخیم اعضای خون‌ساز بدن به ویژه مغز استخوان است که با تکثیر و تکامل ناقص سلول‌های خون و پیش‌سازهای آن در خون و مغز استخوان ایجاد می‌شود(۲). سلول‌های سفید خونی معمولاً در صورت نیاز بدن، به طریقی منظم و کنترل شده رشد کرده و تقسیم می‌شوند. اما بیماری لوسمی در این روند اخلال ایجاد نموده و رشد سلول‌های خونی را از کنترل خارج می‌نماید. در بیماری لوسمی حاد، مغز استخوان مقدار بسیار زیادی سلول‌های سفید خونی نارس تولید می‌کند و تولید طبیعی سلول‌های سفید خونی نیز متوقف می‌شود که منجر به از بین رفتن توانایی بدن در مقابله با بیماری‌ها می‌شود(۳). دو نوع اصلی حاد از این بیماری وجود دارد که عبارتند از لوسمی میلوئیدی حاد Acute Myeloid Leukemia (AML) و لوسمی لنفوسیتی حاد Acute Lymphoblastic Leukemia (ALL)(۸، ۷). تشخیص لوسمی میلوئیدی حاد از لوسمی لنفوسیتی حاد برای درمان موفق، حیاتی است(۹). یکی از دقیق‌ترین و مهم‌ترین روش‌ها برای کشف این بیماری و پیش‌بینی آن، استفاده از DNA افراد و اطلاعات ژنتیکی آن‌ها می‌باشد. تکنولوژی ریز آرایه(DNA Microarray)، برای مطالعه سریع ژن‌ها به وجود آمده است و یک تصویر کلی از میزان بیان ژن را ارایه می‌دهد(۲). این تکنولوژی در تنظیم و تعاملات ژن‌ها و پژوهش‌های بالینی و دارویی کاربرد دارد و برخلاف روش‌های قبلی که تنها مطالعه یک ژن را میسر می‌نمود، امکان بررسی بیان هزاران ژن را در حداقل زمان ممکن فراهم می‌کند(۱۶-۱۰). بنابراین تشخیص دقیق سرطان می‌تواند با طبقه‌بندی داده‌های ریز آرایه عملی باشد(۱۷). مشکل اصلی در تحلیل داده‌های ریز آرایه، بعد بالای آن‌ها است که در نتیجه تعداد بسیار زیاد متغیرها(ژن‌ها) در مقابل تعداد کم نمونه‌ها ایجاد می‌شود. اگر چه تعداد بسیار زیادی از ژن‌ها در داده‌های ریز آرایه وجود دارند، تنها بخش اندکی از آن‌ها تاثیر به‌سزایی در صحت طبقه‌بندی می‌گذارند. از این‌رو، اولین قدم مهم در آنالیز داده‌های ریز آرایه، کاهش تعداد ژن‌ها یا به عبارتی، انتخاب ژن‌های متمایزکننده به‌منظور طبقه‌بندی است(۱۲). انتخاب ژن‌های مرتبط و تفسیر این اطلاعات بدون کمک آنالیز آماری و روش‌های هوشمند تحلیل اطلاعات ممکن نیست. یادگیری ماشین(Machine learning)، شاخه‌ای از هوش مصنوعی (Artificial intelligence) است که با طرح و به کارگیری الگوریتم‌ها به کامپیوترها این امکان را می‌دهد که کارآیی خود را بر اساس یادگیری، بهینه نمایند. الگوریتم‌های مختلف داده کاوی(Data mining) و یادگیری ماشین(Machine learning) می‌توانند در خوشه‌بندی و طبقه‌بندی ژن‌ها مورد استفاده قرار گیرند. هدف از انجام این مطالعه، تشخیص انواع لوسمی ALL و AML با استفاده از مجموعه داده‌های ریز آرایه و روش‌های یادگیری ماشین بود.

مواد و روش‌ها

مطالعه حاضر توصیفی و داده محور بود که به ارایه روشی برای تشخیص لوسمی میلوژنیک حاد(AML) و لوسمی لنفوسیتیک حاد(ALL) با استفاده از داده‌های بیان ژن بیماران لوسمی میلوژنیک و لنفوسیتیک حاد پرداخته است.

توصیف مجموعه داده‌ها:

داده‌های مورد استفاده در این مطالعه، بیان ۷۱۲۹ ژن مربوط به ۷۲ بیمار مبتلا به لوسمی بود که با استفاده از فناوری ریز آرایه توسط گلوب و همکاران به دست آمده است(۹). هر بیمار با برچسب لوسمی میلوژنیک حاد(AML) یا لوسمی لنفوسیتیک حاد(ALL) مشخص

می‌گردد. ۲۵ بیمار مبتلا به لوسمی میلوژنیک حاد و ۴۷ بیمار مبتلا به لوسمی لنفوسیتیک حاد بودند. مجموعه داده‌های مذکور قبلاً به دو دسته داده‌های آموزشی و داده‌های آزمایشی تقسیم شده و در بازه [0,1] نرمال‌سازی شده‌اند. مجموعه داده‌های آموزشی، بیان ژن ۳۸ بیمار(شامل ۱۱ بیمار مبتلا به لوسمی میلوژنیک حاد و ۲۷ بیمار مبتلا به لوسمی لنفوسیتیک حاد) و مجموعه داده‌های آزمایشی بیان ژن ۳۴ بیمار(شامل ۱۴ بیمار مبتلا به لوسمی میلوژنیک حاد و ۲۰ بیمار مبتلا به لوسمی لنفوسیتیک حاد) را مشخص می‌کنند و داده‌های مربوط به بیماران، بالغین و کودکان را شامل می‌شود.

روش پیشنهادی:

تابع توزیع چگالی(Density Distribution Function) ، مفهومی بنیادی در آمار است. متغیر تصادفی X را در نظر بگیرید که تابع توزیع چگالی آن P است. با داشتن تابع توزیع چگالی می‌توانیم تخمینی از توزیع X داشته باشیم. فرض کنید مجموعه‌ای از داده‌های مشاهده شده از نمونه‌ها وجود دارند که تابع توزیع چگالی آن ناشناخته است. برآورد چگالی(Density estimation) به فرآیند تخمین تابع چگالی احتمال یک متغیر تصادفی با استفاده از نمونه‌های مشاهده شده از آن متغیر گفته می‌شود. برآورد چگالی مبتنی بر هسته(kernel density estimation)، روشی ناپارامتر(non-parametric) برای برآورد تابع چگالی احتمال توزیع است که به صورت رابطه زیر تعریف می‌شود:

$$\widehat{P}(x) = \frac{1}{Nh}\sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right)$$

در این رابطه h پارامتر هموارساز یا پهنای باند (bandwidth) است که انتخاب مناسب این پارامتر، مهم‌ترین مسئله در برآورد هسته‌ای است، N تعداد نمونه‌های آموزشی و K(.) تابع هسته است. یکی از توابع هسته معروف، هسته تابع پایه شعاعی(RBF:Radial Basis Function) می‌باشد.

در روش‌های برآورد چگالی هسته‌ای، انتخاب روشی مؤثر برای محاسبه پارامتر هموارسازی از اهمیت خاصی در برآورد چگالی برخوردار است. اغلب روش‌های برآورد چگالی هسته‌ای مقدار ثابتی را برای این پارامتر در نظر می‌گیرند که این مقدار ثابت همیشه می‌تواند با دقت بالایی برای برآورد چگالی مورد استفاده قرار گیرد. به ازای کوچک کردن پارامتر هموارسازی h ، منحنی حاصل از برآورد هسته‌ای ناهموارتر شده و جزئیات جعلی بیشتری را از چگالی واقعی به نمایش می‌گذارد و به ازای بزرگ کردن این پارامتر، منحنی هموار و باعث محو شدن جزئیات واقعی تابع چگالی می‌گردد. از دیگر مشکلات روش‌های برآورد چگالی، احتمال هسته‌ای ابعاد زیاد داده‌ها است.

در این مطالعه، روشی مؤثر برای انتخاب پارامترهای مختلف هموارسازی در هر بعد، پیشنهاد می‌شود. در روش پیشنهادی، هسته تابع پایه شعاعی ناهمسانگرد(anisotropic RBF kernel) مورد استفاده در روش ماشین بردار پشتیبان (Support Vector Machine) برای طبقه‌بندی ناپارامتری بر اساس برآورد چگالی احتمال به کار گرفته می‌شود.

در این بخش یک روش ناپارامتری هسته‌ای مبتنی بر رتبه‌بندی ژن‌ها با استفاده از معیارهای بهره اطلاعاتی (Information Gain) و نسبت بهره(Gain Ration) در هسته تابع پایه شعاعی ناهمسانگرد پیشنهاد می‌شود که از رتبه ژن‌ها برای یادگیری پارامترهای هسته تابع پایه شعاعی ناهمسانگرد استفاده می‌نماید. روش پیشنهادی دارای سه مرحله رتبه‌بندی ژن‌ها، انتخاب ژن‌ها و طبقه‌بندی است.

مرحله رتبه‌بندی ژن‌ها:

اولین مرحله روش پیشنهادی، رتبه‌بندی ژن‌ها بر اساس معیار بهره اطلاعاتی و نسبت بهره است. در این مرحله، بردار رتبه ژن‌ها تشکیل می‌شود. رتبه‌بندی ژن‌ها در روش پیشنهادی برای دو منظور استفاده می‌شود:

– استفاده از رتبه ژن‌ها برای یادگیری پارامترهای هسته تابع پایه شعاعی ناهمسانگرد

– استفاده از رتبه ژن‌ها برای انتخاب ژن بر اساس روش‌های فیلتر

مرحله انتخاب ژن‌ها:

مسئله انتخاب ژن‌ها در واقع شناسایی و انتخاب یک زیر

مجموعه مفید از ژن‌ها از میان مجموعه داده‌های اولیه است که حداکثر توان را در پیشگویی خروجی دارا باشند. برای حل مشکل، ابعاد زیاد داده‌های ریز آرایه در برآورد چگالی هسته‌ای، روش پیشنهادی زیر مجموعه‌ای از ژن‌ها را بر اساس روش انتخاب ویژگی فیلتر انتخاب می‌کند.

در روش پیشنهادی، رتبه ژن‌ها برای انتخاب ژن‌ها و تعیین پارامترهای هموارسازی مورد استفاده قرار می‌گیرند. در این مرحله، ابتدا بردار ژن‌ها به ترتیب نزولی مرتب می‌شوند و ژن‌های با رتبه صفر حذف می‌شوند. سپس ژن‌ها با بالاترین رتبه انتخاب می‌شوند و ژن‌های دارای رتبه پایین حذف می‌شوند.

مرحله طبقه‌بندی:

آخرین مرحله روش پیشنهادی، طبقه‌بندی است. فرض کنید که N نمونه آموزشی شامل d ژن $X = \{x^t, r^t\}_{t=1}^{N}$ وجود دارد که هر نمونه با یک بردار ژن $x^t = (x_1^t, x_2^t, \dots, x_d^t)$ و برچسب r^t مشخص می‌گردد. مرحله انتخاب ژن روش پیشنهادی، k ژن با بالاترین رتبه را به عنوان ورودی مرحله طبقه‌بندی انتخاب می‌کند. تابع جداسازی طبقه‌بندی ناپارامتری هسته‌ای تابع پایه شعاعی ناهمسانگرد به صورت زیر تعریف می‌شود:

$$g_i(x) = \hat{p}(x|C_i)\hat{P}(C_i) = \frac{1}{N\prod_{j=1}^{k} h_j}\sum_{t=1}^{N}\exp\left(-\sum_{j=1}^{k}\frac{(x_j - x_j^t)^2}{2h_j^2}\right) r_i^t$$

در رابطه فوق، x بیان‌گر نمونه جدیدی است که می‌خواهیم نوع لوسمی آن را پیش‌بینی کنیم. با استفاده از این رابطه، نمونه x به کلاسی اختصاص می‌یابد که بالاترین مقدار را داشته باشد. در این رابطه، hj پهنای باند در بعد j را نشان می‌دهد. هم چنین فرض می‌شود که m کلاس C1, C2,..., Cm وجود دارد، اگر نمونه xt متعلق به کلاس Ci باشد، مقدار rti برابر یک و در غیر این صورت صفر است. Ni تعداد نمونه‌های متعلق به کلاس Ci است.

یافته‌ها

برای ارزیابی کارایی روش پیشنهادی، آزمایش‌هایی با استفاده از نرم‌افزار Matlab R2013a انجام می‌شود و عملکرد روش پیشنهادی با استفاده از معیارهای دقت، حساسیت و اختصاصیت با روش ناپارامتری چگالی احتمال هسته‌ای تابع پایه شعاعی مقایسه می‌شود.

میزان دقت یک روش طبقه‌بندی، درصد نمونه‌های طبقه‌بندی شده درست را در میان تمام نمونه‌ها نشان می‌دهد. حساسیت به معنی نسبتی از موارد مثبت است که سیستم آن‌ها را به درستی به عنوان مثبت علامت‌گذاری می‌کند. اختصاصیت به معنی نسبتی از موارد منفی است که سیستم آن‌ها را به درستی به عنوان منفی علامت‌گذاری می‌کند. در آزمایش‌ها، ابتدا مقدار بهینه عرض هسته σ در روش طبقه‌بندی ناپارامتری، هسته تابع پایه شعاعی را با استفاده از اعتبارسنجی عرضی با ده تکرار روی مجموعه داده‌های آموزشی به دست آورده و سپس با استفاده از پارامترهای بهینه تعیین شده، به انجام آزمایش‌ها بر روی مجموعه داده‌های آزمایشی می‌پردازیم. مقدار بهینه پارامتر σ از مجموعه {۰/۲، ۰/۵، ۱، ۲، ۳، ۴، ۵، ۶، ۷، ۸، ۹، ۱۰، ۲۰، ۱۰۰} با استفاده از اعتبارسنجی عرضی با ده تکرار انتخاب می‌شود. نتایج طبقه‌بندی روش پیشنهادی و روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی، در جدول ۱ نشان داده شده است. پارامتر n مقدار ژن‌های استفاده شده در طبقه‌بندی را نشان می‌دهد. مرحله انتخاب ژن روش پیشنهادی توانسته است با استفاده از معیار بهره اطلاعاتی ۲۳۰ ژن و با استفاده از معیار نسبت بهره ۸۶ ژن از ۷۱۲۹ ژن را انتخاب کند. نتایج آزمایش‌ها حاکی از آن است که روش پیشنهادی توانسته است با انتخاب ژن‌های مناسب به عملکرد خوبی دست یابد.

همان گونه که از جدول ۱ مشخص است، کارایی روش پیشنهادی با هر دو معیار بهره اطلاعاتی و نسبت بهره در تشخیص لوسمی میلوژنیک و لنفوسیتیک حاد در مقایسه با روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی، به طور چشمگیری بهبود یافته است. نتایج این جدول هم چنین نشان می‌دهد روش پیشنهادی به طور قابل توجهی تعداد ژن‌ها را کاهش داده که این امر موجب افزایش سرعت و ساده شدن سیستم می‌شود. با مقایسه نتایج به دست آمده توسط روش پیشنهادی با استفاده از معیارهای بهره اطلاعاتی و نسبت بهره مشخص می‌گردد که معیار نسبت

جدول ۱: مقایسه عملکرد روش پیشنهادی و روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی

نام روش	تعداد ژن (n)	دقت (Accuracy)	حساسیت (Sensitivity)	اختصاصیت (Specificity)
روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی	۷۱۲۹	٪۳۵/۲۹	٪۹۰	٪۱۲/۵۰
روش پیشنهادی با معیار بهره اطلاعاتی	۲۳۰	٪۹۷/۰۶	٪۱۰۰	٪۹۵/۸۳
روش پیشنهادی با معیار نسبت بهره	۸۶	٪۹۷/۰۶	٪۱۰۰	٪۹۵/۸۳

جدول ۲: مقایسه عملکرد روش طبقه‌بندی ناپارامتر هسته تابع پایه شعاعی با استفاده از تمام ژن‌ها و ژن‌های تعیین شده توسط روش پیشنهادی

نام روش	تعداد ژن‌ها (n)	دقت (Accuracy)	حساسیت (Sensitivity)	اختصاصیت (Specificity)
روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی	۷۱۲۹	٪۳۵/۲۹	٪۹۰	٪۱۲/۵۰
روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی با معیار بهره اطلاعاتی	۲۳۰	٪۹۴/۱۲	٪۱۰۰	٪۹۱/۶۷
روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی با معیار نسبت بهره	۸۶	٪۹۴/۱۲	٪۱۰۰	٪۹۱/۶۷

بهره توانسته است با تعداد ژن‌های کمتری بـه عملکـردی یکسان با معیار بهره اطلاعاتی در طبقه‌بندی داده‌ها دسـت یابد. به منظور بررسی بیشتر ژن‌های استخراج شده توسط روش پیشنهادی، روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی با اسـتفاده از ژن‌هـای تعیـین شـده توسـط روش پیشنهادی مورد آزمایش قرار می‌گیرد(جدول ۲).

همان گونه کـه در جـدول ۲ نشـان داده شـده اسـت، عملکرد روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی، با استفاده از ژن‌های استخراج شده توسط روش پیشنهادی به طور قابل توجهی بهبود یافته است. هم چنین نتایج این جدول نشان می‌دهند که معیار نسبت بهره با تعداد ژن‌های کمتری قادر به طبقه‌بندی داده‌هـای لوسـمی میلوژنیـک و لنفوسیتیک حاد است.

بحث

در این مطالعه، داده‌های بیان ژن سرطان خون با روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی ناهمسانگرد با استفاده از معیار انتخاب ژن بهره اطلاعاتی و نسبت بهـره طبقه‌بندی گردیدند. روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی ناهمسانگرد، با معیار بهره اطلاعاتی با انتخاب ۲۳۰ ژن به دقت ٪۹۴/۱۲ و با معیار نسبت بهره با انتخاب ۸۶ ژن به دقت٪۹۴/۱۲ رسید. در حالی‌که روش طبقه‌بندی ناپارامتری هسته تابع پایه شعاعی، با استفاده از ۷۱۲۱ ژن به دقت ٪۳۵/۲۹ رسید.

بن دور و همکاران داده‌های ریـز آرایـه لوسـمی را بـا روش نزدیک‌ترین همسایه و ماشین‌بردار پشتیبان با استفاده از هسته درجه دوم مورد بررسی قرار دادند و به ترتیب با دقت ٪۹۱/۶ و ٪۹۴/۴ ، قادر بـه شناسـایی انـواع سـرطان بودند (۱۸). نگون و همکاران با روش جداسازی لجستیک، داده‌های ریز آرایه لوسمی را مورد بررسی قرار دادند و با دقت ٪۹۴/۴ قادر به تشخیص انـواع سـرطان بودنـد، هـم چنین این محققان با روش تحلیل جداسازی درجه دوم به دقـت ٪۹۵/۴ رسـیدند(۱۹). لـی و همکـاران در مطالعـه دیگری با انتخاب روش الگوریتم ژنتیک و طبقه‌بندی کننده KNN بـه دقـت٪۸۴/۶ در ریـز آرایـه لوسـمی دسـت یافتند(۲۰). چن و لین در سال ۲۰۱۱ با انتخاب مجموعـه داده‌های بیان ژن مربوط به سـرطان خـون و انجـام روش BPNN بـا دقـت ٪۹۵/۸۳ قـادر بـه تشخیص انواع سرطان

خون شدند(۲۱).

ونگ و همکاران در سال ۲۰۰۶ از داده‌های بیان ژن و روش KNN و Single NF برای طبقه‌بندی بیماران سرطان خون مبتلا به دو نوع AML و ALL استفاده کردند و به ترتیب به دقت ۷۲/۶۴٪ و ۸۷/۵٪ رسیدند(۲۲).

کای و همکاران در سال ۲۰۱۴ مطالعه‌ای بر روی داده‌های ریزآرایه سرطان خون با استفاده از روشی موسوم به I-RELIEF-NB انجام دادند و به دقت ۹۱/۶۷٪ رسیدند. همین محققان روش I-RELIEF-LDA را بر روی مجموعه داده‌های بالا انجام دادند و به دقت ۹۲/۸۶٪ رسیدند. بالاترین دقت این محققان زمانی بود که آن‌ها از روش RELIEF-KNN استفاده کردند و به دقت ۹۴/۴۴٪ رسیدند (۲۳). ژانگ و همکاران در سال ۲۰۱۲ مطالعه‌ای بر روی مجموعه داده‌های ریزآرایه خون انجام دادند و از روش BMSF-NB استفاده نمودند و به دقت ۹۶/۲۵٪ رسیدند(۲۴).

نتیجه‌گیری

نتایج این مطالعه نشان داد، روش برآورد ناپارامتری هسته، تابع پایه شعاعی ناهمسانگرد با معیار نسبت بهره و انتخاب ژن‌های مناسب، با دقت بالایی قادر به تشخیص سرطان لوسمی است. بنابراین به نظر می‌رسد روش پیشنهادی می‌تواند در تشخیص دقیق‌تر انواع لوسمی‌کمک کند تا تصمیمات مناسب تری در مورد نحوه تشخیص و درمان بیماران گرفته شود.

References :

1- Sheikhpour R, Hekmat Moghadam H. The effect of estrogen on p53 protein in T47D breast cancer cell line. Razi J Med Sci 2015; 22(133): 51-8. [Article in Farsi]
2- Torkaman A, Charkari NM, Aghaeipour M. An approach for leukemia classification based on cooperative game theory. Anal Cell Pathol(Amst) 2011; 34(5): 235-46.
3- Zand AM, Imani S, Saadati M, Borna H, Ziaei R, Honari H. Effect of age, gender and blood group on blood cancer types. Kowsar Med J 2010 15(2): 111-4. [Article in Farsi]
4- Zali H, Amini R, Shiri Haris R. Gene expression analysis of leukemia microarray data by David program . J Ilam Uni Med Sci 2013; 21(2): 92-102. [Article in Farsi]
5- Parsa N. Environmental factors, genes and human cancers. Sci Cultivation J 2012; 2(1): 12-9. [Article in Farsi]
6- 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.
7- Sheikhpour R, Aghaseram M, Sheikhpour R. Diagnosis of acute myeloid and lymphoblastic leukemia using gene selection of microarray data and data mining algorithm. Sci J Iran Blood Transfus Organ 2016; 12 (4): 347-57. [Article in Farsi]
8- Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet 1999; 21(1 Suppl): 10-4.
9- 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.
10- 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]
11- Vahedi M, Alavi Majd H, Mehrabi Y, Naghavi B. Gene expression data clustering and its application in differential analysis of leukemia. J Semnan Uni Med Sci 2008; 9(2): 163-8. [Article in Farsi]
12- 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]
13- Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998; 95(25): 14863-8.
14- Habeck M. DNA microarray technology to revolutionise cancer treatment. Lancet Oncol 2001; 2(1): 5.
15- Alavimajd H, Vahedi M, Mehrabi Y, Naghavi B. Clustering approach in DNA microarray analysis. Research in Medicine 2007; 31(1): 19-25. [Article in Farsi]
16- Alba E, García-Nieto J, Jourdan L, Talbi EG. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. Congr Evol Comput Singapore 2007; 1-7. Available from: file:///C:/Documents%20and%20Settings/m.mokhtari/My%20Documents/Downloads/JMcec2007.pdf.
17- Mahmoud AM, Maher BA, El-Horbaty ES, Salem AB. Analysis of machine learning techniques for gene selection and classification of micoarray data. ICIT

2013. 6th Int Conf Inform Technol 2013; 1-9.
18- Cho SB, Won HH. Machine learning in DNA microarray analysis for cancer classification. Bioinformatics 2003; 19: 189-98.
19- Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002; 18(1): 39-50.
20- Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001; 17(12): 1131-42.
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- Wang Z, Palade V, Xu Y. Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis. In Evolving Fuzzy Systems. In: International Symposium on Evolving Fuzzy Systems 2006; p. 241-6.
23- 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.
24- 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.

Original Article

A new approach for diagnosis of Acute Myeloid and Lymphoblastic Leukemia using gene expression profile and machine learning techniques

Sheikhpour R.[1,2], Sheikhpour R.[3], Aghasaram M.[3]

[1]Department of Physical Activity & Sport Science, Taft Branch, Islamic Azad University, Taft, Iran
[2]Hematology & Oncology Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
[3]Department of Computer Engineering, Yazd University, Yazd, Iran

Abstract

Background and Objectives

Leukemia is a cancer type in the world. One of the most accurate methods for detection and prediction of Acute Myeloid Leukemia and Acute Lymphoblastic Leukemia is to use DNA and genetic information of people. Microarray technology is a tool to study the expression of thousands of genes in shortest possible time. Analyzing the microarray datasets may not be possible without the statistical analysis and machine learning techniques. In this paper, microarray data sets and machine learning techniques are used for the diagnosis of leukemia.

Materials and Methods

The data used in this descriptive study are the expression of 7129 genes of 72 patients with leukemia which have been achieved by the microarray technology. Then, the diagnosis of AML and ALL was performed using the microarray data based on anisotropic radial basis function with the gain ratio and information gain.

Results

The proposed method using information gain with the selection of 230 important genes and using gain ratio with the selection of 86 important genes was able to detect AML and ALL with accuracy of 97.06%, whereas non-parametric kernel classification method based on the radial basis function has the accuracy of 35.29٪ with 7129 genes.

Conclusions

The results of this study showed that the gene expression data and proposed method with gain ratio method are able to detect leukemia with high accuracy. Therefore, it seems that proposed method can help to accurately diagnose leukemia for a better decision making about the diagnosis of diseases and treatment of patients.

Key words: Leukemia, Gene Expression, Microarray Analysis, Machine Learning

Received: 26 Dec 2015
Accepted: 11 May 2016

Correspondence: Sheikhpour R., PhD of Biochemistry. Department of Physical Activity & Sport Science, Taft Branch, Islamic Azad University and Hematology & Oncology Research Center, Shahid Sadoughi University of Medical Sciences.
P.O.Box: 89156-56965, Yazd, Iran. Tel: (+9835) 36282884; Fax: (+9835) 36235958
E-mail: *robab.sheikhpour@iauyazd.ac.ir*