مجله دانشکده پزشکی اصفهان
سال سی و پنجم/شماره‌ی ٤١٩/هفته‌ی سوم فروردین ماه ١٣٩٦

تاریخ دریافت: ١٣٩٥/٩/١٥
تاریخ پذیرش: ١٣٩٥/١١/٣

ارزیابی و به کارگیری مدل‌های مختلف طبقه‌بندی به منظور استخراج ژن‌های شاخص مرتبط با عود سرطان سینه از داده‌های میکروآرایه

محمدرضا صحتی[١]، مینا کاید[٢]

مقاله پژوهشی

چکیده

مقدمه: در این پژوهش، سعی شد تا با به کارگیری مؤثر الگوریتم‌های محاسباتی و مدل‌های ریاضی، اعتبار ژن‌های شاخص مستخرج از داده‌های میکروآرایه و تفسیرپذیری آن‌ها در مطالعات زیستی بهبود یابد.

روش‌ها: ابتدا، بهترین مدل طبقه‌بند با هدف دستیابی به بیشترین صحت پیش‌بینی عود سرطان سینه در داده‌های بیان ژن میکروآرایه انتخاب شد. بدین منظور، از داده‌های بیان ۵۰ ژن شاخص مربوط به تومور سینه در ۱۲۷۱ بیمار مبتلا به سرطان (۳۷۹ نفر با عود سرطان و ۸۹۲ نفر بدون عود) استفاده شد و با مقایسه‌ی عملکرد چند طبقه‌بند مطرح بر روی این ژن‌ها، یک سیستم پیش‌بین مناسب برای عود به دست آمد. در این راستا، ابتدا به منظور افزایش صحت از طریق کاهش بعد و انتخاب مناسب‌ترین ژن‌ها (ویژگی‌ها) روش‌های Correlation-based feature selection (CFS)، Principal component analysis (PCA)، Independent component analysis (ICA)، الگوریتم ژنتیک (Genetic algorithm یا GA) و همچنین، روش انتخاب تصادفی چند ویژگی در ترکیب با انواع مدل‌های طبقه‌بند مورد بررسی قرار گرفت.

یافته‌ها: در مجموع، ۵ ژن شاخص از ترکیب روش الگوریتم ژنتیک، روش Top scoring set (TSS) و انتخاب تصادفی ژن‌ها انتخاب شدند که در اغلب طبقه‌بندها، بهترین نتایج را داشتند. ۵ ژن شاخص نهایی شامل TRIP13، KIF20A، NEK2، RACGAP1 و TYMS، به صورت معنی‌داری در ساختمان Microtubule و Spindle شرکت داشتند و فرایند زیستی اتصال میکروتوبول‌های Spindle به Kinetochore را تنظیم می‌کردند.

نتیجه‌گیری: با استفاده از مدل‌های ترکیبی، علاوه بر اجتناب از بروز خطای انطباق بیش از حد مدل بر داده‌های آموزش، می‌توان به صحت مناسب با ژن‌های شاخصی که از نظر زیست‌شناسی معنی‌دار و تفسیرپذیر باشند، دست پیدا کرد.

واژگان کلیدی: الگوریتم، بیومارکرها، سرطان سینه، طبقه‌بندی، پروفایل بیان ژن

ارجاع: صحتی محمدرضا، کاید مینا. **ارزیابی و به کارگیری مدل‌های مختلف طبقه‌بندی به منظور استخراج ژن‌های شاخص مرتبط با عود سرطان سینه از داده‌های میکروآرایه.** مجله دانشکده پزشکی اصفهان ١٣٩٦؛ ٣٥ (٤١٩): ١٠٣-٩٨

مقدمه

بازگشت سرطان، یک رویداد مرگ‌بار است که با استفاده از درمان‌های کمکی نظیر درمان‌های هورمونی یا پرتودرمانی می‌توان خطر وقوع آن را تا حد زیادی کاهش داد. پایین بودن صحت پیش‌بینی عود سرطان سینه توسط شاخص‌های بالینی و بافت‌شناسی موجب شده است تا در سال‌های اخیر، محققان به دنبال کشف شاخص‌های پیش‌بینی از داده‌های میکروآرایه، مربوط به بیان ژن‌ها در تومورهای سرطان سینه باشند. با کمک صحت بالاتر این شاخص‌ها، می‌توان هزینه‌ی گزاف شیمی‌درمانی‌های مکرر و غیر ضروری را برای بسیاری از بیمارانی که تحت عمل جراحی اولیه‌ی برداشتن تومور سرطانی قرار گرفته‌اند، کاهش داد و همچنین، از مرگ و میر ناشی از عدم استفاده از درمان‌های کمکی تا حد زیادی جلوگیری کرد.

میکروآرایه‌ها با اندازه‌گیری هم‌زمان بیان تعداد زیادی از ژن‌ها گستره‌ی وسیعی از اطلاعات در سطح مولکولی و سلولی را در اختیار ما قرار می‌دهند. van de Vijver و همکاران، داده‌های میکروآرایه‌ی مربوط به ۱۱۷ بیمار جوان مبتلا به سرطان سینه را که در گره‌های

١- استادیار، گروه بیوالکتریک و مهندسی پزشکی، دانشکده‌ی فن‌آوری‌های نوین پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
٢- گروه مهندسی الکترونیک، مؤسسه‌ی علوم و فن‌آوری سپاهان، اصفهان، ایران
نویسنده‌ی مسؤول: محمدرضا صحتی
Email: mr.sehhati@amt.mui.ac.ir

لنفاوی آن‌ها سلول سرطانی مشاهده نشده بود، مورد تحلیل و بررسی قرار دادند (۱). بر اساس نتایج این مطالعه، یک وسیله‌ی سنجش تجاری به نام ماماپرینت برای پیش بینی عود سرطان سینه با استفاده از ۷۰ ژن شاخص، تولید گردید و مورد استفاده قرار گرفت. Wang و همکاران، داده‌های میکروآرایه‌ی مربوط به ۲۸٦ بیمار دیگر را به همین منظور بررسی کردند (۲). در این مطالعه نیز ۷٦ ژن شاخص استخراج و معرفی شدند که تنها ۳ ژن مشترک با مطالعه‌ی قبلی در آن‌ها وجود داشت؛ در حالی که صحت پیش‌بینی در هر دوی آن‌ها بر روی داده‌ی اختصاصی مطالعه‌ی خودشان به طور تقریبی یکسان بود. به طور کلی، ژن‌های شاخص معرفی شده در مطالعات مختلف هم‌پوشانی اندکی دارند و مداخله‌ی همه‌ی آن‌ها در سرطان توجیه پذیر نیست. یک علت ناسازگاری بین مجموعه‌ی ژن‌های شاخص معرفی شده برای عود سرطان سینه در مطالعات مختلف، می‌تواند در نظر نگرفتن تعامل هم‌زمان ژن‌ها در بسیاری از مدل‌های پیش‌بین باشد.

از طرف دیگر، نویز موجود در داده‌های میکروآرایه، نتایج هر مطالعه را به سمت انطباق با مدل‌های مورد استفاده سوق می‌دهد. همچنین، تنوع نمونه‌های مورد استفاده در مطالعات مختلف از نظر مشخصات بیمار مانند سن و همچنین، تنوع نمونه‌ها از نظر وضعیت بافت‌شناسی و هورمونی تومورها، استفاده از الگوریتم‌های متفاوت به منظور پیش‌پردازش و طبیعی‌سازی داده‌ها، استفاده از روش‌های مختلف استخراج ویژگی و نیز استفاده از الگوریتم‌های مختلف طبقه‌بندی برای ساخت مدل پیش‌بینی عود سرطان سینه را می‌توان از عوامل مؤثر در ناسازگاری نتایج مطالعات مختلف دانست (٤-۳). از سوی دیگر، چون ژن‌ها و محصولات آن‌ها در یک لحظه به طور هم‌زمان در رویدادهای مختلف زیستی سلول شرکت دارند، نمی‌توان تأثیر آن‌ها را به طور مستقل بر روی یک رویداد خاص بررسی کرد (٤).

Li و همکاران، روش جدیدی را بر روی داده‌های میکروآرایه‌ی موجود از سرطان سینه معرفی کردند که با استفاده از آن، بتوانند گروه کم خطر را با صحت بالایی در زیر گروه‌های خاص از بیماران پیش‌بینی کنند (۵). Zhao و همکاران، بیان کردند که شاخص‌های گزارش شده برای سرطان سینه جنبه‌های متفاوت زیستی این بیماری را نشان می‌دهد (٦). نتایج نشان داد که با جمع قدرت پیش‌بینی چندین شاخص مختلف، می‌توان صحت پیش‌بینی را در طیف گسترده‌ای از نمونه‌های ناهمگن افزایش داد. صحتی و همکاران، هفت مجموعه‌ی داده‌ی میکروآرایه، شامل ۱۵۳۸ نمونه‌ی تومور سرطان سینه را از نظر شاخص پایداری و صحت طبقه‌بندی مورد بررسی قرار دادند (۳). در نهایت، ۵۰ ژن شاخص معرفی شدند که با تکیه بر پایداری این مجموعه و همچنین، انطباق آن‌ها با ژن‌های شناخته شده‌ی دخیل در سرطان، می‌توان آن‌ها را به عنوان شاخص‌های زیستی عمومی مربوط به بازگشت سرطان سینه معرفی کرد. با وجود پایداری و همچنین، تفسیرپذیری زیستی مجموعه‌ی ژن‌های معرفی شده توسط صحتی و همکاران، صحت پیش‌بینی به دست آمده در مطالعه‌ی مربوط چندان رضایت بخش نبود (۳).

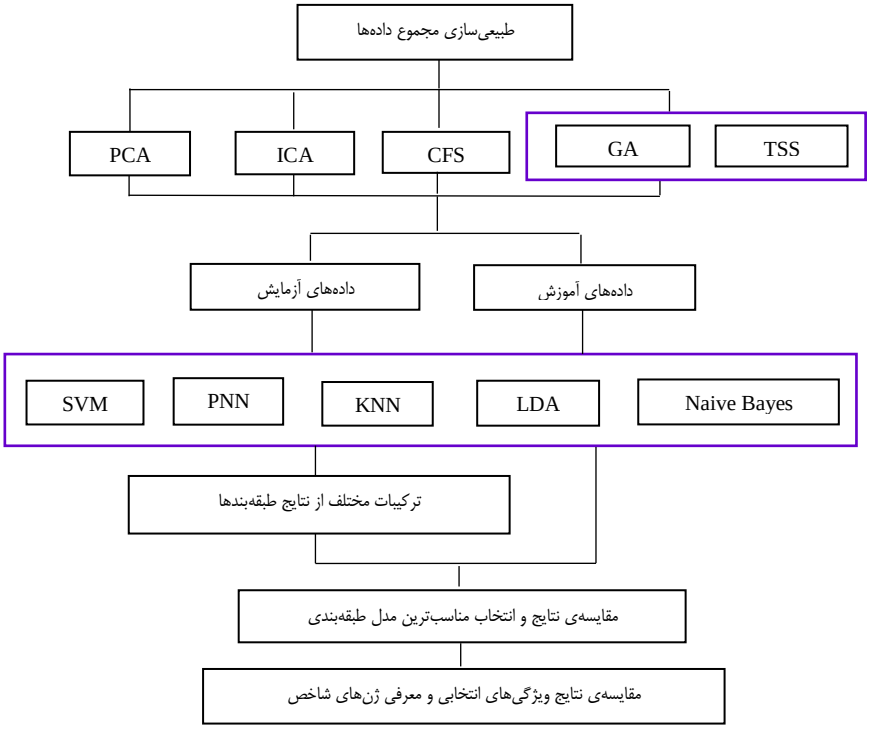
با این مقدمه، در مطالعه‌ی حاضر از ژن‌های پایدار حاصل از مطالعات قبل به عنوان نقطه‌ی شروع استفاده شد و با به کارگیری مدل‌های مختلف مطرح در زمینه‌ی انتخاب ویژگی و طبقه‌بندی، بالاترین صحت به دست آمد. در نهایت، مجموعه‌ی ژن‌های دخیل در مدل بهینه به عنوان ژن‌های شاخص انتخاب و تفسیر شدند.

روش‌ها

روند پردازش داده‌ها: در اغلب کاربردهای یادگیری ماشین به منظور افزایش کارایی مدل، از روش‌های استخراج ویژگی (Feature extraction) مانند روش آنالیز مؤلفه‌های اصلی (PCA یا Principal component analysis) برای کاهش بعد استفاده می‌شود (۵). در بسیاری کاربردها نیز با روش‌های انتخاب ویژگی (Feature selection) اقدام به حذف ویژگی‌هایی می‌شود که دارای اطلاعات مفید نیستند و برای مدل طبقه‌بند گمراه کننده می‌باشند. در این مطالعه، در جهت کاهش بعد و انتخاب مناسب‌ترین ژن‌ها (ویژگی‌ها) چند روش متداول انتخاب ویژگی و همچنین، مدل‌های کارآمد طبقه‌بندی به صورت موازی و هم‌زمان مورد استفاده و ارزیابی قرار گرفتند (شکل ۱). بدین منظور، ویژگی‌های انتخاب شده از هر روش با استفاده از مدل‌های طبقه‌بند مختلف برای محاسبه‌ی صحت پیش‌بینی عود سرطان به کار گرفته شدند.

در ادامه، برای به دست آوردن حداقل تعداد ژن‌ها، از ترکیب روش الگوریتم ژنتیک و همچنین، روش Top scoring set (TSS) استفاده شد (۸-۷). در روش TSS، هدف یافتن دو ژن شاخص است؛ به گونه‌ای که در صورت بیشتر بودن مقدار بیان یکی نسبت به دیگری، بتوان در مورد امکان عود نمونه‌ی سرطانی اظهار نظر کرد. در این بررسی، تمام ترکیبات دوتایی از مجموعه‌ی ۵۰ ژن شاخص منتخب از مطالعه‌ی قبلی، به روش TSS بررسی شد. در نتیجه‌ی این بررسی، چند مجموعه‌ی دوتایی با نتایج نزدیک به هم مشاهده شدند و در نتایج روش الگوریتم ژنتیک نیز این ژن‌ها موجود بودند. همچنین، در ادامه، ترکیبات سه‌تایی و چهارتایی مجموعه‌ی دوتایی‌های پیدا شده در مرحله‌ی اول مورد بررسی قرار گرفتند.

استفاده از الگوریتم ژنتیک برای انتخاب بهترین مدل: در الگوریتم ژنتیک که از فرایند تکامل ژنتیکی الگوبرداری شده است، ترکیبات تصادفی ویژگی‌ها در هر نسل توسط تابع هدف مناسب مورد ارزیابی قرار می‌گیرند و بهترین کاندیداها به نسل بعدی منتقل می‌شوند.



شکل ۱. مراحل پردازش داده‌ها به منظور انتخاب بهترین مدل و تعیین کمترین تعداد از ژن‌های شاخص

PCA: Principal component analysis; ICA: Independent component analysis; CFS: Correlation-based feature selection; GA: Genetic algorithm; TSS: Top scoring set; SVM: Support vector machines; PNN: Probabilistic neural network; KNN: K-nearest neighbor; LDA: Linear discriminant analysis

پس از تولید چندین نسل متوالی، انتظار می‌رود بهترین ترکیب از ویژگی‌ها به آخرین نسل انتقال یافته باشد (۸). به منظور کاربرد الگوریتم مطالعه (۸) متناسب با شرایط مطالعه‌ی حاضر تعداد متغیرها به ۵۰ ویژگی تغییر داده شد که برابر با تعداد کل ژن‌ها بود و در قسمت تابع هدف، هر بار طبقه‌بندهای مختلف جایگزین گردید. طی نسل‌های مختلف، با توجه به نوع طبقه‌بند، مجموعه‌ی ژن‌های متفاوتی به دست آمد که برخی از اعضای این مجموعه‌ها، در اغلب نسل‌ها تکراری بودند و به عنوان ژن‌های پایدار در نظر گرفته شدند. در این مطالعه، همچنین نتایج کاربرد روش پیشنهادی با سایر روش‌های انتخاب ویژگی مطرح در زمینه‌ی انتخاب ژن مقایسه شد.

در این پژوهش، بدون پرداختن به ساخت یک الگوریتم جدید طبقه‌بندی، تنها بر روی انتخاب یک مدل مناسب طبقه‌بند یا ترکیبی از نتایج طبقه‌بندهای مختلف تمرکز شد. در این راستا، طبقه‌بندهای ماشین بردار پشتیبان (Support vector machine) در حالت‌های مختلف خطی (Linear) و غیر خطی (RBF یا Radial basis function) با پارامترهای مختلف (۹)، K-nearest neighbor (KNN)، Naive Bayes، Probabilistic neural network (PNN) و Linear discriminant analysis (LDA) به کار برده شد (۱۱-۱۰).

در کاربرد LDA نیز از دو حالت خطی و هسته‌ی مرتبه‌ی دوم استفاده شد.

در تشخیص عود سرطان با یک مسأله‌ی طبقه‌بندی مشتمل بر دو دسته روبه‌رو هستیم؛ دسته‌ی پرخطر شامل بیمارانی است که رویداد عود سرطان در آن‌ها در محدوده‌ی زمانی کمتر از پنج سال رخ داده است و دسته‌ی کم خطر شامل بیمارانی است که در پنج سال اول، هیچ علامتی از بازگشت سرطان در آن‌ها مشاهده نشده است. با توجه به این که نتیجه‌ی عملکرد نهایی یک طبقه‌بند را به روش‌های متفاوتی می‌توان بیان کرد، برای ارزیابی دستاوردهای این پژوهش از تعاریف زیر استفاده شد:

TP: تعداد نمونه‌هایی که به درستی پرخطر (بازگشت پذیر) تشخیص داده شده‌اند.

TN: تعداد نمونه‌هایی که به درستی کم خطر (بدون بازگشت) تشخیص داده شده‌اند.

FN: تعداد نمونه‌هایی که به اشتباه پرخطر تشخیص داده شده‌اند.

FP: تعداد نمونه‌هایی که به اشتباه کم خطر تشخیص داده شده‌اند.

صحت عبارت از تعداد نمونه‌های درست طبقه‌بندی شده نسبت به تعداد کل نمونه‌ها است.

جدول ۱. نتایج پیش‌بینی عود سرطان سینه بر روی ۷ مجموعه‌ی داده‌ی مستقل با به کارگیری الگوریتم‌های مختلف طبقه‌بندی

طبقه‌بندی کننده	میانگین ± انحراف معیار	ویژگی (درصد)	حساسیت (درصد)	بیشترین صحت پیش‌بینی (درصد)
KNN	۶۱/۰ ± ۳/۵	۷۸	۸۰	۷۵
SVM (linear)	۶۶/۰ ± ۳/۴	۷۴	۸۶	۷۶
SVM (rbf)	۶۵/۰ ± ۳/۲	۷۹	۸۳	۷۸/۵
PNN	۶۴/۰ ± ۳/۱	۷۵	۸۰	۷۸
LDA (QDA)	۶۸/۰ ± ۳/۳	۷۶	۹۰	۸۰/۵
Naive bayes	۶۲/۰ ± ۳/۳	۷۰	۸۶	۷۷

SVM: Support vector machines; RBF: Radial basis function; PNN: Probabilistic neural network; KNN: K-nearest neighbor; LDA: Linear discriminant analysis

حساسیت، عبارت از نسبت تعداد نمونه‌هایی که به درستی پرخطر تشخیص داده شده‌اند، به کل نمونه‌های پرخطر است.

ویژگی، نسبت تعداد نمونه‌هایی که به درستی کم‌خطر تشخیص داده شده‌اند، به کل نمونه‌های کم‌خطر است.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN)$$
$$Sensitivity = TP/(TP + FN)$$
$$Specificity = TN/(TN + FP)$$

یافته‌ها

با مقایسه‌ی نتایج روش TSS و ژن‌های به دست آمده در الگوریتم ژنتیک به ازای تمام طبقه‌بندها مشاهده شد که ژن‌های دارای نتایج بهتر در روش TSS در مجموعه‌ی به دست آمده توسط الگوریتم ژنتیک نیز وجود دارد. همچنین، در ارزیابی جداگانه‌ی هر یک از ۷ مجموعه‌ی داده، با کاربرد الگوریتم ژنتیک با استفاده از یک الگوریتم طبقه‌بند خاص، نتایج مشابهی به دست آمد. با مقایسه‌ی نتایج اعمال الگوریتم‌های طبقه‌بندی بر روی مجموعه‌های مختلف در تعداد متفاوت از دوتایی تا ده‌تایی حاصل از ترکیب ژن‌های معرفی شده در روش الگوریتم ژنتیک و روش TSS و انتخاب تصادفی ژن‌ها، در نهایت مجموعه‌ی پنج‌تایی شامل ژن‌های شاخص دارای بهترین نتیجه بود.

پنج ژن شاخص نهایی عبارت از TRIP13، KIF20A، NEK2، RACGAP1 و TYMS بودند که به صورت معنی‌داری در ساختمان Microtubule و Spindle شرکت داشتند و فرایند زیستی اتصال میکروتوبول‌های Spindle به Kinetochore را تنظیم می‌کردند. روند انتخاب ویژگی توسط الگوریتم ژنتیک برای طبقه‌بندهای مختلف هر بار نتایج متفاوتی نشان داد، اما در هر یک از ۷ مجموعه‌ی داده‌ها به طور مستقل، برای هر طبقه‌بند نتایج یکسان است.

در مرحله‌ی ارزیابی، تمام نمونه‌های مربوط به مجموعه‌ی داده‌های مختلف با هم ترکیب شدند و از مجموعه‌ی حاصل، هر بار ۷۰ نمونه‌ی کم‌خطر و ۷۰ نمونه‌ی پرخطر به عنوان داده‌های آموزش و ۲۰ نمونه‌ی کم‌خطر و ۲۰ نمونه‌ی پرخطر به عنوان داده‌های آزمایش انتخاب شدند؛ به صورتی که داده‌های آموزش و آزمایش با هم هم‌پوشانی نداشتند و هر بار به صورت تصادفی انتخاب می‌شدند. انتخاب تصادفی مجموعه‌ی داده‌ها تا صد مرتبه تکرار شده و نتایج نهایی این ارزیابی متقاطع برای انواع مدل‌های طبقه‌بند در جدول ۱ گزارش شده است.

بحث

در ساخت مدل‌های پیش‌بین، رسیدن به بیشترین صحت پیش‌بینی به عنوان هدف اصلی پی‌گیری می‌شود. در این راستا، استفاده از مدل مناسب به منظور طبقه‌بندی و همچنین، انتخاب بهترین ویژگی‌ها به عنوان ورودی مدل دو جنبه‌ی اصلی موضوع به شمار می‌روند.

در این پژوهش با استفاده از الگوریتم ژنتیک انواع مدل‌های طبقه‌بند در ترکیب با چند روش کاهش بعد بمنظور دستیابی به بهترین مدل ممکن برای پیش‌بینی عود سرطان سینه با استفاده از داده‌های میکروآرایه مورد ارزیابی قرار گرفت. نتایج این ارزیابی نشان می‌دهد که طبقه‌بندهای مختلف از جمله ماشین بردار پشتیبان، LDA و PNN که از طبقه‌بندی کننده‌های پرکاربرد در حوزه‌ی بیوانفورماتیک به شمار می‌روند، پس از انتخاب ژن‌های مؤثر، تفاوت قابل توجهی ندارند. این موضوع در نگرش جدید به شناسایی الگو نیز مطرح شده است که مطابق آن، استفاده از مدل‌ها و الگوهای رفتاری پیچیده برای پیش‌پردازش و طبقه‌بندی داده‌ها همیشه کارآمد نیست (۸).

با این حال، مدل LDA (Pseudo quadratic) صحت بالاتری (۸۰ درصد) نسبت به سایر مدل‌ها داشته است. از طرف دیگر، مطابق با نتایج به دست آمده اهمیت پیدا کردن ویژگی‌های مناسب هم از نظر دستیابی به صحت پیش‌بینی مطلوب و هم از نظر تفسیرپذیری زیستی ژن‌های شاخص، نقش مهم‌تری نسبت به انتخاب نوع طبقه‌بند دارد. به عبارت دیگر، صحت نهایی مدل به ژن‌های شاخص انتخاب شده در مرحله‌ی کاهش بعد، وابستگی بیشتری دارد.

تشکر و قدردانی

این مقاله از پایان نامه‌ی دوره‌ی کارشناسی ارشد کـه بـا تـأمین اعتبـار توسط مؤسسه‌ی علوم و فناوری سپاهان اصفهان انجام شد، اسـتخراج شده است.

References

1. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002; 347(25): 1999-2009.
2. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005; 365(9460): 671-9.
3. Sehhati M, Mehridehnavi A, Rabbani H, Pourhossein M. Stable gene signature selection for prediction of breast cancer recurrence using joint mutual information. IEEE /ACM Trans Comput Biol Bioinform 2015; 12(6): 1440-8.
4. Mehridehnavi A, Zand H, Sehhati M. Dimensionality reduction on topological features of the gene network constructed from microarray data for prediction of breast cancer recurrence. J Isfahan Med Sch 2016; 33(359): 1973-85. [In Persian].
5. Li J, Lenferink AE, Deng Y, Collins C, Cui Q, Purisima EO, et al. Identification of high-quality cancer prognostic markers and metastasis network modules. Nat Commun 2010; 1: 34.
6. Zhao X, Rodland EA, Sorlie T, Naume B, Langerod A, Frigessi A, et al. Combining gene signatures improves prediction of breast cancer survival. PLoS One 2011; 6(3): e17845.
7. Kriti, Virmani J, Dey N, Kumar V. PCA-PNN and PCA-SVM based CAD systems for breast density classification. In: Hassanien AE, Grosan C, Fahmy Tolba M, editors. Applications of intelligent optimization in biology and medicine: Current trends and open problems. New York, NY: Springer; 2016. p. 159-80.
8. Yang S, Naiman DQ. Multiclass cancer classification based on gene expression comparison. Stat Appl Genet Mol Biol 2014; 13(4): 477-96.
9. Babatunde O H, Armstrong L, Leng J , Diepeveen D. A genetic algorithm-based feature selection. International Journal of Electronics Communication and Computer Engineering 2014; 5(4): 899-905.
10. Reddy SVG, Thammi Reddy K, Valli Kumari V, Varma Kamadi VSPR. An SVM based approach to breast cancer classification using RBF and Polynomial kernel functions with varying arguments. International Journal of Computer Science and Information Technologies 2014; 5(4): 5901-4.
11. Amini Z, Mehridehnavi A. Comparison of different classifiers for prediction of breast cancer metastasis in microarray analysis. J Isfahan Med Sch 2014; 32(292): 1028-35. [In Persian].

Journal of Isfahan Medical School

Vol. 35, No. 419, 3rd Week, April 2017

Received: 05.12.2016

Accepted: 22.01.2017

Evaluation of Different Classification Models to Extract Gene Signatures for Breast Cancer Recurrence Using Microarray Data

Mohammadreza Sehhati[1], Mina Kayed[2]

Original Article

Abstract

Background: In this study, we aimed to improve the reliability and biological interpretability of gene signatures selected from microarrays by efficient usage of computational models and mathematical algorithms.

Methods: At the first step, a good model with high accuracy was chosen to predict cancer recurrence in microarray gene expression data on breast tumors. In this regard, microarray gene expression data of breast tumor in 1271 cancer patients (379 with recurrence and 892 people without recurrence) were utilized to construct an appropriate predictive model for recurrence by comparing the performance of multiple classifiers. In the pre-processing stage, different methods like correlation-based feature selection (CFS), principal component analysis (PCA), independent component analysis (ICA), and genetic algorithm as well as a random selection method were used to reduce the dimensions and choose the most appropriate genes (features).

Findings: A total of five gene signatures were selected by combining genetic algorithm, top scoring set (TSS), and random selection method, which showed the best results in most classification models. The final indicator genes were TRIP13, KIF20A, NEK2, RACGAP1 and TYMS, which had significant contribution in the structure of microtubules and spindle and also regulated the attachment of spindle microtubules to kinetochore.

Conclusion: By using hybrid models, we can avoid overfitting in training and achieve acceptable accuracy with biologically interpretable genes.

Keywords: Algorithms, Biomarkers**,** Breast cancer, Classification, Gene expression profiling

Citation: Sehhati M, Kayed M. **Evaluation of Different Classification Models to Extract Gene Signatures for Breast Cancer Recurrence Using Microarray Data.** J Isfahan Med Sch 2017; 35(419): 98-103.

1- Assistant Professor, Department of Bioelectric and Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

2- Department of Electrical Engineering, Sepahan Institute of Higher Education, Isfahan, Iran

Corresponding Author: Mohammadreza Sehhati, Email: mr.sehhati@amt.mui.ac.ir