مجله دانشکده پزشکی اصفهان
سال سی و دوم/شماره ۲۹۲/هفته اول شهریور ماه ۱۳۹۳

تاریخ دریافت: ۱۳۹۳/۱/۳۱
تاریخ پذیرش: ۱۳۹۳/۴/۱۲

مقایسه‌ی عملکرد طبقه‌بندی کننده‌های مختلف در پیشگویی متاستاز سرطان سینه با استفاده از میکرو آرایه

زهرا امینی[۱]، دکتر علیرضا مهری دهنوی[۲]

مقاله پژوهشی

چکیده

مقدمه: هدف از این تحقیق، دستیابی به یک روش قابل اعتماد برای پیشگویی متاستاز در مبتلایان به سرطان سینه بود.

روش‌ها: در این مطالعه از آنالیز میکرو آرایه‌ی DNA مربوط به تومور سینه در ۷۸ بیمار جوان با شرایط یکسان (۳۴ نفر با متاستاز و ۴۴ نفر بدون متاستاز) استفاده و تلاش شد تا با مقایسه‌ی عملکرد چند طبقه‌بندی کننده‌ی مطرح روی بیان ژن‌های آن‌ها، یک سیستم پیشگویی قوی برای متاستاز به دست آید. برای این امر، طبقه‌بندی کننده‌های SWLDA (Stepwise linear discriminate analysis)، ماشین بردار پشتیبان (SVM یا Support vector machine) و K نزدیک‌ترین همسایه (KNN یا K-Nearest Neighbours) با استفاده از روش LOO (Leave one out) بر روی ۲۳۱ ژن انتخابی به کار برده شد تا این نمونه‌ها را به دو گروه با و بدون متاستاز تفکیک کنند.

یافته‌ها: روش ماشین بردار پشتیبان با کرنل خطی از نظر میزان صحت، Sensitivity و Specificity بهترین روش است. ماشین بردار پشتیبان با استفاده از کرنل خطی توانست با Sensitivity بیش از ۸۴ درصد و Specificity نزدیک به ۸۲ درصد به تفکیک دادگان بپردازد.

نتیجه‌گیری: در مورد روش SWLDA، مزیتی که وجود دارد این است که این طبقه‌بندی کننده قبل از دسته‌بندی از یک مرحله‌ی انتخاب ویژگی بهره می‌برد که این مسأله پیچیدگی تابع تصمیم‌گیری را کمتر و تعمیم پذیری طبقه‌بندی کننده را بیشتر می‌کند.

واژگان کلیدی: میکرو آرایه، پیشگویی سرطان سینه، طبقه‌بندی کننده‌ی Support vector machine، K-nearest neighbours، Stepwise linear discriminate analysis

ارجاع: امینی زهرا، مهری دهنوی علیرضا. **مقایسه‌ی عملکرد طبقه‌بندی کننده‌های مختلف در پیشگویی متاستاز سرطان سینه با استفاده از میکرو آرایه.** مجله دانشکده پزشکی اصفهان ۱۳۹۳؛ ۳۲ (۲۹۲): ۱۰۳۵-۱۰۲۸

مقدمه

بسیاری اوقات بیماران مبتلا به سرطان سینه که از نظر سطح و میزان پیشرفت بیماری هم در شرایط مشابه قرار دارند، نسبت به درمان‌ها پاسخ و واکنش متفاوتی نشان می‌دهند و احتمال برگشت بیماری و متاستاز هم در آن‌ها متفاوت می‌باشد (۲-۱). تاکنون قوی‌ترین روش‌های موجود برای پیشگویی متاستاز (مانند Lymph node status و Histological grade) هم در طبقه‌بندی دقیق تومورهای سینه بر اساس رفتار کلینیکی، موفق نبوده‌اند. در این مقاله، از آنالیز میکرو آرایه‌ی DNA مربوط به تومورهای سینه‌ی اولیه از ۷۸ بیمار جوان با شرایط یکسان استفاده و تلاش شد

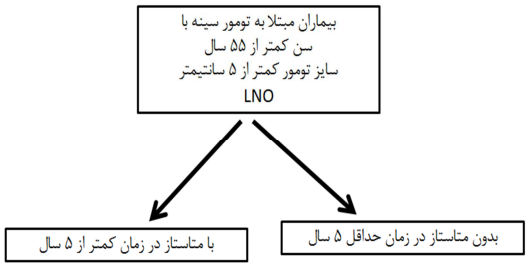
۱- دانشجوی دکتری، گروه مهندسی پزشکی، دانشکده‌ی فناوری‌های نوین، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۲- دانشیار، گروه مهندسی پزشکی- بیوالکتریک، گروه مهندسی پزشکی، دانشکده‌ی فناوری‌های نوین، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
نویسنده‌ی مسؤول: دکتر علیرضا مهری دهنوی
Email: mehri@med.mui.ac.ir

تا با طبقه‌بندی کننده با سرپرست روی بیان ژن‌های آن‌ها، یک پیشگویی قوی برای متاستاز به دست آید.

با توسعه‌ی تکنیک‌های میکرو آرایه، امکان رصد کردن بیان هزاران ژن به صورت همزمان فراهم شده است؛ بدیهی است دادگانی که از این طریق به دست می‌آیند، ابعاد بسیار بزرگی دارند که استفاده از روش‌های طبقه‌بندی روی آن‌ها ممکن است منجر به نتایج خوبی نشود. از این رو بسیاری اوقات قبل از استفاده از طبقه‌بندی کننده، سعی می‌شود تا ژن‌های مهم‌تر و مؤثرتر در بحث مربوط، به نحو مقتضی انتخاب شوند (۳). در این تحقیق، از دادگان مقاله‌ی ۴ استفاده شده است؛ چون همان‌طور که در بخش بعد توضیح داده خواهد شد، محققان تلاش کرده‌اند تا با به کارگیری روش‌های مختلف از میان حدود ۲۵۰۰۰ ژن موجود برای هر فرد، تنها ۲۳۱ ژن مناسب‌تر را برگزینند. این ۲۳۱ ژن بر اساس دامنه‌ی ضریب همبستگی مرتب شدند و در نهایت، برای بهینه کردن تعداد ژن‌های به کار رفته در طبقه‌بندی کننده با سرپرست، مرحله به مرحله به تعداد ژن‌ها اضافه شد؛ به این ترتیب که در هر مرحله ۵ ژن از بالای لیست ژن‌های مرتب شده به مجموعه‌ی ژن‌های مورد استفاده در طبقه‌بندی اضافه گردیده و کارایی طبقه‌بندی کننده در هر مرحله با روش LOO (Leave one out) سنجیده شده است. در نهایت، بهترین جواب طبقه بندی روی ۷۸ نمونه با استفاده از ۷۰ ژن به دست آمده است که در این حالت، در ۸۳ درصد موارد (۶۵ نفر از ۷۸ نمونه) پیشگویی صحیح صورت گرفته است (۴).

هدف از این تحقیق، دسته‌بندی این نمونه‌ها با روشی به غیر از روش به کار رفته در سایر پژوهش‌ها (۴) و مقایسه‌ی کارایی دو روش بود. برای این امر طبقه‌بندی کننده‌های SWLDA (Stepwise linear discriminate analysis)، ماشین بردار پشتیبان (SVM یا Support vector machine) و K نزدیک‌ترین همسایه (KNN یا K-Nearest Neighbours)با استفاده از روش LOO بر روی ۲۳۱ ژن انتخابی به کار برده شدند تا این نمونه‌ها را به دو گروه با و بدون متاستاز تفکیک نمایند.

در این تحقیق از دادگان موجود در پژوهش دیگری (۴) استفاده گردید که در این بخش به بیان جزئیات مربوط به این دادگان و نحوه‌ی استخراج ژن‌های مناسب (به منظور طبقه‌بندی) پرداخته می‌شود. داده‌های مورد بررسی از ۷۸ بیمار با شرایط مشابه به دست آمده است. تمام بیماران Lymph node negative بودند و در زمان تشخیص اولیه، کمتر از ۵۵ سال سن داشتند. از بین بیماران، ۳۴ بیمار در طی ۵ سال بعد از تشخیص اولیه، دچار متاستاز شده بودند (Poor prognosis) و ۴۴ بیمار در زمانی حداقل برابر ۵ سال دچار متاستاز نشده بودند (Good prognosis).



شکل ۱. تفکیک بیماران به دو گروه با و بدون متاستاز (۴)

سپس میکرو آرایه‌ی حاوی حدود ۲۵۰۰۰ ژن

برای هر بیمار تهیه شد و پارامتر Intensity (میانگین هندسی Intensity برای هر دو کانال سبز و قرمز) برای هر ژن محاسبه شد. در مرحله‌ی اول، ژن‌ها به سه دسته‌ی بدون تغییر، با افزایش بیان ژن و با کاهش بیان ژن تقسیم شدند. سپس در بین ژن‌هایی که تغییر بیان داشتند، آن‌هایی که بیش از ۲ برابر تغییر بیان داشتند، انتخاب شدند که به این ترتیب، تعداد ژن‌های مورد بررسی به حدود ۵۰۰۰ ژن کاهش یافت. در مرحله‌ی بعد، میزان همبستگی هر یک از این ژن‌ها با نتیجه‌ی بیماری (Disease outcome) محاسبه شد و ۲۳۱ ژن که همبستگی بیش از ۰/۳ با نتیجه‌ی بیماری داشتند، حفظ شدند و سایر ژن‌ها از بررسی حذف گردیدند.

روش‌ها

هدف نهایی در هر مسأله‌ی شناسایی الگو، تفکیک مجموعه‌ای از نمونه‌ها به دو یا چند طبقه‌ی مختلف است. در این جا هدف نهایی، تفکیک دادگان به دو گروه با متاستاز و فاقد متاستاز است. طبقه‌بندی کننده‌ها در یک دسته‌بندی کلی به دو دسته‌ی با سرپرست و بدون سرپرست تقسیم می‌شوند. در روش‌های با سرپرست از یک مجموعه‌ی داده‌ی برچسب خورده، به عنوان مجموعه‌ی آموزشی برای تنظیم پارامترهای طبقه‌بندی کننده استفاده می‌شود. به عبارت دیگر، ابزار طبقه‌بندی، فضای ورودی و خروجی مسأله و ارتباط بین آن‌ها را از روی یک سری داده‌ی آموزشی یاد می‌گیرد (۵).

در این پژوهش نیز پس از انتخاب ۲۳۱ ژن مؤثر، از طبقه‌بندی کننده‌های زیر برای تفکیک دادگان به دو گروه با/ بدون متاستاز استفاده شد.

طبقه‌بندی کننده‌ی تفکیک خطی با پالایش گام به گام (SWLDA)

LDA (Linear discriminant analysis) ساده‌ترین و پرکاربردترین طبقه‌بندی کننده‌ی آماری است. چنانچه اطلاعاتی مبنی بر چگونگی توزیع ویژگی‌ها در فضای ویژگی در اختیار نداشته باشیم، توابع جداساز خطی اولین انتخاب به شمار می‌آیند. توابع جداساز خطی در ساده‌ترین فرمشان به دلیل عدم نیاز به تنظیم پارامترهای مختلف و اطلاعات جانبی و سادگی در پیاده‌سازی، اولین گزینه‌ی پیشنهادی در مسایل طبقه‌بندی محسوب می‌شوند. این طبقه‌بندی کننده در فرم استانداردش یک طبقه‌بندی کننده‌ی باینری است که با اعمال تخمین MAP (Maximum a posteriori) و با فرض گوسی بودن تابع توزیع احتمال شرطی ویژگی‌ها در فضای ویژگی و همچنین تساوی ماتریس‌های کوواریانس طبقات، محاسبه می‌شود (۷-۶).

طبقه‌بندی کننده‌ی SWLDA نیز در واقع همان LDA است که قبل از انجام عملیات طبقه‌بندی از یک روش پیشرو- پسرو برای انتخاب ویژگی استفاده می‌کند (۹-۸). در شروع، هیچ ویژگی برای مدل در نظر گرفته نمی‌شود، ویژگی که معیار ورود به سیستم را برآورده کند، به مدل اضافه می‌شود. بعد از اضافه شدن هر ورودی جدید به مدل، یک روند بازگشتی گام به گام پسرو برای حذف ویژگی‌هایی که کمترین اهمیت را دارند، به کار می‌رود. این فرایند، تا زمانی که مدل شامل تعداد ویژگی‌های از پیش تعیین شده باشد، یا تا زمانی که هیچ ویژگی اضافه‌ای نمانده باشد که معیار خروج/ ورود را برآورده کند، ادامه می‌یابد.

در جدول ۱، ویژگی‌های انتخاب شده به روش گام به گام در مورد ۲۳۱ ژن اولیه آمده است.

جدول ۱. ویژگی‌های منتخب توسط SWLDA

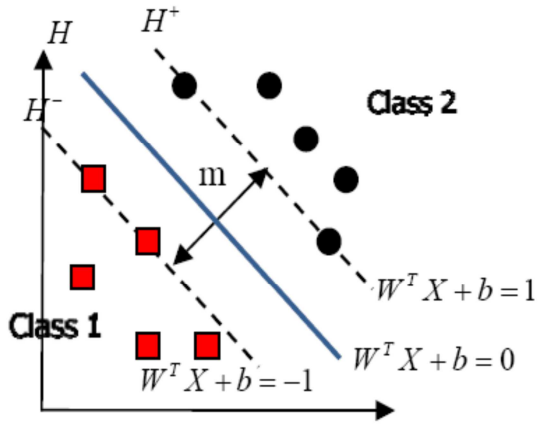
شماره‌ی ویژگی	نام ویژگی
۸	NM-۰۰۳۸۸۲
۷	Contig ۵۵۳۷۷-RC
۵۸	NM-۰۱۸۱۰۴
۱۵۲	Contig ۵۷۸۶۴-RC
۱۰۶	NM-۰۱۴۳۶۳
۲۱۶	Contig ۲۵۹۹۱
۱۸۷	NM-۰۰۷۲۰۳
۱۰۵	Contig ۱۷۷۸-RC
۱۸	Contig ۶۳۱۰۲-RC
۲۱۱	AF-۰۵۲۱۶۲
۲۲۱	NM-۰۰۲۰۱۹
۲۱۷	Contig ۳۵۲۵۱-RC
۲۱	NM-۰۰۰۳۲۰

SWLDA: Stepwise linear discriminate analysis

ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان از طبقه‌بندی کننده‌های پرکاربرد است که به خاطر موفقیتش در تشخیص ارقام دست‌نویس به شهرت رسید. این طبقه‌بندی کننده بر اساس حداقل‌سازی خطر خطا، مرز بین دو طبقه را مشخص می‌کند. شیب این مرز فقط تابع تعدادی از بردارهای ورودی است که روی حاشیه‌ی مرز دو طبقه قرار می‌گیرند و بردارهای پشتیبان مرز نامیده می‌شوند. این طبقه‌بندی کننده مانند LDA یک طبقه‌بندی کننده‌ی باینری است؛ اما هر دو نوع خطی و غیر خطی آن موجود است. نوع غیر خطی SVM از هسته‌های غیر خطی به خصوص توابع RBF (Radial basis function) استفاده می‌کند (۱۰-۱۱). ماشین بردار پشتیبان خطی در حالت جدایی پذیر به فرم زیر است:

فرض می‌کنیم $X = \{X_1, X_2, ..., X_n\}$ مجموعه‌ی نمونه‌ها و $y \in \{1, -1\}$ هم برچسب‌های طبقات باشند. هدف، یافتن مرز تصمیم خطی است که نمونه‌ها را به درستی طبقه‌بندی کند و تا حد امکان از نمونه‌های دو طبقه دور باشد. اگر این طبقات مطابق شکل ۲ به صورت خطی جدایی پذیر باشند، یک ابر صفحه مانند H و تابع تمایز f(x) آن‌ها را از هم جدا می‌کند.



شکل ۲. ماشین بردار پشتیبان در حالت جدایی پذیر خطی

$$H: W.x + b = 0 \quad ,f() = sing\ (W.x + b) \quad (1)$$

چنانچه فاصله‌ی بین ابر صفحات H^- و H^+، $m = 2/\|W\|$ –حاشیه‌ی مرزی– بزرگ‌تر شود، طبقه‌بندی کننده با حاشیه‌ی اطمینان بزرگ‌تر و در نتیجه در برابر نویز مقاوم‌تر خواهد بود. طراحی ابر صفحه‌ی بهینه با بیشترین عرض ناحیه‌ی مرزی، به شرط درست دسته‌بندی شدن تمام نمونه‌ها، یک مسأله‌ی بهینه‌سازی مقید است و با استفاده از روش ضرایب لاگرانژ قابل حل است.

(۲)

$$\begin{cases} \text{Minimize}\ \frac{1}{2}\|W\|^2 \\ \text{Subject to}\ y_i(W^T x + b) \geq 1 \forall i = 1, \ldots, n \end{cases}$$

K نزدیک‌ترین همسایه (KNN)

هدف این تکنیک مشخص کردن طبقه‌ی یک داده‌ی آزمایشی بر اساس طبقه‌ی K داده‌ی آموزش که نزدیک‌ترین همسایه‌های آن هستند، می‌باشد (۱۴-۱۲). این نزدیک‌ترین همسایه‌ها، به طور معمول با استفاده از یک فاصله‌ی قابل اندازه‌گیری به دست می‌آید. با یک K به قدر کافی بزرگ و نمونه‌های آموزش کافی، KNN می‌تواند هر تابعی را تقریب بزند که این مسأله باعث می‌شود تا KNN بتواند مرزهای تصمیم‌گیری غیر خطی ایجاد کند.

در روش K نزدیک‌ترین همسایه، به طور معمول K عددی فرد انتخاب می‌شود. در این تحقیق هم صحت طبقه‌بندی برای چند مقدار مختلف K سنجیده شد که در جدول ۲ نتایج برای سه حالت K = ۱, ۳, ۵ آمده است.

طبق جدول ۲ به نظر می‌رسد مقدار مناسب برای K عدد ۳ می‌باشد که در ادامه از K = ۳ برای این طبقه‌بندی کننده استفاده گردید.

یافته‌ها

با استفاده از سه طبقه‌بندی کننده‌ی SWLDA، SVM و KNN، دادگان مربوط به دو طبقه‌ی با و بدون متاستاز تفکیک شدند.

در جداول ۳ تا ۵، نتایج مربوط به دسته‌بندی دادگان مربوط به ۷۷ بیمار (۳۳ نفر با متاستاز و ۴۴ نفر بدون متاستاز) با استفاده از این سه طبقه‌بندی کننده آمده است. لازم به ذکر است که به دلیل وجود برخی مقادیر Intensity نامطلوب در یکی از دادگان گروه با متاستاز، یکی از نمونه‌های این گروه از روند محاسبات حذف گردید.

در جدول ۶ نیز پارامترهای مربوط به هر یک از طبقه‌بندی کننده‌های مورد استفاده آمده است.

همان‌طور که جدول ۶ نشان می‌دهد، بهترین طبقه‌بندی کننده‌ها به ترتیب SVM، KNN و SWLDA می‌باشند. SVM با استفاده از کرنل خطی توانست با حساسیت بیش از ۸۴ درصد و اختصاصیت نزدیک به ۸۲ درصد به تفکیک دادگان بپردازد.

جدول ۲. مقایسه‌ی اثر مقادیر مختلف K در روش KNN

متغیر / K	میزان صحت	حساسیت	تعیین کنندگی
K = ۱	۰/۷۷۹۲	۰/۸۶۳۶	۰/۶۶۶۷
K = ۳	۰/۸۱۸۲	۰/۸۴۰۹	۰/۷۸۷۹
K = ۵	۰/۸۰۵۲	۰/۸۴۰۹	۰/۷۵۷۶

KNN: K-Nearest Neighbours

جدول ۳. جدول تصمیم‌گیری و دسته‌بندی دادگان با روش KNN و K = ۳

K = ۳	KNN	خرجی صحیح	
		عاقبت خوب	عاقبت بد
خروجی طبقه‌بندی کننده	عاقبت خوب	۳۷	۷
	عاقبت بد	۷	۲۶

KNN: K-Nearest Neighbours

جدول ۴. جدول تصمیم‌گیری و دسته‌بندی دادگان با روش SVM

SVM		خروجی صحیح	
		عاقبت خوب	عاقبت بد
خروجی طبقه‌بندی کننده	عاقبت خوب	۳۷	۶
	عاقبت بد	۷	۲۷

SVM: Support vector machine

جدول ۵. جدول تصمیم‌گیری و دسته‌بندی دادگان با روش SWLDA

SWLDA		خروجی صحیح	
		عاقبت خوب	عاقبت ضعیف
خروجی طبقه‌بندی کننده	عاقبت خوب	۳۲	۷
	عاقبت ضعیف	۱۲	۲۶

SWLDA: Stepwise linear discriminate analysis

جدول ۶. مقایسه‌ی عملکرد طبقه‌بندی کننده‌های مختلف در پیشگویی متاستاز سرطان سینه

پارامتر / طبقه‌بندی کننده	میزان صحت	حساسیت	تعیین کنندگی
KNN	۰/۸۰۵۲	۰/۸۴۰۹	۰/۷۵۷۶
SVM	۰/۸۳۱۲	۰/۸۴۰۹	۰/۸۱۸۲
SWLDA	۰/۷۵۳۲	۰/۷۲۷۳	۰/۷۸۷۹

KNN: K-Nearest Neighbours; SVM: Support vector machine; SWLDA: Stepwise linear discriminate analysis

بحث

هدف از این تحقیق، بررسی عملکرد طبقه‌بندی کننده‌های مختلف در پیشگویی متاستاز در سرطان سینه و با استفاده از دادگان میکرو آرایه بود. بدین منظور، از دادگان میکرو آرایه‌ی مربوط به دو دسته افراد با شرایط یکسان که به سرطان سینه مبتلا شده بودند، استفاده گردید که دسته‌ی اول در بازه‌ی زمانی حداقل برابر ۵ سال دچار متاستاز نشده بودند، اما در دسته‌ی دوم در زمانی کمتر از این مدت متاستاز رخ داده بود.

با توجه به طبقه‌بندی کننده‌های پرکاربرد در حوزه‌ی بیوانفورماتیک، در این تحقیق سه طبقه‌بندی کننده‌ی SWLDA، SVM و KNN مورد قیاس قرار گرفتند. با توجه به جداول دیده می‌شود که برای دسته‌بندی دادگان به دو گروه با/ بدون متاستاز، روش SVM با کرنل خطی از نظر میزان صحت، حساسیت و اختصاصیت بهترین روش است.

در قیاس با کارهای مشابه صورت گرفته روی همین دادگان، تا پیش از این بهترین کار توسط van't Veer و همکاران (۴) صورت گرفته بوده است و طی آن بهترین جواب طبقه‌بندی روی ۷۸ نمونه با استفاده از ۷۰ ژن به دست آمد که در این شرایط صحت ۸۳ درصد موارد (۶۵ نفر از ۷۸ نمونه) مشخص شد؛ در حالی که در روش پژوهش حاضر، SVM با استفاده از کرنل خطی توانست با حساسیت بیش از ۸۴ درصد و اختصاصیت حدود ۸۲ درصد به

تفکیک دادگان بپردازد و نتایج بهتری نسبت به روش‌های گذشته به دست آورد.

نکته‌ی مهم دیگر در مورد روش SWLDA این است که این طبقه‌بندی کننده طبق جدول ۱ تنها با استفاده از ۱۳ ژن به جای ۲۳۱ ژن به این میزان از صحت رسیده است که این مسأله حجم محاسبات را بسیار کاهش می‌دهد و حسن روش SWLDA استفاده از تعداد ویژگی کمتر برای طبقه‌بندی است. این مسأله، پیچیدگی تابع تصمیم‌گیری را کاهش و تعمیم پذیری طبقه‌بندی کننده را افزایش می‌دهد. در صورتی که بسیاری از طبقه‌بندی کننده‌هایی که پیش از این مورد استفاده قرار گرفته اند، همواره با مشکل بزرگی بردار ویژگی ها و بالا بودن حجم محاسبات روبه‌رو بوده‌اند.

تشکر و قدردانی

بدین‌وسیله از معاونت تحقیقات و فناوری دانشگاه علوم پزشکی اصفهان سپاسگزاری می‌گردد.

References

1. Weigelt B, Peterse JL, van't Veer LJ. Breast cancer metastasis: markers and models. Nature Reviews 2005; 5: 591-602.
2. Lujambioa A, Calinc GA, Villanuevad A, Roperoa S, Sanchez-Cespedese M, Blancof D, et al. A microRNA DNA methylation signature for human cancer metastasis. PNAS 2008; 105(36): 13556-61.
3. Mehridehnavi A, Ziaei L. Minimal gene selection for classification and diagnosis prediction based on gene expression profile. Adv Biomed Res 2013; 2: 26.
4. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415(6871): 530-6.
5. Lotte F, Congedo M, Lecuyer A, Lamarche F, Arnaldi B. A review of classification algorithms for EEG-based brain-computer interfaces. J Neural Eng 2007; 4(2): R1-R13.
6. Huerta EB, Duval B, Hao JK. Selection for microarray data by a LDA-based genetic algorithm. Lecture Notes in Computer Science 2008; 5265: 250-61.
7. Sharmaa A, Paliwala KK. Cancer classification by gradient LDA technique using microarray gene expression data. Data and Knowledge Engineering 2008; 66(2): 338-47.
8. Krusienski DJ, Sellers EW, McFarland DJ, Vaughan TM, Wolpaw JR. Toward enhanced P300 speller performance. J Neurosci Methods 2008; 167(1): 15-21.
9. Nijboer F, Sellers EW, Mellinger J, Jordan MA, Matuz T, Furdea A, et al. A P300-based brain-computer interface for people with amyotrophic lateral sclerosis. Clin Neurophysiol 2008; 119(8): 1909-16.
10. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000; 16(10): 906-14.
11. Hernandez JCh, Duval B, Hao JK. SVM-based local search for gene selection and classification of microarray data. Bioinformatics Research and Development Communications in Computer and Information Science 2008; 13: 499-508.
12. Parry RM, Jones W, Stokes TH, Phan JH, Moffitt RA, Fang H, et al. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. Pharmacogenomics J 2010; 10(4): 292-309.
13. Jiangsheng Y. Method of k-nearest neighbors. Beijing, China: Institute of Computational Linguistics Peking University; 2002.
14. Krusienski DJ, Sellers EW, Cabestaing F, Bayoudh S, McFarland DJ, Vaughan TM, et al. A comparison of classification techniques for the P300 Speller. J Neural Eng 2006; 3(4): 299-305.

Journal of Isfahan Medical School Received: 20.04.2014

Vol. 32, No. 292, 1st Week, September 2014 Accepted: 03.07.2014

Comparison of Different Classifiers for Prediction of Breast Cancer Metastasis in Microarray Analysis

Zahra Amini MSc[1], Alireza Mehridehnavi PhD[2]

Original Article

Abstract

Background: In this research, we investigated the performance of some different classifiers for prediction of metastasis in breast cancer.

Methods: We used the DNA microarrays of primary breast tumors of 78 young patients. Among these patients, 34 had developed distant metastases within 5 years (poor prognosis group) and 44 formed good prognosis group. For analysis, we applied three different classifiers including support vector machine (SVM), stepwise linear discriminant analysis (SWLDA) and K-nearest neighbors (KNN) classifier. Each of these classifiers used 231 selected genes as an input feature vector and their performances were estimated via using leave one out (LOO) method to classify patients into two groups namely, good and poor prognosis.

Findings: The best results were obtained by support vector machine with linear kernel. This classifier achieved a sensitivity and specificity of 84% and 82%, respectively, for metastasis prediction.

Conclusion: Our findings provide a strategy to specify patients who would benefit from adjuvant therapy.

Keywords: Microarrays, Prediction of breast cancer, Support vector machine (SVM), Stepwise linear discriminant analysis (SWLDA), k-nearest neighbors (KNN) classifiers

Citation: 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

1- PhD Student, Department of Bioelectric and Biomedical Engineering, School of Advanced Medical Technology, Isfahan University of Medical Sciences, Isfahan, Iran
2- Associate Professor, Department of Bioelectric and Biomedical Engineering, School of Advanced Medical Technology, Isfahan University of Medical Sciences, Isfahan, Iran
Corresponding Author: Alireza Mehridehnavi PhD, Email: mehri@med.mui.ac.ir