مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره پنجم شماره دوم. ص ۲۵۱-۲۴۴، ۱۳۹۷

ارزیابی تأثیر انتخاب ویژگی و توابع کرنل مختلف در عملکرد SVM برای تشخیص سرطان پستان

اعظم اروجی[۱]، مصطفی لنگری زاده[۲]*

• **دریافت مقاله:** ۹۶/۱۱/۲۹ • **پذیرش مقاله:** ۹۷/۴/۲۱

مقدمه: سرطان پستان یکی از رایج‌ترین سرطان‌ها در میان زنان است. در تصاویر ماموگرافی، تشخیص تومورهای خوش‌خیم از بدخیم به دلیل شباهت ساختاری کاری دشوار و زمان‌بر است. یادگیری ماشین یک شاخه از هوش مصنوعی است که می‌تواند به صورت ابزاری کمکی در کنار پزشک قرار گیرد و آن‌ها را در تصمیم‌گیری یاری کند. ماشین بردار پشتیبان SVM یکی از رایج‌ترین روش‌های یادگیری ماشین است که عملکرد آن به نوع تابع کرنل و ویژگی‌های ورودی وابسته است. هدف این مطالعه، بررسی تأثیر انتخاب ویژگی و توابع کرنل مختلف در عملکرد SVM می‌باشد.

روش: این مطالعه از نوع تحلیلی بود و با روش مقایسه‌ای انجام گرفت. انتخاب بهترین ویژگی‌ها توسط الگوریتم ژنتیک انجام شد. سپس SVM با توابع کرنلی مختلف شامل چندجمله‌ای، خطی، توابع شعاعی پایه، درجه دو و پرسپترون چندلایه ابتدا با تمام ویژگی‌ها و سپس با ویژگی‌های منتخب آموزش و ارزیابی شد. به منظور ارزیابی عملکرد طبقه‌بندها از مجموعه داده سرطان پستان ویسکانسین و پیاده‌سازی مدل‌ها در متلب انجام شد.

نتایج: نتایج نشان داد که بعد از انتخاب ویژگی عملکرد SVM با تابع کرنل پرسپترون چندلایه کاهش و با تابع کرنل درجه دو افزایش یافت. با این حال، عملکرد توابع کرنل خطی و تابع شعاعی پایه در هر دو حالت مطلوب بود. به طور کلی، بعد از کاهش بعد، بهترین مقدار دقت، ویژگی، حساسیت و صحت به ترتیب به میزان ۰/۶۶۳، ۰/۸۳۳، ۱/۰۷۷ و ۰/۱۳۸ درصد کاهش یافت.

نتیجه‌گیری: روش‌های مبتنی بر تکنیک‌های یادگیری ماشین می‌توانند پزشکان را در تصمیم‌گیری برای درمان یا تشخیص بیماری یاری کنند.

کلید واژه‌ها: یادگیری ماشین، کاهش بعد، سیستم‌های پشتیبان تصمیم بالینی، تشخیص زودهنگام سرطان

• **ارجاع:** اروجی اعظم، لنگری زاده مصطفی. ارزیابی تأثیر انتخاب ویژگی و توابع کرنل مختلف در عملکرد SVM برای تشخیص سرطان پستان. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۷؛ ۵(۲): ۲۴۴-۲۵۱.

۱. دانشجوی دکترای انفورماتیک پزشکی، گروه مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع رسانی پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران
۲. دکترای تخصصی انفورماتیک پزشکی، استادیار، گروه مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع رسانی پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران
* **نویسنده مسئول:** تهران، خیابان ولیعصر، میدان ونک، خیابان رشید یاسمی، پلاک ۶، دانشکده مدیریت واطلاع رسانی پزشکی
• **شماره تماس:** ۰۲۱۸۸۷۹۴۳۰۱
• **Email:** Langarizadeh.m@iums.ac.ir

مقدمه

سرطان پستان از رایج‌ترین سرطان‌ها، چه در کشورهای در حال توسعه و چه در کشورهای توسعه یافته، میان زنان است [۱،۲]. سالانه نزدیک به ۶۰۰۰۰۰ مرگ بر اثر سرطان پستان و بیش از یک میلیون ابتلا به سرطان پستان در کل جهان گزارش می‌شود [۳،۴]؛ اگرچه سرطان پستان یکی از دلایل اصلی مرگ بر اثر سرطان در میان زنان است، اگر به زودی تشخیص داده شود امکان درمان آن بیشتر خواهد بود [۵]. تشخیص زودهنگام نه تنها منجر به مدیریت و درمان بهتر می‌شود که هزینه‌های درمان را نیز کاهش می‌دهد [۱]. ماموگرافی روش قدیمی تشخیص سرطان پستان است؛ اما تحقیقات نشان داده نه تنها رادیولوژیست‌ها در تفسیر یک تصویر ماموگرافی بسیار متفاوت عمل می‌کنند، بلکه ۹۰٪ رادیولوژیست‌ها کمتر از ۳٪ و تنها ۱۰٪ از آن‌ها حدوداً ۲۵٪ از سرطان‌ها را تشخیص می‌دهند [۶]؛ با وجود آزمایش‌های مختلف، همچنان تشخیص می‌تواند، حتی برای خبرگان، دشوار باشد. ظهور فناوری‌های نوین پزشکی و حجم زیاد داده‌های بیماران مسیر را برای توسعه استراتژی‌های جدید در پیش‌بینی و تشخیص سرطان هموار کرده است [۷]. در سال‌های اخیر، تکنیک‌های یادگیری ماشین برای تشخیص بیماری‌ها و کاهش هزینه‌های پزشکی مورد توجه قرار گرفته‌اند [۳]. (Machine Learning) ML شاخه‌ای از هوش مصنوعی است که به موضوع یادگیری از روی نمونه‌های داده و استنتاج بر اساس آن می‌پردازد [۸]. در سال‌های اخیر، دقت پیش‌بینی سرطان به واسطه استفاده از تکنیک‌های ML، بین ۱۵ تا ۲۰ درصد افزایش داشته است [۸]. مطالعات زیادی به کارگیری رویکردهای هوش مصنوعی در پیش‌بینی سرطان سینه را موفقیت‌آمیز دانسته‌اند [۶].

روش‌های یادگیری ماشین ML به دو گروه اصلی تقسیم می‌شوند: یادگیری باناظر و بدون ناظر: در یادگیری باناظر، مجموعه‌ای از داده‌ها برای آموزش ماشین استفاده می‌شوند که با جواب درست برچسب خورده‌اند؛ اما در یادگیری بدون ناظر، هیچ داده برچسب خورده‌ای وجود ندارد و مشخص نیست که پاسخ مورد انتظار چیست [۷،۸]. طبقه‌بندی یکی از روش‌های یادگیری باناظر است که از داده‌های برچسب‌دار برای توسعه یک مدل پیش‌بینی استفاده می‌کند [۷]. در تحقیقات سرطان، این تکنیک‌ها می‌توانند برای شناسایی الگوهای میان داده‌ها، پیش‌بینی بقا و تشخیص تومورهای خوش خیم از بدخیم مورد استفاده قرار گیرند [۹-۱۲،۷].

SVM(Support Vector Machine) رایج‌ترین تکنیک برای دسته‌بندی داده‌های خطی و غیرخطی است. برای داده‌های خطی SVM با حداکثر کردن فاصله داده‌ها تا یک hyper plane، بهینه‌ترین hyper plane را می‌سازد. برای داده‌های غیرخطی، SVM ابتدا با استفاده از توابع کرنل داده را به فضایی با ابعاد بیشتر نگاشت می‌کند و سپس hyper plane بهینه را می‌سازد [۱۳]. مهم‌ترین مزایای SVM عبارت‌اند از: مقاومت به نویز، تولید طبقه‌بند دقیق، عدم مشکل بیش برازش و یادگیری سریع حتی در صورت بالابودن تعداد ورودی‌ها [۱۳،۱۴]. مطالعات زیادی به اثربخشی SVM در تشخیص سرطان پستان یعنی‌دسته‌بندی تومورهای خوش خیم از بدخیم پرداخته‌اند [۱،۳،۷،۱۳-۱۵].

برای کار با SVM یک سری موارد ضروری است که نوع تابع کرنل و همچنین انتخاب زیرفضای بهینه ویژگی از آن دسته‌اند. به منظور انتخاب زیر فضای بهینه از الگوریتم‌های کاهش بعد استفاده می‌شود. کاهش بعد یکی از مهم‌ترین تکنیک‌های پیش پردازش داده است. در کاهش بعد داده مزایای زیادی وجود دارد برای نمونه الگوریتم‌های ML در ابعاد کم بهتر عمل می‌کنند، به علاوه با کاهش بعد ویژگی‌های نامرتبط حذف می‌شوند، نویز کاهش می‌یابد و در نتیجه مدل بهتری ساخته می‌شود. کاهش بعدی که با انتخاب زیرمجموعه‌ای از ویژگی‌ها انجام می‌شود «انتخاب ویژگی» نام دارد [۸]. در سال‌های اخیر، در میان تمام تکنیک‌های انتخاب ویژگی، رویکردهای مبتنی بر جمعیت مانند الگوریتم کلونی مورچگان و زنبورها و الگوریتم ژنتیک بسیار مورد توجه قرار گرفته‌اند [۱۶]. الگوریتم ژنتیک یک الگوریتم جستجوی تصادفی هوشمند است که از تکامل در طبیعت الهام گرفته شده است. الگوریتم ژنتیک (Algorithm Genetic) GAدارای همه ویژگی‌هایی است که یک الگوریتم جستجو برای حل مسائل دنیای واقعی باید داشته باشد: GA را می‌توان به سادگی برای مسائل جدید تغییر داد، زیرا عملگرهای آن به دانش خاص مسئله وابستگی کمی دارند. به علاوه، GA جامع بوده و نسبت به نویز مقاوم است [۶]. هدف این مقاله ارزیابی اثر توابع کرنل مختلف و انتخاب ویژگی مبتنی بر GA بر عملکرد SVM در پیش‌بینی سرطان پستان است.

روش

به منظور ارزیابی از پایگاه داده سرطان پستان ویسکانسین WBCD(Wisconsin Breast Cancer Database) موجود در پایگاه UCI

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29 استفاده شد. این مجموعه داده شامل ۶۹۹ رکورد و ۱۰ ویژگی است (بدون در نظر گرفتن شناسه) که اطلاعات مربوط به آن‌ها در جدول ۱ نشان داده شد. از این میان، تعداد ۲۴۱ بیمار (۳۴٪/۵) دارای تومور بدخیم و ۴۵۸ بیمار (۶۵٪/۵) دارای تومور خوش‌خیم هستند. در مرحله پیش پردازش داده، ۱۶ مقدار مفقودی با استفاده از میانگین هر صفت پر شدند.

جدول ۱: ویژگی‌های مجموعه داده سرطان پستان Wisconsin

نام متغیر	دامنه	میانگین (انحراف معیار)
ضخامت انبوه (Clump Thickness)	۱-۱۰	۴/۴۲ (۲/۸۲)
یکنواختی اندازه سلول (Uniformity of Cell Size)	۱-۱۰	۳/۱۳ (۳/۰۵)
یکنواختی شکل سلول (Uniformity of Cell Shape)	۱-۱۰	۳/۲۱ (۲/۹۷)
چسبندگی لبه‌ها (Marginal Adhesion)	۱-۱۰	۲/۸۱ (۲/۸۵)
حجم سلول بافت اپیتلیال (Single Epithelial Cell Size)	۱-۱۰	۳/۲۲ (۲/۲۱)
هسته‌های عریان (Bare Nuclei)	۱-۱۰	۳/۵۴ (۳/۶)
کروماتین بلاند (Bland Chromatin)	۱-۱۰	۳/۴۴ (۲/۴۴)
هسته عادی (Normal Nucleoli)	۱-۱۰	۲/۸۷ (۳/۰۵)
میتوز (Mitoses)	۱-۱۰	۱/۵۹ (۱/۷۲)
کلاس (Class)	۲ (خوش خیم) و ۴ (بدخیم)	

در این مطالعه، از الگوریتم ژنتیک برای انتخاب بهترین ویژگی‌ها استفاده شد. تابع برازشی که برای شناسایی ویژگی‌های تأثیرگذار در تشخیص تومورهای خوش خیم و بدخیم، در نظر گرفته شد، K نزدیک‌ترین همسایه (K-nearest neighbor KNN) است. به منظور کاهش ابعاد داده تا حد ممکن، در این مقاله، تعداد همسایه‌ها ۳ در نظر گرفته شد تا بهترین ۳ ویژگی مشخص شوند.

به منظور بررسی اثر توابع کرنل بر عملکرد SVM، ۵ مدل متفاوت با کرنل‌های خطی، درجه دو (Quadratic)، پرسپترون چند لایه (Multi-layer perceptron)MLP، چند جمله‌ای (Polynomial) و توابع شعاعی پایه (Radial basis function)RBF ایجاد شد. هر کدام دو بار، بر اساس تمامی ویژگی‌ها و بر اساس ویژگی‌های انتخاب شده توسط GA، آموزش دیده و سپس تست می‌شوند. الگوریتم ۱ روش پیشنهادی را نشان می‌دهد. به منظور پیاده‌سازی الگوریتم پیشنهادی از نرم‌افزار Matlab استفاده شد.

الگوریتم ۱: الگوریتم روش پیشنهادی

گام اول: پیش پردازش داده و پر کردن نمونه‌های مفقودی
گام دوم: استفاده از داده به عنوان ورودی الگوریتم ژنتیک
گام سوم: ای
جاد جمعیت اولیه به صورت تصادفی
گام چهارم: محاسبه تابع برازش
گام پنجم: بررسی شرایط خاتمه GA (رسیدن به حداکثر تعداد تکرار که دراینجا ۱۰۰ در نظر گرفته شد، یا رسیدن به مقدار بهینه)
بله: گام نهم
خیر: گام ششم
گام ششم: انتخاب والدین (روش: چرخ رولت)
گام هفتم: اعمال عملگرهای تقاطع (روش Arithmetic) و جهش (روش Bit inversion)
گام هشتم: تکرار از گام چهارم
گام نهم: خاتمه GA
گام دهم: استفاده از ویژگی‌های منتخب به عنوان ورودی SVM
گام یازدهم: آموزش و آزمون ۵ مدل SVM با کرنل‌های مختلف بر اساس تمام ویژگی‌ها و ویژگی‌های منتخب
گام دوازدهم: مقایسه عملکرد SVM با و بدون انتخاب ویژگی (روش 10 fold cross-validation)

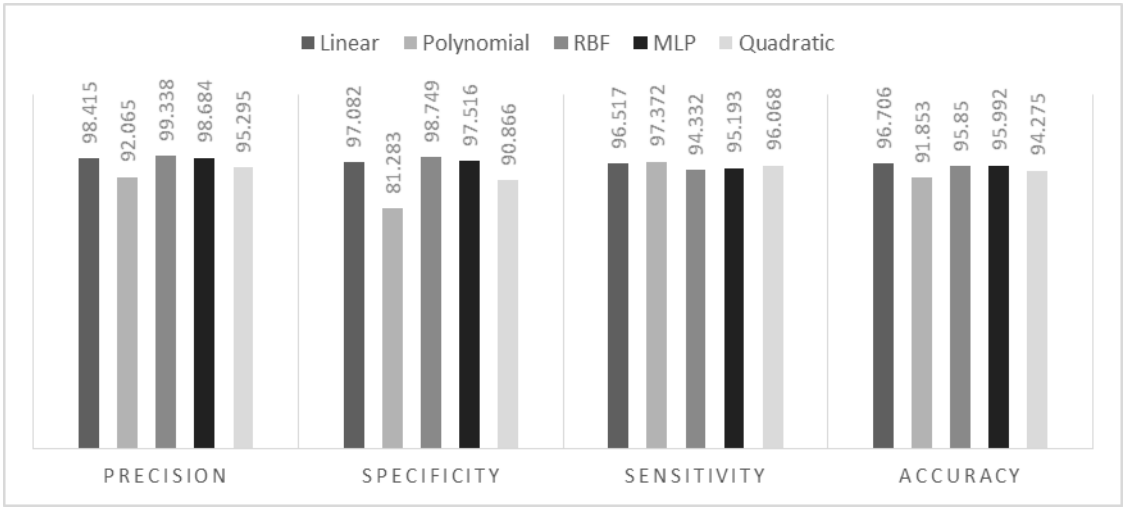
نتایج

پس از پیش پردازش داده، به منظور انتخاب ویژگی از GA با تابع برازش KNN مبتنی بر فاصله اقلیدسی استفاده شد. ویژگی‌های منتخب عبارت‌اند از BN(Bare Nuclei)، CT(Clump Thickness) و (Normal Nucleoli) NN در مرحله بعد، مدل‌های SVM بر اساس توابع کرنل خطی، چند جمله‌ای، RBF، MLP و درجه دو بر اساس تمامی ویژگی‌ها و ویژگی‌های منتخب ساخته شدند. یکی از رایج‌ترین روش‌های ارزیابی عملکرد دسته‌بندها که مجموعه داده برچسب دار را به چند زیرمجموعه تقسیم می‌کنند، Cross-Validation می‌باشد. بدین منظور با استفاده از روش 10 fold cross validation مجموعه داده به ۱۰ زیر مجموعه مستقل تقسیم شد و هر بار یکی به عنوان مجموعه تست و سایر داده‌ها به عنوان داده آموزش در نظر گرفته شدند. بدین صورت هر داده یک بار برای تست و ۹ بار برای آموزش استفاده می‌شود.

در نتیجه کل مجموعه داده برای آموزش و تست پوشش داده می‌شود. معیارهای ارزیابی عملکرد طبقه‌بند شامل دقت، صحت، حساسیت و ویژگی با میانگین‌گیری میان تعداد دفعات تکرار محاسبه شد. جدول ۲ نحوه محاسبه هر یک از این معیارها را نمایش می‌دهد. میانگین نتایج حاصل از ۱۰ مرتبه اجرای SVM با توابع کرنل مختلف برای حالتی که تمام ویژگی‌ها وارد مدل شده‌اند، در شکل ۱ نشان داده شد.

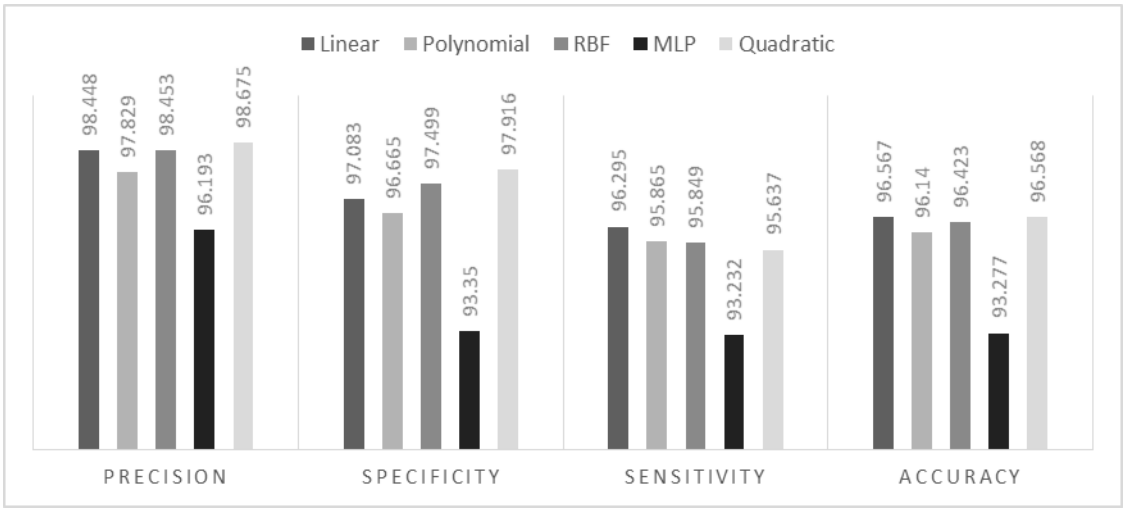
جدول ۲: نحوه محاسبه معیارهای ارزیابی عملکرد طبقه‌بند – TP: مثبت درست، TN: منفی درست، FP: مثبت نادرست و FN: منفی نادرست

صحت	Accuracy= (TP+TN)/(TP+TN+FP+FN)
دقت	Precision= TP /(TP+FP)
حساسیت	Sensitivity = TP / (TP+FN)
ویژگی	Specificity = TN /(TN+FP)



شکل ۱: میانگین معیارهای ارزیابی برای ۵ مدل مختلف SVM

پس از اجرای GA به منظور انتخاب ویژگی، ۵ مدل SVM با توابع کرنل مختلف بر اساس سه ویژگی انتخاب شده، آموزش و سپس تست شدند. میانگین معیارها نیز برای تمامی مقادیر k در شکل ۲ نشان داده شد.



شکل ۲: میانگین معیارهای ارزیابی برای ۵ مدل مختلف SVM بعد از کاهش بعد

بحث و نتیجه‌گیری

باتوجه به نتایج، مدل‌های SVM مبتنی بر توابع کرنل RBF، MLPو Linear به جواب‌های بهتری دست یافته‌اند؛ اگر چه ویژگی در آن‌ها پایین است؛ اما معمولاً در تشخیص بیماری‌هایی چون سرطان مدلی که حساسیت بالاتری دارد ترجیح داده می‌شود. در مدل با حساسیت بالا احتمال این که تومور بدخیم، خوش خیم تشخیص داده شود، کم است.

با توجه به یافته‌های شکل ۲، در دسته‌بندی بر اساس ویژگی‌های انتخاب شده مدل‌های SVM مبتنی بر توابع کرنل درجه دو، خطی و RBF عملکرد مناسبی داشتند. همچنین از مقایسه شکل‌های ۱ و ۲ مشاهده می‌شود که بعد از انتخاب ویژگی عملکرد SVM با تابع کرنل MLP کاهش و با تابع کرنل درجه دو افزایش یافت. با این حال، عملکرد توابع کرنل RBF و خطی در هر دو حالت مطلوب بود. به طور کلی، بعد از کاهش بعد، بهترین مقدار دقت، ویژگی، حساسیت و صحت به ترتیب به میزان ۰/۶۶۳، ۰/۸۳۳، ۱/۰۷۷ و ۰/۱۳۸ درصد کاهش یافت.

در مقاله Kumar و Sitamahalakshmi [۱۳] عملکرد SVM با شبکه عصبی احتمالی (Probabilistic Neural Network)PNN برای دسته‌بندی مقایسه شده است. بدین منظور از سه مجموعه داده پزشکی استفاده شده که عبارت‌اند از: WBCD، ماموگرافی و هابرمن. از روش 10-fold cross validation و معیار صحت برای ارزیابی مدل‌ها استفاده شد. نتایج نشان داد که به ترتیب مجموعه داده‌های WBCD، ماموگرافی و هابرمن SVM با تابع کرنل RBF و صحت ۹۶/۹٪، ۸۲/۲٪ و ۷۱/۷٪ عملکرد بهتری نسبت به PNN با صحت ۷۱/۵٪، ۷۹/۵٪ و ۷۰/۳٪ دارد. در مطالعه حاضر SVM با تابع RBF بعد از کاهش بعد نیز به دقت قابل مقایسه‌ای رسیده است.

Bazazeh و Shubair نیز در مقاله خود [۷] به مقایسه سه روش SVM، شبکه بیزین و جنگل تصادفی در دسته‌بندی WBCD پرداختند. نتایج نشان داد شبکه بیزین بر اساس معیارهای حساسیت (۹۷٪/۱) و دقت (۹۷٪/۲) و جنگل تصادفی با AREA UNDER ROC برابر با ۹۹/۹٪ بهترین روش‌ها هستند.

Gatuha و Jiang [۱] نیز به مقایسه عملکرد شش روش داده‌کاوی پرداخته که عبارت انداز: Bagging، SVM، Decorate، J48، KNN و Naïve Bayes. بدین منظور از WBCD، روش ارزیابی 10-fold cross validation و معیارهای حساسیت، ویژگی و صحت استفاده شده است. بر اساس معیار حساسیت Naïve Bayes (۹۷٪/۵)، معیار ویژگی Bagging (۹۷٪/۵) و معیار صحت SVM (۹۷٪) بهترین عملکرد را داشتند. همچنین از میان ترکیب‌های مختلف این شش روش، ترکیب SVM، KNN، J48 و Naïve Bayes با صحت ۹۷/۳٪ بهترین رویکرد بود. در این مقاله نوع تابع کرنل مورد استفاده تعیین نشده؛ اما صحت مدل SVM

توسعه داده شده حدود ۰/۴ از SVM با انواع تابع کرنل، بعد از کاهش بعد بیشتر است.

در مطالعه Sewak و همکاران [۱۵] نیز از ترکیب دسته‌بندها به منظور بهبود عملکرد تشخیص استفاده شده است. ابتدا بر اساس 10-fold cross validation، ۳۰ مدل SVM با توابع کرنل خطی، RBF و چندجمله‌ای آموزش دیده و سپس بهترین ۵ مدل برای پیاده‌سازی دسته‌بند ترکیبی به کار گرفته شدند. مجموعه داده تشخیص سرطان پستان ویسکانسین با ۵۶۹ رکورد و ۳۲ ویژگی برای ارزیابی مدل استفاده شده است. از میان مدل‌ها تابع کرنل چند جمله‌ای بهترین عملکرد را داشت در حالی‌که برای مجموعه داده مورد استفاده در مطالعه حاضر، تابع چند جمله‌ای عملکرد مناسبی نداشت؛ اما روش ترکیبی با صحت ۹۹/۲۹٪، دقت ۹۸/۸۹٪، ویژگی ۹۸/۱۱٪ و حساسیت ۱۰۰٪ عملکردی بهتر از ۵ مدل انتخابی داشت.

در مطالعه Mert و همکاران [۱۴] نیز از مجموعه داده ویسکانسین استفاده شده؛ اما با استفاده از آنالیز اجزای مستقل ICA (Independent component Analysis) ابعاد داده به یک ویژگی کاهش یافته است. در بخش ارزیابی، عملکرد روش‌های KNN، ANN، RBF-NN و SVM قبل و بعد از کاهش بعد بر اساس روش‌های 5/10-fold cross validation و بخش‌بندی ۲۰٪ و ۴۰٪ بررسی شده است. در 10-fold cross validation با در نظر گرفتن تمامی ویژگی‌ها، ANN با صحت ۹۷/۵۳٪، ویژگی ۹۳/۳۹٪، حساسیت ۱۰۰٪، معیار-F برابر با ۹۸/۰۷٪، شاخص Youden برابر با ۰/۹۳۴، قدرت تشخیصی (Discriminant power) بی‌نهایت و AUC برابر با ۰/۹۵۶ بهترین روش بود. بعد از کاهش بعد با ICA، روش SVM با تابع کرنل RBF با معیار-F برابر با ۹۳/۰۴٪، قدرت تشخیصی ۲/۷۶۹ و حساسیت ۹۷٪/۴۷ و روش KNN با شاخص Youden برابر با ۰/۷۹۵، صحت ۹۱/۰۳٪ و ویژگی ۸۴/۹٪ و RBF-NN با AUC مساوی با ۰/۸۸۱ بهترین عملکرد را داشتند؛ اگرچه مدل SVM با تابع RBF حساسیت بالاتری از مدل مطالعه حاضر بعد از انتخاب ویژگی دارد؛ اما از نظر معیارهای صحت (۹۰٪/۸۶) و ویژگی (۷۹٪/۷۱) مدل پژوهش حاضر عملکرد بهتری داشته است. از نظر زمان CPU کاهش بعد باعث کاهش زمان به اندازه ۱/۲۵ ثانیه برای KNN، ۴۱/۴۹ ثانیه برای ANN، ۳۹/۳۵ ثانیه برای RBF-NN و ۸/۳۳ ثانیه برای SVM شده است.

دسته‌بند پیشنهادی Mittal و همکاران [۳] ترکیبی از شبکه عصبی خودسازمان‌ده و stochastic gradient descent است که عملکرد آن با سه روش درخت تصمیم، SVM و جنگل تصادفی مقایسه شده است. برای ارزیابی عملکرد این روش‌ها از 10-fold cross validation و معیار صحت استفاده شده است. نتایج نشان داد که برای هر دو مجموعه داده، تبلیغات اینترنتی و سرطان پستان Wisconsin، روش SVM برای داده‌های آموزش با صحتی برابر ۹۶/۷۴۵٪ و ۹۹٪/۷۸۲ بهترین روش بوده است؛ اما در مرحله تست، روش پیشنهادی با صحت ۹۶/۸۷۲٪ و ۹۹/۶۹۲٪ بهتر از SVM عمل کرده است. در این مقاله تابع کرنل SVM ذکر نشده به همین دلیل نتایج آن با مطالعه حاضر قابل مقایسه نیست.

تابع کرنل RBF رایج‌ترین تابعی است که در طراحی مدل‌های SVM استفاده شده است و معمولاً از نظر صحت از سایر روش‌های یادگیری ماشین بهتر عمل کرده است. تنها در مطالعه Mert و همکاران [۱۴] به موضوع انتخاب ویژگی پرداخته شد که در آن نیز کاهش اندک در معیارهای ارزیابی و افزایش سرعت نشان دهنده مؤثر بودن الگوریتم‌های کاهش بعد است. در این مطالعه همچنین مشخص شد که SVM از نظر زمانی نیز پیچیدگی کمی دارد. هیچ کدام از مطالعات قبلی به بررسی تأثیر انواع توابع کرنل بر عملکرد SVM نپرداخته‌اند. در این مطالعه به بررسی تأثیر انتخاب ویژگی و توابع کرنل بر عملکرد طبقه‌بند SVM در دسته‌بندی تومورهای خوش‌خیم و بدخیم سرطان پستان پرداخته شد؛ اگر چه نتایج نشان داد که توابع کرنل RBF و خطی با هر تعداد ویژگی به نتایج مطلوبی دست یافتند، اما در مطالعات آینده بایستی این موارد با مجموعه داده‌های بزرگ و واقعی نیز آزمایش شود. با توجه به این که معمولاً ابعاد داده‌های واقعی بالا است، بررسی تأثیر انواع روش‌های انتخاب ویژگی بر عملکرد طبقه‌بندها می‌تواند مفید باشد.

تضاد منافع

بدین‌وسیله نویسندگان تصریح می‌نمایند که هیچ‌گونه تضاد منافعی در خصوص پژوهش حاضر وجود ندارد.

References

1. Gatuha G, Jiang T. Evaluating Diagnostic Performance of Machine Learning Algorithms on Breast Cancer. In: He X, Gao X, Zhang Y, Zhou ZH, Liu ZY, Fu B, et al, editors. Intelligence Science and Big Data Engineering Big Data and Machine Learning Techniques: 5th International Conference, IScIDE 2015 Jun 14-16; Suzhou, China: Springer International Publishing; 2015. p. 258-66.

2. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136(5):E359-86.

3. Mittal D, Gaurav D, Roy SS. An effective hybridized classifier for breast cancer diagnosis. IEEE International Conference on Advanced Intelligent Mechatronics (AIM); 2015 Jul 7-11; Busan, South Korea: IEEE; 2015. p. 1026-31.

4. Stewart BW, Kleihues P. World Cancer Report. International Agency for Research on Cancer Press; 2003. Available from: http://publications.iarc.fr/Non-Series-Publications/World-Cancer-Reports/World-Cancer-Report-2003

5. Dramicanin T, Lenhardt L, Zekovic I, Dramicanin MD. Support Vector Machine on fluorescence landscapes for breast cancer diagnostics. J Fluoresc 2012;22(5):1281-9.

6. Liou DM, Chang WP. Applying data mining for the analysis of breast cancer data. Methods Mol Biol 2015;1246:175-89.

7. Bazazeh D, Shubair R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA); 2016 Dec 6-8; Ras Al Khaimah, United Arab Emirates: IEEE; 2016.

8. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014;13:8-17.

9. Wang KJ, Makond B, Chen KH, Wang KM. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. Applied Soft Computing 2014;20:15-24.

10. Thongkam J, Xu G, Zhang Y, Huang F. Toward breast cancer survivability prediction models through improving training space. Expert Systems with Applications 2009;36(10):12200-9.

11. Kharya S. Using data mining techniques for diagnosis and prognosis of cancer disease International Journal of Computer Science, Engineering and Information Technology 2012; 2(2): 55-66.

12. Krishnaiah V, Narsimha DG, Chandra NS. Diagnosis of lung cancer prediction system using data mining classification techniques. International Journal of Computer Science and Information Technologies 2013;4(1):39-45.

13. Kumar KS, Sitamahalakshmi T. Performance variation of support vector machine and probabilistic neural network in classification of cancer datasets. International Journal of Applied Engineering Research 2016;11(4):2224-34.

14. Mert A, Kılıç N, Bilgili E, Akan A. Breast cancer detection with reduced feature set. Computational and Mathematical Methods in Medicine 2015;2015:11.

15. Sewak M, Vaidya P, Chan CC, Zhong-Hui D. SVM Approach to Breast Cancer Classification. Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS); 2007 Aug 13-15; Iowa City, IA, USA: IEEE; 2007.

16. Goodarzi M, dos Santos Coelho L. Firefly as a novel swarm intelligence variable selection method in spectroscopy. Anal Chim Acta 2014;852:20-7.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2018; 5(2): 244-251

Evaluation of the Effect of Feature Selection and Different kernel Functions on SVM Performance for Breast Cancer Diagnosis

Orooji Azam[1], Langarizadeh Mostafa[2]*

• **Received:** 18 Feb, 2018 • **Accepted:** 12 Jul, 2018

Introduction: Breast cancer is one of the most common cancers affecting women. In mammography, differentiating a malignant tumor from a benign one is a very tedious task due to their structural similarities. Machine Learning (ML) is a subfield of Artificial Intelligence that can be used as an effective tool to help physicians to make decisions. Support vector machine (SVM) is one of the most common ML techniques that its performance depends on kernel parameters tuning and input features. The aim of this study was to investigate the effect of feature selection and different kernel functions on SVM performance.

Methods: This analytic study was performed through comparative method. Genetic algorithm was used for feature selection. SVM models based on different kernel functions, including polynomial, Linear, Radial Basis Function (RBF), Quadratic and Multi-Layer Perceptron (MLP), were first performed with all features and then, with the selected features. The Wisconsin original breast cancer data set was used as a training set to evaluate the performance of the classifiers. All implementations were done in MATLAB environment.

Results: According to the obtained results, by applying feature selection, the performance of SVM with MLP kernel function decreased and with quadratic kernel function increased. However, the performances of the linear and RBF kernels were desirable in both conditions. Generally, after the dimension reduction, the best accuracy, specificity, sensitivity and accuracy were dropped by 0.663, 0.833, 1.077 and 0.138 percent respectively.

Conclusion: The ML-based methods can help physicians in diagnosis and decision makings for treatment.

Keywords: Machine Learning, Dimension reduction, Clinical decision support systems, Early cancer diagnosis

• **Citation:** Orooji A, Langarizadeh M. Evaluation of the Effect of Feature Selection and Different kernel Functions on SVM Performance for Breast Cancer Diagnosis. Journal of Health and Biomedical Informatics 2018; 5(2): 244-251.

1. PhD Student in Medical Informatics, Health Information Management Dept., School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran .
2. PhD in Medical Informatics, Assistant Professor, Health Information Management Dept., School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran.

***Correspondence**: No. 6, Shahid Yasami St., Vali-e-Asr St., School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran.
• **Tel:** 02188794301 • **Email:** Langarizadeh.m@iums.ac.ir