روش‌های ارزیابی سیستم‌های رایانه‌ای تشخیصی پزشکی: مروری بر روش‌های موجود

مهسا منصوریان، حمید رضا مراتب

مقاله مروری

چکیده

مقدمه: در حال حاضر سیستم‌های تشخیصی رایانه‌ای دارای کاربرد وسیعی در علوم پزشکی می‌باشند. این سیستم‌ها می‌توانند در تشخیص درست و به‌موقع بیماری‌ها به پزشک یاری رسانند. عملکرد این‌گونه سیستم‌ها باید به‌صورت مناسبی ارزیابی شود. در این مقاله، معیارهای ارزیابی سیستم‌های تشخیصی پزشکی مورد بررسی قرار می‌گیرد.

روش‌ها: با استفاده از داده استاندارد در هر روش تشخیصی، میزان خطای سیستم بر اساس پارامترهای حساسیت، ویژگی، صحت، دقت ، سطح زیر منحنی (ROC یا receiver operating characteristic) ، F-measure، Matthews correlation coefficient و ... محاسبه شد و نقاط قوت و ضعف هر معیار مورد بررسی قرار گرفت. معیارهای ارزیابی در یک مثال پزشکی بر روی مقایسه دو روش در تشخیص عارضه قلبی بیماران، محاسبه و ارزیابی شدند. هم‌چنین ارجحیت روش‌های تشخیصی به یکدیگر با استفاده از تست McNemar مشخص گردید.

یافته‌ها: به دلیل برابر بودن نسبی تعداد افراد واقعا سالم و مریض در مجموعه داده مورد استفاده، پارامتر صحت، معیار قابل قبولی برای ارزیابی کلی دو روش بود که در دو روش به ترتیب ۸۴ و ۸۶٪ محاسبه شد. هر دو روش از نظر پزشکی، قابل اعتماد نبودند چون میزان خطای نوع اول بیشتر از۰/۰۵ بود. با این حال، توان تشخیصی روش دوم چون بالای ۸۰٪ بود، قابل قبول می‌باشد. چون سطح زیر منحنی ROC در دو روش بین ۰/۸ و ۰/۹ بود، قدرت تشخیصی آن‌ها "بسیار خوب" می‌باشد. در نهایت، دو روش تشخیصی از نظر آماری معادل شناخته شدند.

نتیجه‌گیری: ارزیابی سیستم‌های تشخیصی باید با روش‌های مناسب و معیارهای مرتبط انجام گیرد. با بکارگیری معیارهای متعدد قادر خواهیم بود توانایی روش تشخیصی را از زوایای مختلف ارزیابی نماییم.

واژه‌های کلیدی: سیستم‌های تشخیصی رایانه‌ای، داده کاوی، اعتبار سنجی، بازشناخت الگو

ارجاع: منصوریان مهسا، مراتب حمید رضا. **روش‌های ارزیابی سیستم‌های رایانه‌ای تشخیصی پزشکی:مروری بر روش‌های موجود .** مجله تحقیقات نظام سلامت ۱۳۹۴؛ ۱۱(۳): ۴۵۸-۴۴۵

تاریخ دریافت: ۱۳۹۴/۰۶/۲۳ **تاریخ پذیرش: ۱۳۹۴/۰۸/۱۱**

۱. گروه فیزیک پزشکی، دانشکده ی پزشکی، دانشگاه تربیت مدرس، تهران، ایران

۲. استادیار گروه مهندسی پزشکی، دانشکده فنی مهندسی، دانشگاه اصفهان، اصفهان، ایران (نویسنده مسؤول)

Email: h.marateb@eng.ui.ac.ir

مقدمه

در جامعه‌ی پزشکی امروز ضرورت استفاده از انواع روش‌های تشخیصی که به درمان، کمک شایان توجهی می‌کنند امری اجتناب‌ناپذیر به شمار می‌رود چرا که تشخیص در پزشکی، امری بس پیچیده و دشوار است که مبتنی بر تجربه بوده و کمتر به عنوان یک کار فکری تلقی می‌شود (۱). با این‌حال در این عرصه، همواره شکافی بین اطلاعات موجود برای تشخیص و آن‌چه که از حافظه قابل دسترسی است، حتی

برای پزشکی که به خوبی آموزش دیده و روزانه با انواع قابل توجهی از بیماری‌ها و اختلالات سر و کار دارد به طور محسوسی موجود است. بنابرین لزوما استفاده‌ی بهینه، از اطلاعات حاصله از روش‌های تشخیصی، توسط پزشکان صورت نمی‌پذیرد (۳-۲). محدودیت‌های سیستم چشم– مغز انسان، محدودیت‌های یادگیری و تجربه، ظرفیت کم حافظه‌ی کوتاه مدت و عواملی چون خستگی و حواس پرتی، استفاده از اطلاعات در دسترس را به کمتر از حد مطلوب می‌رساند (۶-۴). در جهت جبران این محدودیت‌ها، سیستم‌هایی با استفاده از رایانه ایجاد شدند.

مزیت سیستم‌های تشخیصی رایانه‌ای نسبت به روش‌های سنتی تشخیصی، آن است که بدون استفاده از آزمایشات بالینی متداول حاوی آیتم‌های فراوان هزینه بر مبتنی بر سعی و خطا، پارامترهایی از بیمار را مورد پایش قرار دهد که در تشخیص کارآمدتر باشند. بر خلاف سیستم‌های سنتی موجود که در آن‌ها تصمیم‌گیری با عملکرد تکراری و متداول آیتم‌های مختلف از بیمار بدون در نظرگیری روابط منطقی میان آن‌ها صورت می‌گیرد، انعطاف‌پذیری این سیستم‌ها به نحوی است که قادر خواهد بود با ترکیب دانش‌های ریاضی، کامپیوتر و پزشکی مرتبط با بیماری مورد بررسی آیتم‌های ثبت‌شده را به نحوی مدیریت نماید که حداقل زمان، هزینه و خطا در تشخیص بیماری حاصل گردد.

سیستم‌های تشخیص و آشکارسازی به کمک کامپیوتر (CAD یا Computer-Aided Diagnosis) در فرایند تشخیص می‌تواند زمان و انرژی صرف‌شده را در مقایسه با روش دستی کاهش دهند و اطلاعات موقعیتی آسیبی خاص و یا سایر تحلیل‌های تشخیصی را با هدف کمک به پزشکان تأمین نمایند (۷). مزایای اصلی کامپیوتر، توانایی آن در ذخیره‌ی مقدار زیادی از اطلاعات بدون تغییر در زمان‌های طولانی، خوانش اطلاعات، دریافت پیام مناسب دقیقا همان‌طور که ذخیره می‌شود، انجام عملیات ریاضی با سرعت بسیار بالا و نمایش بسیاری از امکانات تشخیصی به شکل منظم است (۸). علاوه بر این، ارزیابی بیماری، نوع بیماری، شدت، مرحله، پیشرفت و یا برگشت بیماری، هدف دیگری است که توسط سیستم CAD برای پزشکان تأمین می‌شود. این سیستم‌ها خروجی‌ها را بر اساس اطلاعات به دست آمده از منابع مختلف، که تصاویر پزشکی در رأس آن‌ها قرار دارند و با استفاده از روش‌های مختلف گرفته می‌شوند، محاسبه می‌کند. سیستم‌های CAD تفسیرهای مداومی از تصاویر پزشکی فراهم می‌کند تا دقت یک تشخیص را بهبود بخشد (۷).

ارزیابی سیستم‌های CAD، یکی از مشکلات عمده‌ای است که این سیستم‌ها درتوسعه‌ی خود با آن مواجه هستند. از آن‌جایی که نتایج می‌تواند بسته به مجموعه اطلاعات و یا تصاویر استفاده شده تغییر کند، تعیین اثربخشی یک تکنیک خاص کار ساده‌ای نیست. علاوه بر این، برای تعیین امکان‌پذیر بودن اجرای یک تکنیک، آزمون‌ها بایستی با مجموعه‌ای از داده‌ها و یا تصاویری که ترجیحا با ویژگی‌های مختلفی به دست آمده‌اند انجام شود. هم‌چنین این مجموعه تصاویر باید الزامات هدف تکنیک را برآورده سازند به این معنا که بایستی شامل ساختارهایی باشند که سیستم کامپیوتری آن‌ها را جستجو می‌کند (۹). با توجه به پیچیدگی سیستم‌های تشخیصی، ارزیابی آن‌ها از اهمیت خاصی برخوردار است. ارزیابی مناسب این سیستم‌ها، به قابل قبول بودنشان در جامعه پزشکی کمک می‌کند.

مطالعات زیادی تاکنون انجام شده‌اند که از روش‌های متعددی جهت ارزیابی سیستم‌های CAD استفاده کرده‌اند. این روش‌ها شامل همپوشانی، ضریب Dice، منحنی ROC، منحنی FROC ، صحت، حساسیت و سایر روش‌ها است (۱۰-۹). هریک در کنار داشتن مزایای خاص خود، معایبی نیز به همراه دارد. از جمله معایب این روش‌ها لزوم مشارکت مکرر و خسته‌کننده‌ی پزشکان و رادیولوژیست‌ها در تأیید صحت هریک از نسخه‌های سیستم CAD است. در جهت کاهش این محدودیت، مطالعه‌ی حاضر قصد دارد تا روش‌های ارزیابی سیستم‌های CAD را در تشخیص‌های دو کلاسه و چند کلاسه بررسی کرده و خطاهای مختلف آماری

را در ارزیابی سیستم‌ها آزموده و با معیارهای مناسب معرفی کند. در فرایند تست فرضیه آماری، اگر یک فرد سالم به اشتباه توسط سیستم تشخیصی رایانه‌ای بیمار شناسایی شود خطای اول آماری (Type I error) رخ داده است. در حالتی که فرد بیمار، سالم شناسایی شود خطای دوم آماری (Type II error) رخ داده است. در حالت پیشرفته، اگر بیمار به درست شناسایی ولی دلیل این تشخیص اشتباه بلشد، خطای سوم آماری (Type III error) پدید آمده است (۱۱) . در این مطالعه، نقاط قوت و ضعف و همچنین موارد کاربرد هر روش (معیار) ذکر شده و نکات مهم در ارزیابی صحیح سیستم‌های تشخیصی بیان می‌شود. لازم به ذکر است که معیارهای مورد بررسی در حالت دو کلاسه (خروجی سالم و مریض) بیان می‌شود که تعمیم آن‌ها به حالت چند کلاسه با ذکر منبع به صورت مختصر توضیح داده می‌شود.

روش‌ها

سودمندی تست‌های تشخیصی که توانایی آن‌ها در آشکار کردن فردی با بیماری یا جدا کردن افراد سالم است توسط روش‌های متعددی صورت می‌گیرد. مفید بودن یک آزمون تشخیصی در پیشگویی به موارد زیر وابسته است: ۱- تعداد موارد بیماری که به درستی تشخیص داده شده است (مثبت درست یا TP (True Positive)) ۲- تعدادی که نداشتن بیماری به درستی تشخیص داده شده است (منفی درست یا TN (True Negative)) ۳- تعدادی که داشتن بیماری به اشتباه تعیین شده است (مثبت نادرست یا FP (False Positive)) ۴- تعدادی که نداشتن بیماری به اشتباه مشخص شده است (منفی نادرست یا FN (False Negative)). بر اساس این چهار نتیجه می‌توان یک جدول احتمال وقوع ۲×۲ برای هر آزمون تشخیصی ایجاد کرد که توانایی آن را در پیشگویی درست وجود بیماری به طور مختصر بیان می‌کند. موارد بالا به صورت خلاصه در جدول (۱) نشان داده شده‌اند. در این جدول، مثبت یا منفی بودن جواب یک سیستم تشخیصی رایانه‌ای، در سطرها به صورت مثبت یا منفی و سالم یا بیمار بودن توسط روش استاندارد در ستون‌ها مشخص می‌شود. با استفاده از این پارامترها، معیارهای ارزیابی (۱۲،۹) تعریف می‌شوند که در جدول (۲) معرفی شده و در ادامه توضیح داده می‌شوند. تعمیم جدول ۱ و روش‌های اعتبار سنجی آن (جدول ۲) به حالت چند کلاسه (خروجی سیستم تشخیصی بیش از سه حالت باشد مثلا موارد حاد بیماری، موارد متوسط و موارد ضعیف) در بخش بحث و نتیجه‌گیری توضیح داده می‌شود.

جدول ۱. جدول احتمال وقوع Contingency Table (Confusion Matrix) (پارامترهای ارزیابی) در سیستم تشخیصی با دو خروجی بیمار بودن (مثبت یا منفی)

	مورد بیمار	مورد سالم
مثبت	**مثبت درست (TP)**	مثبت نادرست(FP)
منفی	**منفی نادرست (FN)**	منفی درست (TN)

جدول ۲. معیارهای ارزیابی سیستم تشخیصی دو کلاسه و فرمول‌های مربوطه

معیار مورد بررسی	رابطه ریاضی	توضیح
حساسیت Sensitivity (=Recall, True Positive Rate, Statistical Power,1-type II error)	$\frac{TP}{TP+FN}\times 100\%$	درصد بیمارانی که به درستی شناسایی شده‌اند.
ویژگی Specificity	$\frac{TN}{TN+FP}\times 100\%$	درصد افراد سالم که به درستی شناسایی شده‌اند.

صحت Accuracy (=Correct Classification, Detection Rate)	$\frac{TP+TN}{TP+TN+FP+FN}\times 100\%$	درصد افراد سالم و بیماری که به درستی تشخیص داده شده‌اند.
دقت Precision (=Positive Predictive Value i.e. PPV)	$\frac{TP}{TP+FP}\times 100\%$	درصد افراد شناسایی شده که واقعا بیمار هستند.
NPV Negative Predictive Value	$\frac{TN}{TN+FN}\times 100\%$	درصد افراد شناسایی شده که واقعا سالم هستند.
نمودار ROC Received Operating Characteristic Curve	---	نموداری که حساسیت سیستم تشخیصی را نسبت به ۱–ویژگی با تغییر آستانه سیستم نشان می‌دهد.
سطح زیر منحنی ROC (AUC)´	$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$	نشان دهنده قدرت سیستم تشخیصی
نرخ مثبت نادرست False Positive Rate (=False Alarm)	$\frac{FP}{FP+TN}$	سطح معنی‌داری (Significance level) که برابر با ۱– ویژگی است.
نرخ منفی نادرست False Negative Rate	$\frac{FN}{TP+FN}$	خطای نوع دوم آماری که برابر با ۱–حساسیت است.
نسبت شانس مثبت Positive likelihood Ratio	$LR^{+}=\frac{Sensitivity}{1-Specificity}$	برای افرادی که نتیجه تست آن‌ها مثبت است، نسبت احتمال آن‌که فرد واقعا مریض باشد به سالم بودن فرد.
نسبت شانس منفی Negative likelihood Ratio	$LR^{-}=\frac{1-Sensitivity}{Specificity}$	برای افرادی که نتیجه تست آن‌ها منفی است، نسبت احتمال آن‌که فرد واقعا مریض باشد به سالم بودن فرد.
همپوشانی Overlap	$\frac{\lvert A_{seg}\cap A_{ref}\rvert}{\lvert A_{seg}\cup A_{ref}\rvert}$	میزان سطح نسبی همپوشانی دو ناحیه

خطای همپوشانی سطح Area Overlap Error	$1-\frac{\lvert A_{seg}\cap A_{ref}\rvert}{\lvert A_{seg}\cup A_{ref}\rvert}$	تصویر میزان خطای سطح نسبی دو ناحیه تصویر
تفاوت نسبی سطح Relative Area Difference	$\frac{\lvert A_{seg}\rvert-\lvert A_{ref}\rvert}{\lvert A_{ref}\rvert}$	تفاوت نسبی سطوح تفکیک شده خودکار و دستی
ضریب Dice Dice's Coefficient	$\frac{2\times\lvert A_{seg}\cap A_{ref}\rvert}{\lvert A_{seg}\rvert+\lvert A_{ref}\rvert}$	میزان همپوشانی دو ناحیه تصویر تفکیک شده خودکار و دستی
معیار F_1 F_1 measure	$\frac{2\times TP}{2\times TP+FN+FP}$	میانگین هارمونیک حساسیت و دقت
DOR Diagnostic Odds Ratio	$\frac{LR^+}{LR^-}$	نسبت شانس آن‌که تست مثبت است اگر فرد واقعا مریض باشد به آن‌که فرد واقعا سالم باشد.

A_{ref} ناحیه‌ی تقسیم‌بندی شده توسط سیستم و A_{seg} ناحیه‌ی تقسیم‌بندی درست است. معیارهایی که در آن‌ها سطح ناحیه استفاده شده است در حقیقت در سیستم‌های تفکیک تصویر (Image Segmentation Systems) به منظور تطابق ناحیه تقسیم‌بندی شده توسط سیستم خودکار و ناحیه درست که توسط متخصص انجام شده است، استفاده می‌شوند.

حساسیت

حساسیت یک تست به عنوان نسبتی از افراد بیمار که دارای نتیجه مثبت هستند را تعریف می‌کند. حساسیت یک تست تنها می‌گوید که یک تست به چه اندازه توانایی مشخص کردن افراد بیمار را در میان افرادی که برای آن‌ها بیماری تشخیص داده شده دارا است. حساسیت، درباره این که آیا برخی از افراد بدون بیماری، تست مثبت داشته‌اند یا نه، و در صورت وجود به چه نسبتی، چیزی نمی‌گوید. بر مبنای حساسیت، خطای نوع دوم آماری و هم‌چنین توان تشخیصی قابل تعریف است. حساسیت با توان تشخیصی آماری برابر است. یکی از شرایط تست تشخیصی خوب، آن است که توان آن بالاتر از ۸۰٪ باشد.

ویژگی

در تعریف ویژگی تنها نسبتی از افراد بدون بیماری که تست آن‌ها منفی بوده است مورد نظر است. ویژگی تنها می‌تواند از افرادی محاسبه شود که بیماری ندارند. بنابراین درباره‌ی این‌که آیا برخی از افراد با بیماری که نتیجه‌ی آن‌ها منفی بوده و در صورت وجود به چه نسبتی، چیزی نمی‌گوید (۱۳). جمع میزان خطای نوع اول آماری با پارامتر ویژگی برابر با یک است. یکی دیگر از شرایط قابل قبول بودن یک سیستم تشخیصی، آن است که میزان خطای نوع اول کمتر از ۰/۰۵ باشد.

محدودیت اصلی پارامترهای حساسیت و ویژگی آن است که در تخمین احتمال بیماری برای هریک از بیماران، کاربرد عملی ندارد. وقتی که بیمار با نتیجه‌ی مثبت از یک تست تشخیصی به پزشک مراجعه می‌کند، انتظار می‌رود که به این سؤال پاسخ داده شود که "احتمال بیماری در تست مثبت چقدر است؟ " علت این که این دو کمیت قادر نیستند تا به این سؤال پاسخ

گویند آن است که این دو بر اساس افراد دارای بیماری یا بدون بیماری تعریف می‌شود. اما از آن‌جایی که بیمار با مجموعه‌ای از علائم و نشانه‌ها مراجعه می‌کند نه یک تشخیص، پزشک نمی‌داند که آیا در آن زمان بیمار دارای بیماری بوده یا نه و بنابراین این پارامترها نمی‌تواند برای آن‌ها اعمال شود (۱۴). حساسیت و ویژگی هیچ یک متأثر از شیوع بیماری نیست، به این معنی که نتایج از یک مطالعه می‌تواند به آسانی به مجموعه‌ای دیگر با شیوع متفاوتی از بیماری در جمعیت انتقال یابد. با این‌حال حساسیت و ویژگی ممکن است بسته به طیف بیماری در گروه مطالعه شده تغییر یابند (۱۵). حساسیت و ویژگی یک تست کمی وابسته به مقدار نقطه برش (Cut-off) است. این مقدار، تعیین‌کننده‌ی حد بین نتایج مثبت و منفی است. در یک موقعیت وقتی نقطه برش کاهش می‌یابد در صورتی‌که نقطه برش کاهش یابد، افراد بیشتری با بیماری، به درستی مشخص می‌شوند اما در عین حال مثبت‌های اشتباه افزایش خواهند یافت. با افزایش مقدار نقطه برش منفی‌های اشتباه بیشتری نشان داده می‌شود اما تعداد مثبت‌های اشتباه کاهش می‌یابد (۱۶). حساسیت و ویژگی به ترتیب به خطاهای نوع دوم و اول بستگی دارند که در یک سیستم تشخیصی دارای اهمیت فراوان هستند. آن‌ها بر خلاف معیارهای NPV و PPV به میزان شیوع بیماری در جامعه وابسته نیستند (۱۷). در یک تست تشخیصی معمولا مقدار حساسیت بالا همراه با ویژگی پایین است و بالعکس. میزان بالای هر معیار بر اساس آن‌که تشخیص افراد بیمار مهم‌تر است یا سالم، معین می‌شود.

صحت

این معیار برابر است با نسبت افرادی که به درستی دسته‌بندی شده‌اند به کلیه‌ی افراد. با حساسیت و ویژگی مشابه، صحت یک تست خاص با کاهش شیوع بیماری افزایش می‌یابد. اما این به این معنی نیست که اگر آزمون را در جمعیتی با شیوع بیماری کم اعمال کنیم بهتر است. بلکه تنها به این معنی است که در یک تعداد مشخص، آزمون، افراد را به شکل درست‌تری طبقه‌بندی می‌کند. این پارامتر، زمانی که تعداد واقعی نمونه‌های سالم و مریض خیلی متفاوت باشد اصلا معیار مناسبی نخواهد بود چرا که بر اساس تعریف (جدول ۲) صحت را بیش از آنچه هست نشان می‌دهد. به عنوان مثال فرض کنید TP=1، TN=50، FN=10 و FP=9 باشد. مقدار صحت برابر است با ۷۳٪ و این در حالی است که حساسیت و ویژگی این روش تشخیصی به ترتیب برابر است با ۹٪ و ۸۵٪. این تست به هیچ عنوان افراد مریض را به درستی شناسایی نمی‌کند. دلیل این امر آن است که تعداد افراد واقعا سالم و مریض به ترتیب برابر ۵۹ و ۱۱ نفر هستند و داده های مورد بررسی متوازن نیست. در این مواقع، از معیار F1 یا Matthews correlation coefficient (۱۸) می‌توان استفاده کرد. تعریف آن به صورت زیر است:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$$

در مثال مورد بررسی، پارامتر MCC برابر است با ۰/۰۶− . این معیار بین ۱− تا ۱ بوده و هرچه به یک نزدیک‌تر باشد بهتر است و مقادیر نزدیک به صفر، نشان‌دهنده آن است که سیستم تشخیصی همانند یک سیستم تصادفی کار می‌کند و در نتیجه اصلا قابل قبول نمی‌باشد.

نسبت نابرابری تشخیصی (DOR یا Diagnostic Odds Ratio)

DOR یا نسبت نابرابری تشخیصی نیز یک معیار جهانی برای درستی تشخیص است که تخمینی از توان تفکیک روش‌های تشخیصی و هم‌چنین مقایسه‌ی دقت‌های تشخیصی بین دو یا چند آزمون تشخیصی است. DOR یک آزمون نسبت نابرابری مثبت در افراد با بیماری به نابرابری در افراد بدون بیماری است. DOR به طور معنی‌داری به حساسیت و ویژگی وابسته است. آزمونی با حساسیت و ویژگی بالا و مقادیر FPR و FNR پایین، دارای DOR بالایی است. در صورت وجود حساسیت مشابه، DOR آزمون با افزایش ویژگی افزیش می‌یابد (۱۹). DOR وابسته به شیوع بیماری نیست اما مشابه حساسیت و ویژگی وابسته به معیارهای استفاده شده در تعیین بیماری و طیف شرایط پاتولوژیکی گروه مورد مطالعه‌اش (شدت بیماری، فاز، مرحله و ...) است (۱۵).

شاخص Youden

شاخص Youden یکی از قدیمی‌ترین اندازه‌ها جهت تعیین دقت تشخیص است. این شاخص یک مقیاس جهانی است که جهت ارزیابی توان تفکیک کلی یک روش تشخیصی و مقایسه‌ی این تست با تست‌های دیگر مورد استفاده قرار می‌گیرد (۲۰). شاخص youden با کسر ۱ از مجموع حساسیت و ویژگی، محاسبه شده و نه به شکل درصد، بلکه به صورت قسمتی از تعداد کل بیان می‌شود. برای آزمونی که دقت تشخیصی ضعیفی دارد شاخص youden برابر صفر و در یک آزمون کامل برابر ۱ است. این شاخص به اختلافات در حساسیت و ویژگی تست حساس نیست که از مهم‌ترین عدم مزایای آن به شمار می‌رود. به عنوان مثال یک آزمون با حساسیت ۰/۹ و ویژگی ۰/۴ شاخص Youdin مشابه ۰/۳ را با آزمونی دارد که دارای حساسیت ۰/۶ و ویژگی ۰/۷ است. واضح است که این تست‌ها ازدرستی تشخیص قابل مقایسه‌ای برخوردار نیستند. این شاخص از شیوع بیماری متأثر نمی‌شود، اما از طیف بیماری و همچنین ویژگی، حساسیت، نسبت‌های همسایگی و DOR تأثیر می‌پذیرد (۱۵).

منحنی ROC

یک جفت از مقادیر ویژگی و حساسیت برای هر مقدار نقطه برش وجود دارد. برای تشکیل یک منحنی ROC، این جفت از مقادیر را روی یک گراف با 1- specificity بر روی محور x و حساسیت بر روی محور Y رسم می‌کنیم. به عبارت دیگر در این گراف دو بعدی, TPR (True Positive Rate) که برابر نسبت مثبت‌های درست طبقه‌بندی شده به کل مثبت‌ها است، روی محور Y، و FPR (False Positive Rate) که برابر نسبت منفی‌های نادرست دسته‌بندی شده به کل منفی‌ها است روی محور x قرار دارد (۱۵).

یک نقطه در فضای ROC در صورتی‌که در بخش شمال غربی قرار گیرد (TPR بالاتر، FPR پایین‌تر) بهتر از سایر نقاط است. طبقه‌بندی کننده که در سمت چپ یک گراف ROC نزدیک محور X قرار گیرد دسته‌های مثبت را تنها با شاهد قوی تشکیل می‌دهند و بنابراین خطاهای مثبت اشتباه کمی داشته و هم‌چنین TPR کمی دارند. دسته‌کننده‌هایی که در سمت راست بالای یک گراف ROC قرار می‌گیرند دسته‌های مثبت را تنها با شاهد ضعیف تشکیل می‌دهند اما اغلب FPR بالایی دارند (۲۱).

در تصمیمات پزشکی، آنالیز ROC عموما در مسائل تشخیصی دو کلاسه (وجود یا عدم شرایط ناهنجار) مورد استفاده قرار می‌گیرد. دو محور، بین خطاها (FP) و فوایدی (TP)که یک دسته‌کننده، بین دو کلاس تشکیل می‌دهد مصالحه‌ای (trade off) را انجام می‌دهد. بیشتر آنالیزها به دلیل تقارنی که در مسائل دو کلاسه وجود دارد سرراست است.

با بیش از دو کلاس، وضعیت پیچیده‌تر می‌شود. با n کلاس، ماتریس confusion یک ماتریس n×n می‌شود که حاوی n دسته‌ی صحیح (مؤلفه‌های قطر اصلی) و $n^2 - n$ خطای ممکن (مؤلفه‌های غیر قطری) است. به جای مصالحه‌ی بین TP و FP ، n فایده و $n^2 - n$ خطا وجود دارد. n کلاس، ایجاد n گراف ROC متفاوت (یک گراف برای هر کلاس) است که این روش، فرمولاسیون مرجع کلاس (Class Reference) نامیده می‌شود (۲۱).

سطح زیر منحنی ROC (AUC)

شکل منحنی ROC و سطح زیر منحنی آن (AUC) کمک می‌کند تا میزان توان تفکیک یک تست تخمین زده شود. سطح زیر منحنی مقداری بین صفر و یک دارد و بیانگر میزان خوب بودن تست است. یک تست تخیصی کامل دارای مقدار سطح زیر منحنی ۱ است در حالیکه یک تست غیر تفکیکی مقدار ۰/۵ دارد. در صورتی که AUC نزدیک به صفر باشد، نشان دهنده آن است که خروجی سیستم تشخیصی عکس میزان واقعی است. عموما ارتباط بین سطح زیر منحنی و درستی تشخیص مطابق جدول ۳ اعمال می‌شود:

جدول ۳. تفسیر میزان درستی تشخیص با استفاده از سطح زیر منحنی ROC

درستی تشخیص	سطح زیر منحنی
عالی	۰/۹-۱/۰
بسیار خوب	۰/۸-۰/۹
خوب	۰/۷-۰/۸
کافی	۰/۶-۰/۷
بد	۰/۵-۰/۶
آزمون مفید نیست	< ۰/۵

ارتباط بین سطح زیر منحنی ROC و درستی تشخیص

مساحت زیر منحنی ROC اندازه‌ای ازدرستی تشخیص است که چیزی در مورد پارامترهایی مثل حساسیت و ویژگی نمی‌گوید. از دو تست با سطح زیر منحنی مشخص و مشابه، یکی می‌تواند حساسیت بالایی داشته باشد در حالی‌که دیگری ویژگی بالایی را دارا باشد. علاوه بر این، سطح زیر منحنی درباره‌ی مقادیر پیشگویانه و سهم تست در تشخیص Ruling in)و یا رد (Ruling out) بیماری چیزی بیان نمی‌کند.

با مقایسه‌ی سطح زیر دو منحنی ROC میتوان تخمین زد که کدام‌یک از دو تست برای تشخیص بیماری از غیر بیماری و یا هر نوع از دو شرایط مورد نظر مناسب‌تر است. بایستی به این نکته اشاره کرد که این مقایسه نباید بر اساس ارزیابی بصری یا حسی باشد. برای دستیابی به این هدف می‌توان از آزمون‌های آماری که سطوح معنی‌داری آن از قبل مشخص شده‌است جهت تخمین معنی‌داری اختلاف بین دو منحنی استفاده نمود (۱۵).

همان طور که گفته شد AUC اندازه‌ای از افتراق‌پذیری یک جفت کلاس است. در یک مسأله‌ی دو کلاسه، AUC یک مقدار واحد است اما در یک مسأله چند کلاسه، ترکیبی دو به دو از چندین مقدار افتراق‌پذیری است. یک روش برای محاسبه‌ی AUC های چندکلاسه، تولید منحنی ROC برای هرکلاس، اندازه‌گیری سطح زیر منحنی و سپس جمع AUC هایی است که با توجه به درجه‌ی کلاس در داده‌ها وزن‌دهی شده‌اند (۲۱).

مقادیر پیشگویانه و نرخ‌های حاشیه‌ای

هدف کلی یک تست تشخیصی آن است که نتایجش برای یک تشخیص مورد استفاده قرار گیرد. بنابراین نیاز است تا از احتمال این‌که نتیجه‌ی تست تشخیص درست را بدهد آگاه باشیم. مقادیر پیشگویی‌کننده‌ی مثبت و منفی،احتمال بیماری فرد را به محض مشخص شدن نتیجه‌ی تست وی توصیف می‌کند (۲۲).

مقدار پیشگویانه مثبت (PPV) ، احتمال داشتن بیماری مورد نظر در یک نمونه با نتیجه‌ی مثبت است و بنابراین نسبتی از بیماران با نتیجه‌ی تست مثبت در کل نمونه‌ها با نتیجه‌ی مثبت اراده می‌کند.

مقدار پیشگویانه منفی (NPV) ، احتمال نبود یک بیماری با نتیجه‌ی تست منفی را توصیف می‌کند. NPV به عنوان نسبتی از موارد بدون بیماری با نتیجه‌ی تست منفی، در کل نمونه‌ها با نتایج تست منفی است (۱۰).

برخلاف حساسیت و ویژگی، مقادیر پیشگویانه به مقدار زیادی وابسته به شیوع بیماری در جمعیت مورد بررسی است. بنابراین مقادیر پیشگویانه از یک مطالعه نباید به مجموعه دیگری با شیوع متفاوت بیماری در جمعیت انتقال یابد. شیوع به طور متفاوتی بر PPV و NPV اثر می‌گذارد. با افزایش شیوع بیماری در جمعیت PPV افزایش می‌یابد در حالی‌که NPV کاهش می‌یابد. با این‌که، تغییر در PPV قابل توجه‌تر است، با این‌که NPV تا حدی ضعیف‌تر از شیوع بیماری متأثر می‌شود (۱۵). این دو معیار بر روی سطرهای ماتریس

CONFUSION متمرکز هستند. دو معیار دیگر، نرخ مثبت حقیقی (TPR) و نرخ منفی حقیقی (TNR)، ستون‌های ماتریس confusion را در نظر می‌گیرند (کلاس‌های درست).

TNR و PPV مربوط به خطای نوع اول هستند در حالی‌که TPR و NPV مرتبط با خطای نوع دوم است. هر چهار پارامتر در محدوده‌ی صفر تا یک قرار دارند (۱۰).

نرخ مثبت کاذب (FPR) و نرخ منفی کاذب (FNR) دو معیار دیگر هستند که برای اندازه‌گیری خطاهای انجام‌شده توسط روش تقسیم‌بندی به کار می‌رود. FPR تعداد مواردی است که به طور نادرست، متعلق به دسته‌ی مثبت برچسب‌گذاری شده‌اند و FNR مواردی است که متعلق به دسته‌ی مثبت برچسب‌گذاری نشده‌اند.

نسبت شانس

نسبت همسایگی LR اندازه‌گیری مفیدی از درستی تشخیص است. این معیار به صورت نسبت نتیجه‌ی تست مورد نظر در نمونه‌ها با بیماری مشخص به نمونه‌ها بدون بیمار بیان می‌شود. LR مستقیما احتمال پیش از تست و بعد از تست بیماری را در یک بیمار مشخص می‌کند.

نسبت همسایگی برای نتایج تست مثبت (+ LR) بیانگر آن است که احتمالا چه مقدار نتیجه تست مثبت در افراد با بیماری در مقایسه با افراد بدون بیماری رخ می‌دهد. نسبت همسایگی مثبت معمولا مقداری بیشتر از یک دارد چرا که احتمال این‌که نتیجه‌ی تست مثبت در افراد با بیماری رخ دهد احتمالا بیشتر از افراد بدون بیماری است.

نسبت همسایگی برای نتیجه‌ی تست منفی(- LR) نسبتی از احتمال وقوع نتیجه منفی در نمونه‌ها با بیماری به احتمال وقوع همان نتیجه در نمونه‌های بدون بیماری است. نسبت همسایگی منفی معمولا مقداری کمتر از یک دارد چرا که احتمال رخداد نتیجه‌ی تست منفی در نمونه‌های بیمار کمتر از افراد بدون بیماری است.

از آن‌جایی که هر دو معیار حساسیت و ویژگی برای محاسبه‌ی نسبت همسایگی به کار می‌رود هیچ یک از نسبت شانس‌ها به شیوع بیماری در نمونه‌های مورد آزمایش بستگی ندارد (۲۳).

F_1-measure

معیار دیگر ارزیابی، F_1-measure است که به عنوان اندازه‌ای از همپوشانی بین کلاس‌های تخمین زده شده و کلاس‌های درست می‌تواند تفسیر شود. F_1-measure ، میانگین‌هارمونیک دقت و حساسیت بوده و محدوده‌ای از صفر (بدون همپوشانی) تا 1 (همپوشانی کامل) را دارا است (۱۰).

یافته‌ها

پارامترهای ارزیابی مورد بررسی به منظور بررسی دو روش تشخیصی عارضه قلبی (۲۵-۲۴) که دیتاست یکسانی استفاده کرده اند مورد استفاده قرار گرفت. در این دو روش، با استفاده از داده‌های تست‌های غیر تهاجمی و مشخصات فردی از قبیل سن و جنسیت، گرفتگی عروق کرونری قلب تخمین زده می‌شود و نتایج به دست آمده با روش مرجع آنژیوگرافی مقایسه می‌شود. نتایج ارزیابی در جدول ۴ خلاصه شده است. در دو روش مورد بررسی، پارامترهای جدول احتمال وقوع به صورت زیر است:

روش اول: TP=95; TN=134; FP=17; FN=26

روش دوم: TP=102; TN=131; FP=20; FN=19

جدول ۴: ارزیابی دو روش تشخیصی شناسایی بیماری قلبی

معیار ارزیابی	روش اول	روش دوم
حساسیت	٪۷۹	٪۸۴
ویژگی	٪۸۹	٪۸۷
صحت	٪۸۴	٪۸۶

دقت	٪۸۵	٪۸۴
NPV	۰/۸۵	٪۸۷
AUC	۰/۸۳	٪۸۶
FPR	۰/۱۱	۰/۱۳
FNR	۰/۲۲	۰/۱۶
LR+	۶/۹۷	۵/۱۴
DOR	۲۸/۲۷	۳۵/۰۸
توان تشخیصی آماری	٪۷۸/۵	٪۸۴/۳
میزان خطای نوع اول آماری	۰/۱۱	۰/۱۳

بر اساس جدول ۴، هیچ کدام از دو سیستم تشخیصی به طور کامل قابل اطمینان نیستند چرا که خطای نوع اول در آن‌ها کمتر از ۰/۰۵ نیست. با این حال، سیستم دوم از نظر توان تشخیصی قابل قبول است چراکه توان تشخیصی آن از ٪۸۰ بیشتر است. صحت دو سیستم بسیار به هم نزدیک است و چون تعداد افراد واقعا سالم و مریض به ترتیب برابر با ۱۵۱ و ۱۲۱ است، این پارامتر قابل قبول است. سطح زیر منحنی در هر دو روش بین ۰/۸ تا ۰/۹ بوده، پس درستی تشخیص "بسیار خوب" است (جدول ۳). معیار F_1 در این دو تست به ترتیب ۰/۸۱۵ و ۰/۸۳۹ می‌باشد. پارامتر MCC در این دو روش به ترتیب برابر با ۶۸ و ٪۷۱ می‌باشد که نزدیک به هم است.

اکنون آیا می‌توان به این سؤال، پاسخ گفت که کدام روش نسبت به دیگری ارجحیت دارد ؟

برای پاسخ به این سؤال باید از تست آماری مناسب که تست Gillick یا McNemar می‌باشد استفاده کرد (۲۷-۲۶). این تست به این سؤال پاسخ می‌دهد که از بین دو سیستم تشخیصی دو کلاسی A و B که در یک دیتاست یکسان استفاده شده‌اند، کدام یک از لحاظ آماری معنی‌دار در سطح خطای ۰/۰۵ بر دیگری برتری دارد. تست مورد نظر دارای خطای نوع اول و دوم مناسبی است که می‌تواند به صورت استاندارد در مقایسه آماری سیستم‌های تشخیصی استفاده شود (۲۷). تست مورد نظر به صورت زیر است:

فرض کنید می‌خواهیم دو سیستم تشخیصی A و B را با هم مقایسه کنیم. پارامترهای x و y به صورت زیر تعریف می‌شوند:

x= تعداد نمونه‌هایی که توسط A نادرست و توسط سیستم B درست شناسایی شده‌اند.

y تعداد نمونه‌هایی که توسط B نادرست و توسط سیستم A درست شناسایی شده‌اند.

سپس پارامتر z به صورت زیر تعریف می‌شود:

$$z=\frac{|x-y|-1}{\sqrt{x+y}}$$

در صورتی که شرط $|z|>1.96$ صادق باشد، دو سیستم A و B با سطح معنی‌داری ۰/۰۵ با هم متفاوت خواهند بود و در غیر این‌صورت معادل هستند. با مقایسه دو سیستم تشخیصی عارضه قلبی مورد بررسی، مشخص شد که هیچ یک از این دو تست به دیگری ارجحیت ندارد.

مقایسه چند کلاسه

در سیستم‌های تشخیصی چند کلاسه، پارامترهای صحت، حساسیت و دقت به صورت متفاوتی تعریف می‌شود (۲۸). به بیان دیگر، چون پارامتر TN همواره مقدار زیادی دارد از

فرمول‌های شامل این پارامتر استفاده نمی‌شود. به جای آن از F_1 باید استفاده کرد. به عنوان مثال به جدول احتمال ۵ توجه کنید که مربوط به یک سیستم تشخیصی فرضی سه کلاسه می‌باشد. در این سیستم، حالات سالم بودن و بیماری‌های ۱ و ۲ شناسایی می‌شوند.

جدول ۵. جدول احتمال وقوع (پارامترهای ارزیابی) در سیستم تشخیصی با سه خروجی سالم بودن و بیماری‌های ۱ و ۲

		داده‌های مرجع		
		سالم	بیماری ۱	بیماری ۲
خروجی سیستم تشخیصی	بیماری ۲	۲	۱۰	۸۸
	بیماری ۱	۶	۴۰	۱۴
	سالم	۳۲	۱۰	۱۸

در صورتی‌که به دنبال تعریف صحت برای خروجی سالم سیستم تشخیصی باشیم، بر اساس تعریف معیار صحت تمامی داده‌های موجود در جدول احتمالی ۵ به غیر از سطر و ستون آخر، پارامتر TN خواهد بود (حالت بیماری ۱ و ۲ ترکیب و مسأله به فرم دوحالتی تبدیل می‌شود). بر این اساس، صحت این سیستم تشخیصی بیش از مقدار واقعی تخمین می‌شود چون مقدار پارامتر TN زیاد بوده و در صورت و مخرج کسر استفاده می‌شود. در این مثال، صحت تشخیصی برای افراد سالم برابر با ۸۴٪ است که با در نظر گرفتن خطاهای موجود در شناسایی حالت سالم (خطاهای نوع اول و دوم)، منطقی نیست. می‌توان میزان صحت کلی سیستم را با محاسبه نسبت بین جمع عناصر روی قطر اصلی به مجموع همه عناصر ماتریس احتمالی به دست آورد (۷۳٪ در مثال مربوط به جدول ۵) ولی این معیار به هیچ وجه اطلاعاتی از عملکرد سیستم در هر خروجی تشخیصی نمی‌دهد.

برای رفع این مشکل، می‌توان معیار F_1 را برای هر کلاس محاسبه و ارائه داد (۲۸). در این صورت، از پارامتر TN استفاده نشده و مشکل ایجاد شده برای پارامتر صحت، مرتفع می‌شود. در مثال بالا، برای خروجی سالم بودن فرد، پارامترهای TP ، FP و FN به ترتیب برابر با ۳۲، (۱۰+۱۸) و (۲+۶) می‌باشند. معیار F_1 برابر با ۶۴٪ می‌باشد. یکی دیگر از مزایای استفاده از این معیار آن است که اگر پارامتر TP برای یک خروجی صفر بود، معیار F_1 نیز صفر خواهد بود که این نکته به دلیل وجود TN در تعریف معیار صحت برای هر کلاس، صدق نمی‌کند.

بحث

با توجه به نکات مطرح شده در این مقاله، ارزیابی دقیق سیستم تشخیصی رایانه‌ای از اهمیت بسیار زیادی برخوردار است. این سیستم‌ها به دلیل دارا بودن ساختار پیچیده، باید به درستی ارزیابی شوند تا کارکرد درست آن‌ها در عمل تضمین شود. در غیر این صورت، نه تنها به پزشک کمک نمی‌کند بلکه مشکلات زیادی را در فرایند تشخیص درست، ایجاد می‌کند. یکی از نکاتی که در این مقاله ذکر نشد، روش درست

آموزش سیستم تشخیصی و ارزیابی آن می‌باشد. سیستم‌های تشخیصی معمولا از یک سری داده آموزش (Training Set) برای تنظیم پارامترهای داخلی خود استفاده می‌کنند. ارزیابی این‌گونه سیستم‌ها باید بر روی داده‌هایی انجام گیرد که در فرایند آموزش، استفاده نشده‌اند. به این داده‌ها، مجموعه داده‌های اعتبارسنجی (Validation Set) گفته می‌شود.

در صورتی‌که تعداد نمونه‌های دو کلاس خروجی، قابل مقایسه نباشد باید به جای معیار صحت از معیارهای جایگزین F_1 یا MCC استفاده کرد. البته تأکید معیار F_1 بر روی شناسایی درست نمونه‌های مثبت است چون در آن از TP استفاده می‌شود. این درحالی است که نمونه‌های منفی نیز در بسیاری از موارد مهم هستند. بر این اساس، معیار MCC ترجیح داده می‌شود. نکته دیگر آن است که همراه با اینگونه معیارها، باید میزان خطای آماری نوع اول و توان تشخیصی آماری سیستم نیز ذکر شود. در این‌صورت، سیستم به صورت کامل ارزیابی می‌شود. در برخی از مقالات، از معیار توان تفکیک (Discriminant Power) ، در سیستم تشخیصی در حالت خروجی مثبت و منفی نیز استفاده شده است (۳۰-۲۹). با استفاده ازاین معیار، ارزیابی دقیقی از سیستم تشخیصی به عمل می‌آید. در سیستم‌های چند کلاسه، علاوه بر گزارش میزان صحت کلی، برای هر گروه تشخیصی نیز باید معیار F_1 به درستی محاسبه و گزارش شود.

نتیجه‌گیری

ارزیابی سیستم‌های تشخیصی باید با روش‌های مناسب و معیارهای مرتبط انجام گیرد. با بکارگیری معیارهای متعدد قادر خواهیم بود توانایی روش تشخیصی را از زوایای مختلف ارزیابی نماییم. همچنین ، استفاده از معیارهای نادرست می تواند کارایی یک سیستم تشخیصی را بیش از حد نشان دهد. در این مقاله ، این نکات با ذکر جزییات بررسی شده است .

References

1. Rogers W, Ryack B, Moeller G. Computer-aided medical diagnosis: literature review. International journal of bio-medical computing 1979;10(4):267-89. Epub 1979.08.01.
2. Lusted LB. Logical Analysis in Roentgen Diagnosis: Memorial Fund Lecture 1. Radiology 1960;74(2):178-93.
3. Tuddenham WJ. Visual Search, Image Organization, and Reader Error in Roentgen Diagnosis: Studies of the Psychophysiology of Roentgen Image Perception Memorial Fund Lecture 1. Radiology 1962;78(5):694-704.
4. Kundel HL, Revesz G. Lesion conspicuity, structured noise, and film reader error. American Journal of Roentgenology 1976;126(6):1233-8.
5. BERBAUM KS, FRANKEN Jr EA, DORFMAN DD, ROOHOLAMINI SA, KATHOL MH, BARLOON TJ, et al. Satisfaction of search in diagnostic radiology. Investigative Radiology 1990;25(2):133-40.
6. Renfrew D, Franken Jr E, Berbaum K, Weigelt F, Abu-Yousef M. Error in radiology: classification and lessons in 182 cases presented at a problem case conference. Radiology 1992;183(1):145-50.
7. Petrick N, Sahiner B, Armato SG, Bert A, Correale L, Delsanto S, et al. Evaluation of computer-aided detection and diagnosis systemsa). Medical Physics 2013;40(8):087001.
8. Gorry G, Barnett G. SEquential diagnosis by computer. JAMA 1968;205(12):849-54.
9. Gonçalves VM, Delamaro ME, Nunes FdLdS. A systematic review on the evaluation and characteristics of computer-aided diagnosis systems. Revista Brasileira de Engenharia Biomédica. 2014;30(4):355-83.
10. Labatut V, Cherifi H. Accuracy measures for the comparison of classifiers. 2012. [online].available from: http://arxiv.org/abs/1207.3790 .
11. Onwuegbuzie AJ, Daniel LG. Typology of analytical and interpretational errors in quantitative and qualitative educational research. Current Issues in Education 2003;6(2).

12. Kohavi R, Provost F. Glossary of terms. Machine Learning 1998;30(2-3):271-4.
13. Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar, Ravi Thomas. Understanding sensitivity, specificity and predictive values. Indian J Ophthalmol 2014;71(11):1062-5.
14. Wales NS. Moving beyond sensitivity and specificity: using likelihood ratios to help interpret diagnostic tests. Australian prescriber 2003;26(5): 111-13
15. Šimundić A-M. Measures of diagnostic accuracy: basic definitions. Med Biol Sci 2008;22(4):61-5.
16. Spitalnic S. Test properties I: Sensitivity, specificity, and predictive values. Hospital Physician. 2004;40:27-36.
17. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care & Pain 2008;8(6):221-3.
18. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 1975;405(2):442-51.
19. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. Journal of clinical epidemiology 2003;56(11):1129-35.
20. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. Biometrical journal Biometrische Zeitschrift 2008;50(3):419-30.
21. Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006;27(8):861-74.
22. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta paediatrica 2007;96(3):338-41.
23. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004;329(7458):168-9.
24. Marateb HR, Goudarzi S. A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system. Journal of Research in Medical Sciences 2015;20(3):214-23..
25. Negahbani M, Joulazadeh S, Marateb H, Mansourian M. Coronary Artery Disease Diagnosis Using Supervised Fuzzy C-Means with Differential Search Algorithm-based Generalized Minkowski Metrics. Peertechz J Biomed Eng 1 (1): 006. 2015;14(006).
26. Webb AR. Statistical pattern recognition. New York: John Wiley & Sons; 2003.
27. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation 1998;10(7):1895-923.
28. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management 2009;45(4):427-37.
29. Blakeley DD, Oddone EZ, Hasselblad V, Simel DL, Matchar DB. Noninvasive carotid artery testing: a meta-analytic review. Annals of internal medicine 1995;122(5):360-7.
30. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. AI 2006: Advances in Artificial Intelligence. Berlin: Springer; 2006. p. 1015-21.

Performance assessment of computer-aided diagnosis systems: A review of methodologies

Mahsa Mansorian [1], Hamid Reza Marateb [2]

Review Article

Abstract

Background: Nowadays, computer-aided diagnosis systems are widely used in medicine. These systems could assist medical doctors in early correct diagnosis of diseases. The performance of such systems must be correctly assessed. In this paper, the performance criteria of these diagnosis systems are taken into account.

Methods: The diagnosis error of such systems was estimated based on the gold standard data using measures such as Sensitivity, Specificity, Accuracy, Precision, Area Under curve ROC (receiver operating characteristic), F-measure, Matthews correlation coefficient , etc. The advantages and disadvantages of those criteria were also discussed. Then, the performance of two Coronary Artery Disease diagnosis systems was assessed. The statistically significant superior method was identified using the McNemar's test.

Findings: Since the analyzed dataset was balanced, the overall performance of the diagnosis methods was assessed using the accuracy measure. The accuracy of the methods was 84% and 86%, respectively. The entire systems were not reliable since Type I error (Alpha) was not less than 0.05. However, the second system had acceptable statistical power (>80%). The diagnosis performance of those systems was "very good" (0.8<AUC<0.9). None of the methods was better than the other.

Conclusion: The performance of the diagnosis systems must be assessed using the proper methods and criteria. Using different suitable performance measures, it is possible to assess the diagnosis performance of such systems in details.

Key Words: Computer-Aided Diagnosis, Data Mining, Pattern Recognition, Validation

Citation: Mansorian M, Marateb H R. **Performance assessment of computer-aided diagnosis systems: A review of methodologies** J Health Syst Res 2015; 11(3):445-458

Received date: 14.09.2015 **Accept date: 03.11.2015**

1. Medical Physics, Medical Sciences Schools, Tarbiat Modares University, Tehran, Iran
2. Assistant Professor, Biomedical Engineering Department, Faculty of Engineering, University of Isfahan, Iran (Corresponding Author) Email: h.marateb@eng.ui.ac.ir