افزایش صحت طبقه‌بندی بیماران دیابتی از لحاظ محدودیت عملکردی با استفاده از ترکیب خطی و غیرخطی بیومارکرها: روش Ramp AUC

لیلی فرجی گاوگانی [۱]، پروین سربخش [۲]، محمد اصغری جعفرآبادی [۳] مرتضی شمشیر گران [۴]

تاریخ دریافت: ۹۷/۸/۲۷
تاریخ پذیرش: ۹۹/۳/۱

چکیده:

سطح زیر منحنی راک یک معیار مرسوم برای ارزیابی عملکرد طبقه‌بندی بیومارکرها است. در عمل یک بیومارکر قدرت طبقه‌بندی محدودی دارد لذا برای بهبود عملکرد طبقه‌بندی، علاقه‌مند به ترکیب مقادیر مربوط به بیومارکرها به صورت خطی و غیرخطی هستیم در این مطالعه ضمن معرفی انواع توابع زیان، به معرفی روش Ramp AUC و برخی ویژگی‌های آن به عنوان یک مدل آماری مبتنی بر سطح زیر منحنی راک پرداخته می‌شود. این مدل جهت ترکیب بیومارکرها به شکل خطی یا غیرخطی باهدف بهبود عملکرد طبقه‌بندی و مینیمم کردن تابع زیان تجربی بر اساس تابع زیان Ramp AUC ارائه‌شده است. به‌عنوان‌مثال کاربردی، در این مطالعه از داده‌های ۳۷۸ بیمار دیابتی مراجعه‌کننده به مراکز دیابتی اردبیل و تبریز در سال ۱۳۹۴-۱۳۹۳ استفاده‌شده است. جهت طبقه‌بندی بیماران دیابتی از لحاظ وضعیت محدودیت عملکردی بر مبنای بیومارکرهای جمعیت شناختی و بالینی از روش RAUC استفاده گردید. اعتبارسنجی مدل به روش آموزش و آزمایش انجام شد. بر اساس نتایج گروه آزمایش، مقادیر سطح زیر منحنی به‌دست‌آمده برای مدل RAUC با ترکیبات خطی از بیومارکرها در قالب هسته خطی برابر ۰٫۸۱ و با هسته تابع پایه شعاعی برابر ۱٫۰۰ می‌باشد. نتایج بیانگر وجود یک الگوی غیرخطی قوی در داده‌ها می‌باشد به طوری که ترکیبات غیرخطی از بیومارکرها عملکرد طبقه‌بندی بالاتری نسبت به ترکیبات خطی را دارا می‌باشند.

واژه‌های کلیدی: تابع هسته، تابع زیان، سطح زیر منحنی راک، طبقه‌بندی، محدودیت عملکردی.

۱ مقدمه

در عمل یک بیومارکر قدرت طبقه‌بندی محدودی دارد لذا برای بهبود عملکرد طبقه‌بندی، علاقه‌مند به ترکیب مقادیر مربوط به بیومارکرها به صورت خطی و غیرخطی هستیم. سطح زیر منحنی راک [۵] یک معیار مرسوم برای ارزیابی عملکرد طبقه‌بندی بیومارکرها می‌باشد. این معیار بیانگر میزان پیشگویی متغیر وابسته توسط مدل به‌کاربرده شده، می‌باشد. مقدار این معیار مابین ۰٫۵ تا ۱ می‌باشد که مقدار ۰٫۵ نشان‌دهنده تصادفی بودن مدل، مقدار بالای ۰٫۷ بیانگر دقت خوب مدل و مقدار بالای ۰٫۹ بیانگر دقت بالای مدل به‌کاربرده شده می‌باشد [۱۹].

بررسی روابط بین بیومارکرها به صورت یک ترکیب خطی، عملکرد طبقه‌بندی را محدود می‌کنند. یک راه‌حل برای بهبود عملکرد طبقه‌بندی، می‌تواند در نظر گرفتن فضایی از ترکیبات چندجمله‌ای یا اسپلاین (توابع چندجمله‌ای، چند ضابطه‌ای با تعدادی گره‌های به‌هم‌پیوسته) برای روابط بین بیومارکرها و انتخاب بسط مناسبی از آن‌ها به عنوان ترکیب بهینه (خطی یا غیرخطی) از بیومارکرها باشد [۱۰]. مدل‌های آماری زیادی مانند رگرسیون لوژستیک و ماشین بردار پشتیبان [۶] و مدل جمعی تعمیم‌یافته [۷] برای ماکزیمم کردن AUC بر مبنای ترکیباتی از بیومارکرها معرفی‌شده‌اند [۱۱، ۲۳، ۲۶، ۱۷، ۲۸] ولی این مدل‌ها دو محدودیت عمده دارند: اول اینکه اکثر این مدل‌ها به منظور پیدا کردن بهترین ترکیب خطی از بیومارکرها مورداستفاده قرار می‌گیرند که ممکن است در صورت وجود رابطه غیرخطی قوی بین آن‌ها یا در صورت وجود اثر متقابل بین بیومارکرها، کارایی لازم را نداشته باشند. مورد دوم اینکه بسیاری از این مدل‌ها برای پیدا کردن بهترین ترکیبات از

[۱] کارشناس ارشد آمار زیستی، دانشکده پزشکی، مرکز تحقیقات پزشکی مبتنی بر شواهد، دانشگاه علوم پزشکی تبریز، تبریز، ایران
[۲] دکتری آمار زیستی، گروه آمار و اپیدمیولوژی دانشگاه علوم پزشکی تبریز، تبریز، ایران
[۳] دانشیار مرکز تحقیقات آموزش علوم پزشکی، دانشگاه علوم پزشکی تبریز، تبریز، ایران
[۴] دکتری اپیدمیولوژی، گروه آمار و اپیدمیولوژی دانشگاه علوم پزشکی تبریز، تبریز، ایران

[5] Area Under the ROC Curve: AUC
[6] Support Index Vector Machine
[7] Generalized Additive Model

الگوریتم گرادیانت استفاده می‌کنند که تابع زیان مدنظر را به صورت موضعی [8] و نه مطلق بهینه می‌کنند [۱۰] مورد سوم اینکه این روش در مقایسه با روش‌های موجود نسبت به داده‌های پرت حساس نیست و برای تعداد کمی از داده‌ها نیز دارای اعتبار کافی می‌باشد. اخیراً در سال ۲۰۱۶ مدلی به نام RAUC [9] توسط فونگ [10] و همکاران برای مرتفع نمودن ایرادات ذکرشده در مورد مدل‌های مذکور معرفی‌شده است [۱۰]. این مدل با استفاده از تابع زیان RAMP و الگوریتمی غیر از گرادیانت یعنی الگوریتم تفاضل توابع محدب [11] همچنین روش تابع هسته اقدام به ترکیب بهینه بیومارکرها می‌کند به طوری که توان طبقه‌بندی متغیر پاسخ با استفاده از ترکیبات ساخته‌شده نسبت به وجود یک بیومارکر یا ترکیبات ساده آن‌ها بیشتر باشد و با مطالعات شبیه‌سازی نشان داده‌شده است که در مجموع نسبت به مدل‌های مرسوم کارایی بهتری (طبقه‌بندی صحیح‌تری) دارد [۱۰]. RAUC با استفاده از الگوریتم تفاضل توابع محدب فضای ترکیبات بیومارکرها را جستجو می‌کند که این فضا ترکیبات غیرخطی را نیز شامل شود [۱۰]. هدف مدل RAUC کاهش میزان تابع زیان سطح زیر منحنی راک تجربی می‌باشد [۱۰].

تعریف ۱.۱. در نظریه برآورد نقطه‌ای و علم آمار و بهینه‌سازی، تابع هزینه یا تابع زیان تابعی است که میزان زیان در یک رویداد تصادفی را نشان می‌دهد. اغلب مقدار مورد انتظار برای این تابع زیان، ریسک آماری گفته می‌شود که به منظور مقایسه دو یا چند برآوردگر مورداستفاده قرار می‌گیرد. در چنین مقایسه‌هایی، برآوردگر با کمترین مقدار زیان مورد انتظار معمولاً ترجیح داده می‌شود. [۶].

اکثر روش‌های مرسوم مانند رگرسیون لوژستیک، ماشین بردار پشتیبان و مدل جمعی تعمیم‌یافته برای بهینه کردن تابع زیان از الگوریتم گرادیانت استفاده می‌کنند که منجر به بهینه‌سازی مکانی و نه بهینه‌سازی مطلق در کل فضای موردبررسی می‌شوند [۱۰].

رویکرد سطح زیر منحنی راک برای به حداکثر رساندن سطح زیر منحنی راک تجربی، یا معادل آن به حداقل رساندن تابع زیان سطح زیر منحنی راک، به صورت رابطه (۱) تعریف می‌شود:

$$\frac{1}{n_D n_{\bar{D}}}\sum_{i=1}^{n_D}\sum_{j=1}^{n_D}\left[1-\mathrm{I}\left\{\left(\mathbf{x_i}-\mathbf{x_j}\right)^T\right\}\beta>0\right] \equiv \frac{1}{n_D n_{\bar{D}}}\sum_{p=1}^{n_D n_{\bar{D}}}\left\{1-\mathrm{I}\left(\eta_p^{\Delta}>0\right)\right\} \quad (1)$$

$\eta_p^{\Delta}=\left(\mathbf{x_i}-\mathbf{x_j}\right)^T\beta$ و $n_D n_{\bar{D}}$ برابر تعداد موردها و شاهدها در نمونه تحت بررسی و □، □ اندیس مربوط به نمونه موردها و شاهدها می‌باشد. $\mathbf{x_i},\mathbf{x_j}$ بردار مربوط به مقادیر بیومارکرها می‌باشد و β بردار ضرایب و p اندیس مربوط به تمام زوج‌های متشکل از نمونه موردها و شاهدها می‌باشد این تابع زیان تجربی غیر هموار و مشتق ناپذیر است لذا برای مینیمم کردن آن به جای تابع پله‌ای $1-\mathbf{I}(X)$ از تقریب‌های توابع سیگموئید [12] شامل الگوریتم تفاضل توابع محدب نرمال یا لوژستیک ، تابع هینگ [13] یا تابع زیان درجه ۲ استفاده می‌شود [۱۰].

۲ انواع طبقه‌بندی‌ها بر پایه سطح زیر منحنی راک

دو گروه عمده از روش‌ها برای طبقه‌بندی بر پایه سطح زیر منحنی راک وجود دارد. .چون هر مشاهده به‌طور مستقیم در محاسبه سطح زیر منحنی راک نقش دارند، در مجموع شامل □ جمله خواهند بود که □ بیانگر اندازه نمونه می‌باشد و به رویکرد تک شاخصه [14] معروف هستند و تابع زیان استفاده‌شده در خود روش‌ها را بهینه می‌کنند. روش‌های رگرسیون لوژستیک که تابع درستنمایی و مدل ماشین بردار پشتیبان که تابع معیار هینگ را بهینه می‌کنند جزء این گروه می‌باشند [۲۲].

گروه دوم به منظور حداکثر کردن تابع سطح زیر منحنی راک تجربی، به عبارتی حداقل کردن تابع زیان سطح زیر منحنی تجربی، مورداستفاده قرار می‌گیرد. در این تابع داده‌ها به دو گروه کنترل و شاهد تقسیم می‌شوند و بر مبنای مقایسه زوج مشاهدات می‌باشد، و در مجموع شامل n^2 جمله خواهند بود. به این گروه رویکرد دو شاخصه یا رویکرد سطح زیر منحنی راک، گفته می‌شود [۱۰] رویکرد سطح زیر منحنی راک نسبت به داده‌های پرت نیرومندتر از رویکرد تک شاخصه می‌باشد. تابع زیان سطح زیر منحنی راک به صورت رابطه (۲) می‌باشد:

$$\frac{1}{n_D n_{\bar{D}}}\sum_{i=1}^{n_D}\sum_{j=1}^{n_D}\left[1-\mathrm{I}\left\{\left(\mathbf{x_i}-\mathbf{x_j}\right)^T\right\}\beta>0\right] \equiv \frac{1}{n_D n_{\bar{D}}}\sum_{p=1}^{n_D n_{\bar{D}}}\left\{1-\mathrm{I}\left(\eta_p^{\Delta}>0\right)\right\} \quad (2)$$

$\eta_p^{\Delta}=\left(\mathbf{x_i}-\mathbf{x_j}\right)^T\beta$ و $n_D n_{\bar{D}}$ برابر تعداد موردها و شاهدها در نمونه تحت بررسی می‌باشد. زمانی که اندازه داده‌های آموزش، بزرگ باشد و سطح زیر منحنی راک به عنوان معیار عملکرد طبقه‌بندی انتخاب شود، رویکرد

[8] local
[9] Ramp AUC
[10] Fong
[11] Difference of Convex functions Algorithm: DCA
[12] Sigmoid
[13] Hing
[14] single-indexed approach

سطح زیر منحنی راک، رویکردی بهتر از رویکرد تک شاخصه خواهد داشت [۲۲]. روش‌هایی مانند رگرسیون لوژستیک و ماشین بردار پشتیبان نسبت به داده‌های پرت حساس هستند. نیرومندی رویکرد سطح زیر منحنی راک نتیجه بهینه‌سازی تابع معیار استفاده‌شده در ارزیابی عملکرد طبقه‌بندی است که به تنظیم پارامتر خاص در مورد نیرومند بودن، نیاز ندارد و تنها از تابع پله‌ای که حالت رتبه‌بندی دارد استفاده‌شده است. سطح زیر منحنی راک بیشتر از رویکرد تک شاخصه مشکل بهینه‌سازی دارد به‌عنوان‌مثال رگرسیون لوژستیک که دارای رویکرد تک شاخصه می‌باشد تابعی محدب و هموار هست اما رویکرد سطح زیر منحنی راک تجربی این‌چنین نیست و باید طی مراحلی بهینه شود [۱۰].

هم‌چنین دو گروه اصلی از روش‌ها بر پایه رویکرد AUC وجود دارد. اکثر روش‌ها بر پایه رویکرد سطح زیر منحنی راک، روش‌های سطح زیر منحنی راک هموار [15] هستند در این روش‌ها از تابع هموار سیگموئید مانند $\mathbf{S_s(x)}$ به عنوان تقریبی از تابع پله‌ای $\mathbf{1-I}(x>0)$ در تابع زیان سطح زیر منحنی راک استفاده می‌شود. به‌عنوان‌مثال تقریب پله‌ای تابع توزیع چگالی نرمال مانند $\mathbf{S_s(x)}=\varphi\left(-\frac{x}{s}\right)$ [۲۰، ۲۷] و تقریب تابع لوژستیک مانند $\mathbf{S_s(x)}=\frac{1}{1+\exp\left(\frac{x}{s}\right)}$ [۲۹، ۱۳، ۲۱] استفاده می‌شوند. با تغییر دادن تقریب استفاده‌شده، تابع معیار نیز تغییر خواهد کرد. هر دو تقریب نرمال و لوژستیک، پارامتر مقیاس s که بیان‌گر دقت تقریب می‌باشد، کنترل می‌کنند که s کوچک‌تر منجر به‌تقریب بهتر می‌شود اما در این صورت بهینه‌سازی سخت خواهد بود [۸] تقریب‌های هموار دیگری مانند تقریب چندجمله‌ای نیز پیشنهادشده‌اند [۸]. این گروه به دلیل اینکه از الگوریتم تفاضل توابع محدب برای بهینه‌سازی استفاده می‌کنند بیشتر از گروه دوم کاربرد دارند.

دومین گروه از روش‌های مربوط به رویکرد سطح زیر منحنی راک، مدل SVM-AUC [16] می‌باشد [۵، ۳۰، ۱۶] در این روش تابع پله‌ای $\mathbf{1-I(x>0)}$ در تابع زیان سطح زیر منحنی راک به وسیله تابع زیان هینگ که به صورت $\mathbf{(1-x)_+}$ است، جایگزین می‌شود. SVM-AUC تقریب خوبی از تابع زیان سطح زیر منحنی راک نمی‌باشد به خصوص زمانی که در بین داده‌های آموزش، داده‌های پرت وجود داشته باشد، هم‌چنین از الگوریتم گرادیانت برای بهینه‌سازی استفاده می‌کند.

مینیمم RAUC دارای مشکل شناسایی پذیری می‌باشد زیرا به ازای $c>1$ ،$h_s(x)\geq h_s(cx)$ می‌باشد. یک روش پیشنهادی جدید برای حل این مسئله، جایگذاری $s=1$ و جریمه $\|\beta\|^2$ می‌باشد. در نهایت حداقل مقدار

۳ معرفی روش RAUC

RAUC یک روش جدید در پیدا کردن بهترین ترکیب از بیومارکرها مبتنی بر سطح زیر منحنی راک، بر پایه الگوریتمی غیر از گرادیانت یعنی الگوریتم تفاضل توابع محدب و استفاده از روش تابع هسته است که ترکیبات غیرخطی را نیز شامل می‌شود. برای این منظور استفاده از تقریب تابع پله‌ای $\mathbf{1-I}(x)$ در تابع زیان سطح زیر منحنی راک موجود در رابطه (۲)، به صورت تابع شیب‌دار در رابطه (۳) پیشنهادشده است:

$$\mathrm{h_s(u)}=\begin{cases}1 & \text{اگر} \quad \frac{u}{s}\leq -\frac{1}{2}\\ -\frac{u}{s}+\frac{1}{2} & \text{اگر} \quad -\frac{1}{2}\leq \frac{u}{s}\leq \frac{1}{2}\\ 0 & \text{اگر} \quad \frac{u}{s}\geq \frac{1}{2}\end{cases} \qquad (۳)$$

$s>0$ یک پارامتر مقیاس می‌باشد و $\sum_{p=1}^{n_D n_{\bar{D}}}\left(\frac{\mathrm{h_s}(\eta_p^{\Delta})}{n_D n_{\bar{D}}}\right)$ را تابع زیان RAMP AUC می‌نامند.

برای هر مقدار $\eta_p^{\Delta}\neq 0$ مقدار این تابع زیان بین ۰ تا ۱ تغییر می‌کند و در بعضی نقاط که s به سمت ۰ میل کند، مقدار آن ۰ یا ۱ می‌شود. شکل ۱ روش‌های مختلف طبقه‌بندی بر مبنای سطح زیر منحنی راک بر اساس توابع زیان مختلف را نشان می‌دهد که بیانگر تقریب بهتری از تابع پله‌ای سطح زیر منحنی راک در صورت استفاده از روش RAUC می‌باشد.

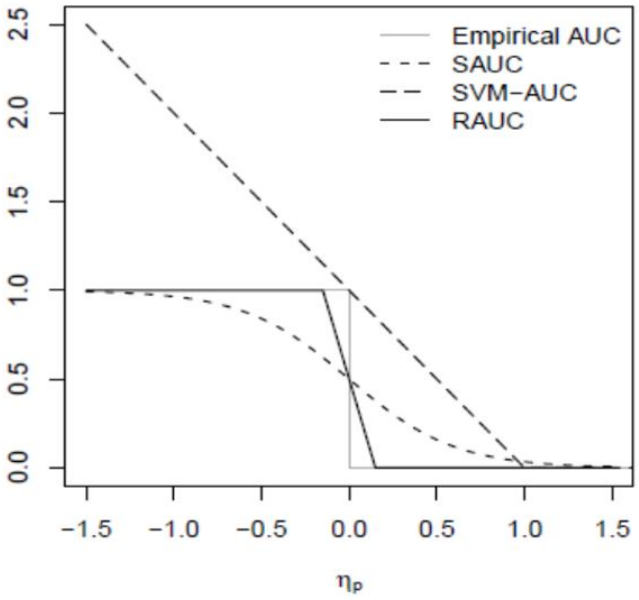
تعریف ۱.۳. در صورتی که برای هر یک از پارامترهای آزاد مدل از روی داده‌های مشاهده‌شده متفاوت، یک مقدار یگانه به دست بیاید آن مدل دارای مسئله شناسایی پذیری [17] خواهد بود.

مسئله شناسایی پذیری در همه روش‌های مبتنی بر رویکرد سطح زیر منحنی راک وجود دارد. این مشکل در مینیمم تابع زیان رویکرد سطح زیر منحنی راک، به ازای مقدار ثابت c به صورت $\mathbf{1-I(x>0)=1-I(cx>0)}$ نمایش داده می‌شود [۱۰].

تاوان تابع زیان RAMP به صورت رابطه (۴) می‌باشد:

$$\Sigma h(\eta_p^{\Delta})+\frac{1}{2}\lambda_n\|\beta\|_2^2 \qquad (۴)$$

[15] Smoothed Area Under the ROC Curve: SAUC
[16] Support Vector Machine-AUC
[17] identifiability
[18] adjust



شکل ۱: نمودار انواع توابع زیان مورداستفاده در سطح زیر منحنی راک [۱۰]

این در مینیمم کردن تابع زیان RAUC مؤثر است، همچنین باعث محدودیت و ثبات مقدار $\|\beta\|_2^2$ می‌شود. λ_n به عنوان یک اهرم برای تعدیل [۱۸] کردن $\|\beta\|^2$ به‌کاربرده می‌شود به طوری که وقتی λ_n کوچک باشد $\|\beta\|^2$ مقدار بزرگی در نظر گرفته می‌شود و نسبت اشباع افزایش می‌یابد بنابراین اگر λ_n به‌اندازه کافی کوچک در نظر گرفته شود به طوری که حداقل ۷۵٪ از $h(\eta_p^{\Delta})$ برابر ۰ یا ۱ باشد عملکرد طبقه‌بندی رضایت‌بخش و مشکل بهینه‌سازی حل خواهد شد [۲۷].

۴ بهینه‌سازی و ترکیب غیرخطی بیومارکرها

تابع زیان RAUC یک تابع غیر محدب و غیر هموار است برای حل مشکل غیر محدب بودن و پیدا کردن بهترین ترکیب، تابع شیب‌دار h به صورت تفاضل دو تابع محدب h_1 و h_2 نوشته می‌شود که در شکل ۲ نشان داده‌شده است و به صورت رابطه (۵) می‌توان نوشت:

$$h(x) = h_1(x) - h_2(x)$$
$$h_1(x) = \left(\frac{1}{2} - x\right)_+, h_1(x) = \left(-\frac{1}{2} - x\right)_+ \quad (۵)$$

این نحوه نوشتن، امکان استفاده از الگوریتم تفاضل توابع محدب را به وجود می‌آورد [۱۸]. اساس الگوریتم تفاضل توابع محدب در این مسئله، استفاده از یک سری مسائل بهینه‌سازی بر حسب توابع محدب برای بهینه کردن یک تابع غیر محدب است و این به وسیله تقریب تکراری h_2 با بسط تیلور مرتبه اول به دست می‌آید و به تقریبی برای h_1 هم نیازی نیست زیرا تفاوت بین یک تابع محدب و یک تابع خطی یک تابع محدب است. شکل ۲ نشان می‌دهد تابع شیب‌دار h به صورت تفاضل دو تابع محدب h_1 و h_2 نوشته می‌شود.

فرایند بهینه‌سازی الگوریتم تفاضل توابع محدب به شرح زیر می‌باشد:

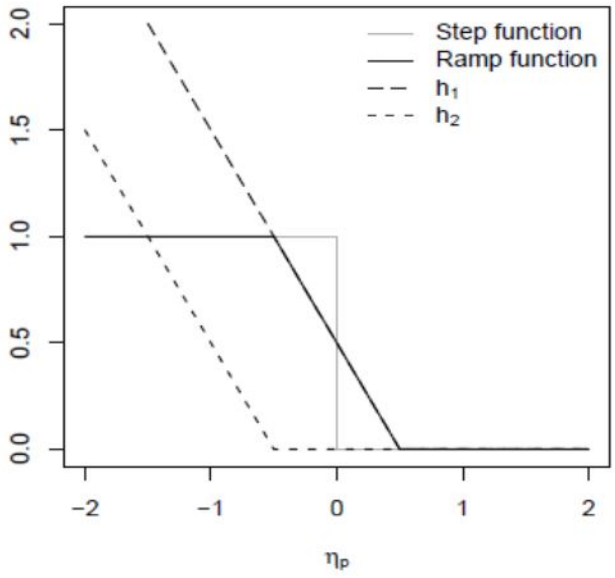
- گام اول: یک مقدار اولیه برای η_p^{Δ} در نظر می‌گیریم و آن را با $\eta_p^{(\Delta,0)}$ نشان می‌دهیم. برای مثال اگر $\hat{\theta}^{rob}$ را به عنوان برآورد اولیه رگرسیون لوژستیک نیرومند [۱۹] به کار ببریم رابطه (۶) را خواهیم داشت:

$$\left(1, \hat{\theta}^{rob}\right)^T \left(x_{ip} - x_{jp}\right)^T \to \eta_p^{\Delta,0} \quad (۶)$$

- گام دوم: حل معادله موجود در رابطه (۷) می‌باشد:

$$\hat{\beta} = \arg\min \sum_{p=1}^{n_D n_{\bar{D}}} \left\{ h_1\left(\eta_p^{\Delta}\right) - h_2\left(\eta_p^{\Delta}, \eta_p^{\Delta,0}\right) + \frac{1}{2}\lambda_n \|\beta\|_2^2 \right\} \quad (۷)$$

[19] Robust Logistic Regression



شکل ۲: تفاضل دو تابع محدب به عنوان برآوردی از تابع RAUC [۱۰]

به طوری که به ازای تمام مقادیر x مشتق مرتبه اول $h_2(x)$ به جز در نقطه $x=-\frac{1}{2}$ که مقدار آن برابر صفر می‌باشد، وجود داشته باشد، $\hat{h}_2\left(\eta_p^{\Delta},\eta_p^{\Delta,0}\right)=h'_2\left(\eta_p^{\Delta,0}\right)\eta_p^{\Delta}+h_2\left(\eta_p^{\Delta,0}\right)$ و $h'_2(x)=-I\left(x<-\frac{1}{2}\right)$ می‌باشد.

- گام سوم: اگر مجموعه $\hat{\beta}\left(x_{ip}-x_{jp}\right)^T$ به $\eta_p^{\Delta,0}$ میل کند آنگاه به گام دوم بازمی‌گردیم و این مراحل را تکرار می‌کنیم تا زمانی که تغییرات ایجادشده در پنالتی تابع زیان RAUC کمتر از یک مقدار ثابت از قبل تعیین‌شده باشد.

در گام دوم از الگوریتم فوق، یک مشکل بهینه‌سازی محدب حل می‌شود. به‌عنوان‌مثال h_1 و $\hat{h}_2$ توابعی غیر هموار هستند که با استفاده از روش استاندارد و معرفی متغیرهای ε_p به جای h_1، به یک مسئله بهینه‌سازی هموار تبدیل می‌شود.

در مواردی که $\varepsilon_p\geq 0$ و $\varepsilon_p\geq\frac{1}{2}-\eta_p^{\Delta}$ منجر به رابطه (۸) خواهد شد:

$$\min_{\varepsilon,\beta}\sum_{p=1}^{n_D n_{\bar{D}}}\left\{\varepsilon_p-h_2\left(\eta_p^{\Delta,0}\right)\right\}+\frac{1}{2}\lambda_n\|\beta\|_2^2 \qquad (۸)$$

و این زمانی روی خواهد داد که محدودیت‌های مربوط به ε و β را در نظر نگیریم. چون در حل مشکل بهینه‌سازی، کار کردن با محدودیت‌ها سخت می‌باشد [۴] هینگ به دنبال پیدا کردن حل مشکل دوگانه (β,ε) از طریق روش لاگرانژ است. روش لاگرانژ اولیه در رابطه (۹) نشان داده‌شده است:

$$\begin{aligned}L_p=\sum_{p=1}^{n_D n_{\bar{D}}}&\left\{\varepsilon_p-h_2\left(\eta_p^{\Delta,0}\right)\eta_p^{\Delta}\right\}+\frac{1}{2}\lambda_n\|\beta\|_2^2\\&-\Sigma_p\alpha_p\left(\varepsilon_p+\eta_p^{\Delta}-\frac{1}{2}\right)-\Sigma_p\gamma_p\varepsilon_p\end{aligned} \qquad (۹)$$

که α و γ، لاگرانژ چندگانه غیر منفی می‌باشند و مربوط به دو مجموعه از محدودیت‌های موجود در مسئله بهینه‌سازی می‌باشند. اگر مشتق مرتبه اول لاگرانژ را به دست آورده و برابر ۰ قرار دهیم، مقادیر قابل‌قبول برای متغیرهای فضای اولیه ε و β را به صورت رابطه (۱۰) به دست خواهیم آورد:

$$\beta=\frac{1}{\lambda_n}\sum_{p=1}^{n_D n_{\bar{D}}}\left\{h_2\left(\eta_p^{\Delta,0}\right)+\alpha_p\right\}x_p^{\Delta}\quad,1=\alpha_p+\gamma_p \qquad (۱۰)$$

همچنین اگر مقادیر به‌دست‌آمده را در L_p جایگذاری کنیم شرایط سخت مربوط به ε_p از بین خواهد رفت. در بعضی مواقع در حل جبری ممکن است با دو برابر مشکل موجود مواجه شویم، مثلاً در مواردی که $0\leq\alpha_p\leq 1$ باشد رابطه (۱۱) را خواهیم داشت:

$$\min_{\alpha}\alpha^T Q\alpha-b^T\alpha \qquad (۱۱)$$

در عبارت بالا Q ماتریس توان دوم به صورت $[p,p']^{th}$ با اجزاء $\left\langle x_p^{\Delta},x_{p'}^{\Delta}\right\rangle/\lambda_n$ و $b=1-2\times Q\times h'_2\left(\eta^{\Delta,0}\right)$ می‌باشد. برای یافتن بهترین ترکیب و پیدا کردن طبقه مربوطه به یک مشاهده‌ی جدید [۱۰] از رابطه (۱۲) استفاده می‌شود:

$$\eta_{new}=\sum_{p=1}^{n_D n_{\bar{D}}}\left\{h_2\left(\eta_p^{\Delta,0}\right)+\alpha_p\right\}\left\langle x_{new},x_p^{\Delta}\right\rangle \qquad (۱۲)$$

[20] non-support pairs

$\eta^{(\Delta,\circ)}$ و α از آخرین تکرار الگوریتم تفاضل توابع محدب به دست می‌آید، طبق اصطلاحات ماشین بردار پشتیبان [۷] دو عبارت x_p^{Δ} و $\mathrm{h}_2\left(\eta_{\mathrm{p}}^{\Delta,0}\right)+\alpha_{\mathrm{p}}=0$ را جفت غیر پشتیبان [۲۰] نامیده می‌شود زیرا که در امتیازدهی شرکت نمی‌کنند.

۵ استفاده از فضای ویژگی برای جداسازی ترکیبات غیرخطی

گاهی ممکن است بهترین طبقه‌بندی بین ترکیبات خطی از متغیرهای بیومارکر ورودی صورت نگیرد و نتوانیم داده‌ها را به صورت خطی جدا کنیم در این صورت آن‌ها را به فضایی با ابعاد بیشتر نگاشت می‌دهیم تا بتوان آن‌ها را در این فضا به صورت خطی جدا کرد بدین منظور برای به دست آوردن بهترین ترکیب از فضای ویژگی استفاده می‌شود. فضای ویژگی برای مورد i را با φ_i نشان می‌دهند. در این صورت مشکل دوگانه موجود در رابطه (۱۲) با استفاده از Q که یک ماتریس توان دوم به صورت $[\mathrm{p},\mathrm{p}']^{\mathrm{th}}$ با اجزاء $\left\langle \mathrm{x}_{\mathrm{p}}^{\Delta},\mathrm{x}_{\mathrm{p}}^{\Delta}\right\rangle/\lambda_{\mathrm{n}}$ می‌باشد می‌توان حل کرد که $\left\langle \varphi_{\mathrm{p}'}^{\Delta},\varphi_{\mathrm{p}'}^{\Delta}\right\rangle$ را به صورت رابطه (۱۳) نوشت:

$$\begin{aligned}\left\langle \varphi_{\mathrm{p}}^{\Delta},\varphi_{\mathrm{p}'}^{\Delta}\right\rangle &= \left\langle \varphi_{\mathrm{ip}},\varphi_{\mathrm{ip}'}\right\rangle+\left\langle \varphi_{\mathrm{jp}},\varphi_{\mathrm{jp}'}\right\rangle-\left\langle \varphi_{\mathrm{jp}},\varphi_{\mathrm{jp}'}\right\rangle \\ &\quad -\left\langle \varphi_{\mathrm{ip}},\varphi_{\mathrm{jp}'}\right\rangle\end{aligned} \qquad (۱۳)$$

مشکل دوگانگی موجود در مرحله سوم را با معرفی $\hat{\eta}^{\Delta}$ در رابطه (۱۴) می‌توان حل کرد، به طوری‌که نیازی به برگشتن به مرحله دوم نباشد:

$$\hat{\eta}^{\Delta}=\left(\varphi_{\mathrm{i}}^{\Delta},\ldots,\varphi_{\mathrm{p}}^{\Delta}\right)\hat{\beta}=\mathrm{Q}\times\left\{\mathrm{h}'_2\left(\eta^{\Delta,0}\right)+\hat{\alpha}\right\} \qquad (۱۴)$$

φ_i ممکن است بی نهایت-بعدی باشد و $\hat{\eta}^{\Delta}$ معمولاً نماینده یک-بعدی می‌باشد. بنابراین نمره یک مشاهده جدید را می‌توان بر اساس رابطه (۱۵) به دست آورد:

$$\begin{aligned}\eta^{new} &= \sum_{p=1}^{n_D n_{\bar{D}}}\left\{h'_2\left(\eta_p^{\Delta,\circ}\right)+\alpha_p\right\}\left\langle \varphi\left(x_{new}\right),\varphi_p^{\Delta}\right\rangle \\ &= \sum_{p=1}^{n_D n_{\bar{D}}}\left\{h'_2\left(\eta_p^{\Delta,\circ}\right)+\alpha_p\right\} \\ &\quad\times\left\{\left\langle \varphi\left(x_{new}\right),\varphi_{ip}\right\rangle-\left\langle \varphi\left(x_{new}\right),\varphi_{jp}\right\rangle\right\}\end{aligned} \qquad (۱۵)$$

$\eta_p^{\Delta,\circ}$ و α از آخرین تکرار الگوریتم تفاضل توابع محدب به دست می‌آیند. این رویکرد از نگاشت از بردار ورودی در یک فضای ویژگی بدون نیاز به مشخص کردن دقیق مقدار نگاشت شده در ادبیات ماشین بردار پشتیبان، تکنیک هسته‌ای [۲۱] نامیده می‌شود [۱، ۱۲].

وقتی داده‌ها در فضای ورودی به صورت خطی تفکیک‌پذیر نباشند، یعنی کلاس داده‌ها هم‌پوشانی داشته باشند در این صورت با استفاده از توابع غیرخطی تحت عنوان فضای ویژگی، داده‌ها به فضایی با ابعاد بیشتر انتقال داده می‌شود تا در این فضا به صورت خطی از هم جدا شوند و در فضای جدید با جایگزینی $\mathrm{x_k}$ با $\varphi(\mathrm{x_k})$ مرز بهینه را می‌توان محاسبه نمود. فضای ویژگی دارای ابعاد بزرگ می‌باشد به خاطر همین دارای محاسبات زیاد و پرهزینه خواهد بود برای غلبه بر این مشکل از تابع هسته استفاده می‌شود که مبتنی بر ضرب داخلی می‌باشند [۱۵، ۳]. هسته‌های متفاوتی وجود دارد که انتخاب هسته مناسب بستگی به ماهیت داده‌ها دارد. در حالت کلی دو نوع هسته داریم: هسته خطی و هسته غیرخطی که شامل سیگموئید چندجمله‌ای و تابع شعاعی پایه [۲۲] می‌باشد که متداول‌ترین آن‌ها تابع شعاعی پایه می‌باشد [۹] انواع توابع هسته همراه با الگوریتم‌های مربوطه در جدول ۱ نشان داده‌شده است:

جدول ۱. انواع تابع هسته

الگوریتم	تابع هسته
$\mathrm{K}\left(\mathrm{x_i},\mathrm{x_j}\right)=\left\langle \mathrm{x_i},\mathrm{x_j}\right\rangle$	خطی
$\mathrm{K}\left(\mathrm{x_i},\mathrm{x_j}\right)=\left(\mathrm{r}+\gamma\left\langle \mathrm{x_i},\mathrm{x_j}\right\rangle\right)^{\mathrm{d}}$	چندجمله‌ای
$\mathrm{K}\left(\mathrm{x_i},\mathrm{x_j}\right)=\exp\left(-\gamma\left\|\mathrm{x_i}-\mathrm{x_j}\right\|^2\right)$	تابع شعاعی پایه
$\mathrm{K}\left(\mathrm{x_i},\mathrm{x_j}\right)=\tanh\left(\gamma\mathrm{X^T{}_i X_j}+\mathrm{r}\right)$	سیگموئید

۶ مثال کاربردی

مثال ۱.۶. مثال کاربردی مورداستفاده در این مطالعه ترکیبی از دو مطالعه مقطعی- تحلیلی صورت گرفته در شمال غرب ایران در سال ۱۳۹۴-۱۳۹۳ می‌باشد جزییات بیشتر در مورد این مطالعات در چندین مطالعه ارائه‌شده‌اند [۲، ۲۵، ۲۴]. در مطالعه حاضر با ترکیب داده‌های دو مطالعه و حذف موارد دارای داده گمشده برای متغیرهای موردنیاز، تعداد ۳۷۸ نفر با داده‌های کامل برای فرایند مدل‌سازی باقی ماندند. سن بیماران دیابتی بین ۲۵ و ۷۰ سال با میانگین ۵۶/۱۶ سال (با انحراف معیار ۸/۷۹ سال) و میانگین طول مدت ابتلا ۸/۹۱ سال (با انحراف معیار ۶/۱۱ سال) بود. از بین ۳۷۸ بیمار دیابتی، ۲۵۵ نفر (۶۷/۴۶ درصد) دارای محدودیت عملکردی بودند. از ۲۴۹ زن تحت مطالعه نیز ۷۲/۳ درصد دارای محدودیت عملکردی بودند درحالی‌که این نسبت در مردان ۵۸/۱ درصد بود و بین زنان و مردان تفاوت معنی‌داری از لحاظ احتمال ابتلا به محدودیت عملکردی مشاهده‌شده است ($p-value=\circ/\circ\circ5$).

مدل $RAUC$ برای بررسی نحوه ارتباط بین بیومارکرهای جنسیت، سن، طول مدت ابتلا، نمایه توده بدن، HDL، کلسترول، تری گلیسیرید، قند خون ناشتا، فشارخون سیستولیک و دیاستولیک و هموگلوبین $A1C$ به عنوان عوامل خطر مرتبط و محدودیت عملکردی به عنوان پیامد به کار گرفته‌شده است. برای تحلیل بر اساس مدل $RAUC$ از هسته تابع پایه شعاعی به عنوان

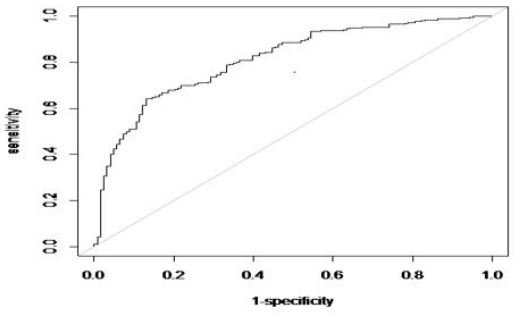
[21] Kernel trick

[22] Radial Basis Function) RBF

پرکاربردترین و مهم‌ترین هسته غیرخطی به منظور بررسی ترکیبات غیرخطی و تابع هسته خطی به منظور بررسی ترکیبات خطی استفاده‌شده است. برای مقایسه قدرت تشخیص و افتراق این مدل با دو تابع هسته متفاوت استفاده‌شده، از سطح زیر منحنی راک استفاده‌شده است. تحلیل‌های لازم برای برازش مدل *RAUC* از نرم‌افزار *R 3.3.1* با بسته نرم‌افزاری *aucm* استفاده گردیده است. به منظور اعتبارسنجی نتایج از روش آزمون و آزمایش مورداستفاده قرارگرفته است. داده‌ها به‌طور تصادفی به دو گروه، داده آموزشی (۷۰ درصد داده‌ها) و داده آزمایش (۳۰ درصد داده‌ها) تقسیم شدند. مقدار احتمال ابتلا به محدودیت عملکردی پس از برازش مدل *RAUC* بر روی داده‌های گروه آزمایش، برای داده‌های گروه آموزش محاسبه گردید و با استفاده از مقدار احتمال به‌دست‌آمده، شاخص سطح زیر منحنی راک برای تعیین صحت طبقه‌بندی بیماران دیابتی مبتلا به محدودیت عملکردی محاسبه گردید. مدل *RAUC* با استفاده از تابع هسته خطی و تابع شعاع پایه بر روی داده‌های آزمایش برازش داده‌شده است. مقادیر سطح زیر منحنی به‌دست‌آمده از نتایج حاصل از داده‌های آزمایش، برای مدل *RAUC* با تابع هسته خطی برابر ۰/۸۱ به‌دست‌آمده که همین مقدار بر اساس هسته تابع شعاعی برابر ۱/۰۰ می‌باشد لذا با توجه به نتایج به‌دست‌آمده، هسته تابع پایه شعاعی بیش‌ترین مقدار سطح زیر منحنی راک را نسبت به هسته خطی دارا می‌باشد.

نمودارهای مربوط به سطح زیر منحنی راک برای مدل *RAUC* با هسته خطی و هسته غیرخطی تابع شعاعی پایه در شکل‌های ۳ و ۴ نشان داده‌شده است.

نتایج حاصل از این پژوهش نشان‌دهنده این موضوع می‌باشد که ترکیبات غیرخطی عملکرد طبقه‌بندی یا پیشگویی بالاتری نسبت به ترکیبات خطی را دارا می‌باشند، این بیان‌گر وجود یک الگوی غیرخطی در داده‌های این مطالعه می‌باشد.



شکل ۳:
نمودار سطح زیر منحنی راک برای مدل RAUC با هسته خطی

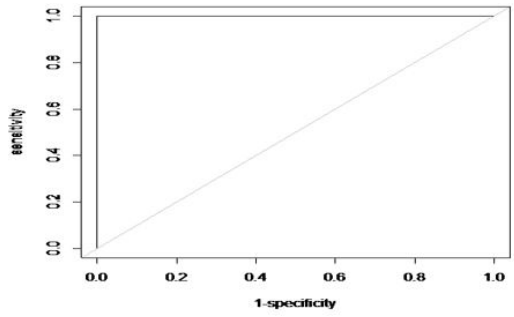
۷ بحث و نتیجه‌گیری

ارزیابی توانایی طبقه‌بندی در تشخیص یا پیشگویی صحیح طبقه ضروری است. معیارهای مختلفی برای ارزیابی عملکرد طبقه‌بندی وجود دارد ولی در طبقه‌بندی‌های دو طبقه‌ای بیشتر سطح زیر منحنی راک توصیه‌شده است. علاقه‌مندی به استفاده از سطح زیر منحنی راک برای ارزیابی عملکرد طبقه‌بندی منجر به طراحی الگوریتم‌هایی شده است که هدف آن‌ها به حداکثر رساندن سطح زیر منحنی راک می‌باشد [۱۴].

سطح زیر منحنی راک اغلب برای اندازه‌گیری عملکرد برآوردگر در مسائل مربوط به طبقه‌بندی دو سطحی استفاده می‌شود. این معیار بیانگر این موضوع می‌باشد که مدل به‌کاربرده شده، به چه اندازه متغیر وابسته را می‌تواند به خوبی پیشگویی کند.

در پژوهش حاضر میزان درستی طبقه‌بندی افراد دیابتی از لحاظ محدودیت عملکردی، بر مبنای بیومارکرهایی مانند جنسیت، سن، طول مدت ابتلا، و بیومارکرهای کیفیت مراقبت شامل فشارخون، قند خون، نمایه توده بدن، کلسترول و هموگلوبین A1C تری گلیسیرید موردبررسی قرار گرفت و با توجه به نتایج حاصل از این پژوهش و مقدار سطح زیر منحنی راک به‌دست‌آمده از مدل RAUC با تابع هسته غیرخطی پایه شعاعی که بیان‌گر یک الگوی غیرخطی بین داده‌ها می‌باشد می‌توان بیان کرد که ترکیبات غیرخطی عملکرد طبقه‌بندی یا پیشگویی بالاتری نسبت به ترکیبات خطی را دارا می‌باشند.

با توجه به اهمیت طبقه‌بندی صحیح مطالعات پزشکی، تشخیص و تعیین ارتباط پیچیده و غیرخطی بین بیومارکرها، بر اساس روش‌هایی که برای یافتن چنین ترکیبات بهینه استفاده می‌گردد مدل RAUC می‌تواند مفید باشد.



شکل ۴:
نمودار سطح زیر منحنی راک مربوط به مدل RAUC با تابع هسته غیرخطی تابع شعاعی پایه

مراجع

[1] Aizerman, M. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control,* **25**, 821–37.

[2] Ataei, J., Shamshirgaran, S., Iranparvar Alamdari, M. and Safaeian, A. (2015). Evaluation of Diabetes Quality of Care Based on a Care Scoring System among People Referring to Diabetes Clinic in Ardabil, 2014. *Journal of Ardabil University of Medical Sciences,* **15**(2), 207–19.

[3] Bellotti, T. and Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications,* **36**(2), 3302–8.

[4] Boyd, S. and Vandenberghe, L.(2004). *Convex Optimization,* Cambridge University Press.

[5] Brefeld, U. and Scheffer, T. (2005). AUC maximizing support vector learning. *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning,* Citeseer semanticscholar.

[6] Bühlmann, P. and Van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. *Springer Heidelberg Dordrecht London*. New York.

[7] Burges, C.J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery,* **2**(2), 121–67.

[8] Calders, T. and Jaroszewicz, S. (2007). Efficient AUC optimization for classification, *Knowledge Discovery in Databases*: PKDD 2007, 42–53.

[9] Daugherty, C., Dewhurst, W. and Spertus, J. (1998). Health related quality of life outcomes in clinical research. *American Journal of Emidemiology,* **51**(7), 569–75.

[10] Fong, Y., Yin, S. and Huang, Y. (2016). Combining biomarkers linearly and nonlinearly for classification using the area under the ROC curve. *Statistics in medicine,* **35**(21), 3792–809.

[11] Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm, *in Proceedings of the Thirteenth International Conference on Machine Learning*, 148–156, Morgan Kaufmann.

[12] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Second Edition, Springer Series in Statistics, Springer. New York.

[13] Herschtal, A. and Raskutti, B. (2004). Optimising area under the ROC curve using gradient descent, *in Proceedings of the Twenty-First International Conference on Machine Learning*, 49, ACM.

[14] Holst, A. (2008). Efficient AUC maximization with regularized least-squares. *Tenth Scandinavian Conference on Artificial Intelligence,* SCAI 2008, IOS Press. Amsterdam.

[15] Huang, C.L., Chen, M.C. and Wang, C.J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications,* **33**(4), 847–56.

[16] Joachims, T. (1998). Making large-scale SVM learning practical. *Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund*.

[17] Komori, O. (2011). A boosting method for maximization of the area under the ROC curve. *Annals of the Institute of Statistical Mathematics,* **63**(5), 961–79.

[18] Le Thi Hoai, A. and Tao, P.D. (1997). Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. *Journal of global optimization,* **11**(3), 253–85.

[19] Linkie, M., Smith, R.J. and Leader-Williams, N. (2004). Mapping and predicting deforestation patterns in the lowlands of Sumatra. *Biodiversity & Conservation,* **13**(10), 1809–18.

[20] Lloyd, C.J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association,* **93**(444), 1356–64.

[21] Ma, S. and Huang, J. (2007) Combining multiple markers for classification using ROC. *Biometrics,* **63**(3), 751–7.

[22] Pepe, M.S., Cai, T. and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics,* **62**(1), 221–9.

[23] Pepe, M.S. and Thompson, M.L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics,* **1**(2), 123–40.

[24] Shamshirgaran, S.M., Ataei, A., Malek, A., Iranparvar-Alamdari, M. and Aminisani, N. (2017). Quality of sleep and its determinants among people with type 2 diabetes mellitus in Northwest of Iran. *World J Diabetes,* **8**(7), 358–364.

[25] Shamshirgaran, S.M., Mamaghanian, A., Aliasgarzadeh, A., Aiminisani, N., Iranparvar-Alamdari, M. and Ataei, J. (2017). Age differences in diabetes-related complications and glycemic control. Under press. *BMC Endocrine Disorders*, **17**, 25.

[26] Tutz, G. and Binder, H. (2006). Generalized Additive Modeling with Implicit Variable Selection by Likelihood-Based Boosting. *Biometrics,* **62**(4), 961–71.

[27] Vexler, A., Liu, A., Schisterman, E.F. and Wu, C. (2006). Note on distribution-free estimation of maximum linear separation of two multivariate distributions. *Nonparametric Statistics,* **18**(2), 145–58.

[28] Xu-hui, W., Ping, S., Li, C. and Ye, W. (2009). A ROC curve method for performance evaluation of support vector machine with optimization strategy. *Computer Science-Technology and Applications*, 2009 IFCSTA'09 International Forum on; IEEE.

[29] Yan, L., Dodier, R.H., Mozer, M. and Wolniewicz, R.H. (2003). Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. *Proceedings of the 20th International Conference on Machine Learning (ICML-03*.

[30] Yu, L. and Yao, X. (2013). A total least squares proximal support vector classifier for credit risk evaluation. *Soft Computing,* **17**(4), 643–50.