تعیین حجم نمونه در رگرسیون لجستیک

مهدی یوسفی نژاد عطاری [۱]، سعید کلاهی رنجی [۲]، ویدا کرباسی [۳]

چکیده:

مسئله تخمین حجم نمونه در کاربردهای پزشکی به خصوص در موارد آزمایشات گران‌قیمت نشانگرهای زیستی دارای اهمیت است. این مقاله به توصیف مسئله تحلیل رگرسیون لجستیک با الگوریتم های تخمین حجم نمونه که عبارتند از روش‌های آماری تک متغیری، رگرسیون لجستیک، تقاطع اعتبار و استنباط بیز می‌پردازد. نویسندگان با پارامترهای مدل رگرسیونی به عنوان متغیر چند متغیری با هدف تخمین حجم نمونه با استفاده از فاصله بین توابع توزیع پارامتر در مجموعه‌های داده تقاطع اعتبار رفتار می‌کنند. در اینجا نویسندگان کمکی جدید برای داده‌کاوی و آموزش آماری با حمایت ریاضیات کاربردی ارائه می‌دهند.

واژه‌های کلیدی: رگرسیون لجستیک، حجم نمونه، انتخاب ویژگی، استنباط بیز، واگرایی کالبک لایبرر.

۱ مقدمه

این مقاله به تحلیل رگرسیون لجستیک به کار رفته در طبقه بندی مشکلات در مسائل پزشکی اختصاص داده شده است که گروهی از بیماران به عنوان گروه نمونه مورد تحقیق و بررسی قرار گرفته‌اند [۱]. هر مریض با دسته‌ای از ویژگی‌ها که نشانگر زیستی نامیده می‌شود توصیف شده و افراد به دو کلاس طبقه بندی شده‌اند.

به دلیل هزینه بالای آزمایشات، تعداد بیماران در نمونه مورد مطالعه در این مقاله تقریباً کوچک است. دو کلاس به ترتیب شامل ۱٤ و ۱۷ مریض می‌باشد. در این مورد سازگاری بیش از حد مدل طبقه بندی غیر قابل اجتناب است. این منجر به مسئله تخمین حجم نمونه می شود [۲، ۳]. تعیین حجم نمونه به این دلیل حائز اهمیت است که در مسائلی نظیر طبقه‌بندی مشکلات در مسائل پزشکی که در این مقاله مورد بحث قرار گرفته، هزینه آزمایش یک مریض بسیار گران تمام می‌شود بنابراین باید تخمینی دقیق در مورد حجم نمونه صورت گیرد.

مطالب مورد بحث در این مقاله عبارتند از: توصیفی مختصر از رگرسیون لجستیک، در بخش ۲ این مقاله نمایش تابع خصوصیت. متغیر هدف بر اساس یک توزیع برنولی فرض شده. پارامترهای مدل رگرسیونی تخمین زده شده‌اند. نمونه مورد مطالعه شامل ۳۱ مریض با اختلال در سیستم قلبی- عروقی است. کارشناسان ۲۰ ویژگی که نمونه را توصیف می‌کند را نام گذاری کرده‌اند. در بخش ۳ مسئله انتخاب ویژگی بحث شده و در بخش ٤ روش‌های تعیین کمترین حجم نمونه مورد بحث قرار گرفته‌اند [٤، ٥]. با این مجموعه از ویژگی‌ها مدل بسیار پیچیده می‌باشد. به همین دلیل، قبل از برآورد حجم نمونه، ما مجموعه‌ای از ویژگی‌های که اندازه کوچکتری از آن است به طبقه‌بندی موثر بیماران پرداخته شد. در رگرسیون لجستیک، ویژگی‌ها با استفاده از روش رگرسیون گام به گام انتخاب شده است [٦، ۷]. در محاسبات آزمایشی ما یک جستجوی جامع پیاده سازی شده است. این باعث می‌شود که هر ترکیب ممکن از ویژگی‌ها در نظر گرفته شده است [۸، ۹، ۱۰].

در سال ۲۰۱۵ باش در تحقیقی به بررسی انواع روش‌های تعیین حجم نمونه توسط یک مطالعه شبیه‌سازی پرداخت که به امتیاز بندی روش‌های مختلف از جمله رگرسیون لجستیک پرداخت [۱٤]. فرانک و هارل فصلی از کتاب خود با عنوان استراتژی‌های مدل رگرسیون را به رگرسیون لجستیک دودویی (باینری) اختصاص داده‌اند [۱۵].

[۱] عضو هیات علمی دانشگاه آزاد اسلامی واحد بناب، گروه مهندسی صنایع، دانشگاه آزاد اسلامی، بناب، ایران

[۲] کارشناس ارشد مهندسی صنایع، باشگاه پژوهشگران جوان و نخبگان، واحد ایلخچی، دانشگاه آزاد اسلامی، ایلخچی، ایران

[۳] دانشجوی کارشناسی ارشد مهندسی دانشگاه آزاد اسلامی واحد بناب، گروه مهندسی صنایع، دانشگاه آزاد اسلامی، بناب، ایران

۲ شرح مسئله رده‌بندی

مجموعه نمونه به صورت $D = \{(x_i, y_i) : i = 1, \dots, m\}$ با m موضوع (مریض) در نظر گرفته می‌شود و هر مریض با n ویژگی توصیف می‌شود. $x_i \in \mathbb{R}^n$ و مربوط به یکی از دو کلاس است و $y_i \in \{0, 1\}$. هم‌چنین فرض شده y_i دارای توزیع برنولی با تابع چگالی احتمالی به صورت رابطه (۱) است.

$$p(y|B) = \prod_{i=1}^{m} \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \qquad (۱)$$

احتمال θ_i به صورت رابطه (۲) تعریف شده.

$$\theta_i = f(x_i^T B) = \frac{1}{1 + \exp(-x_i^T B)} \qquad (۲)$$

از روش حداکثر درست نمایی، تابع خطا برای رابطه (۱) به صورت رابطه (۳) به دست می‌آید.

$$E(B) = -\ln p(y|B) = -\sum_{i=1}^{m} (y_i \ln \theta_i + (1 - y_i) \ln(1 - \theta_i)) \qquad (۳)$$

برآورد حداکثر درست‌نمایی (MLE[٤]) روشی است برای برآورد کردن پارامترهای یک مدل آماری. وقتی بر مجموعه‌ای از داده‌ها عملیات انجام می‌شود یک مدل آماری به دست می‌آید آنگاه درست‌نمایی بیشینه می‌تواند تخمینی از پارامترهای مدل ارائه دهد. در حالت کلی روش MLE در مورد یک مجموعه مشخص از داده‌ها عبارتست از نسبت دادن مقادیری به پارامترهای مدل که در نتیجه آن توزیعی تولید شود که بیشترین احتمال را به داده‌های مشاهده شده نسبت دهد (یعنی مقادیری از پارامتر که تابع درست‌نمایی را بیشینه کند). MLE یک ساز و کار مشخص را برای تخمین ارائه می‌دهد که در مورد توزیع نرمال و بسیاری توزیع‌های دیگر عمل می‌کند.

با حل مسئله بهینه سازی رابطه (٤) الگوریتم رده بندی به صورت رابطه (۵) به دست می‌آید.

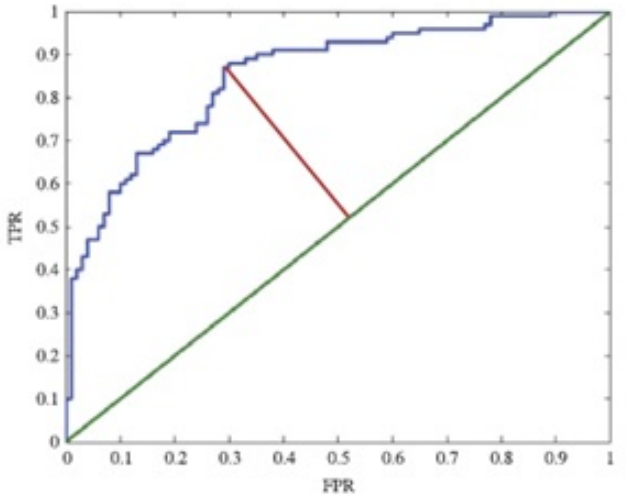
$$\hat{B} = \arg\min E(B) \qquad (٤)$$

$$a(x, c_0) = (f(x, B) - c_0) \qquad (۵)$$

تابع خصوصیت رده‌بندی (AUC) یا همان ناحیه زیر منحنی ROC با معرفی مقادیر $\mathrm{TRP}(\varepsilon)$ یا نرخ مثبت حقیقی و $\mathrm{FPR}(\varepsilon)$ یا نرخ مثبت ساختگی و با به کارگیری تابع شاخص

$$[y = 1] = \begin{cases} 1, & \text{if } y = 1 \\ 0, & \text{if } y \neq 1 \end{cases}$$

بهترین رده‌بندی را، به وسیله بیشترین مقدار از مقادیر AUC مشخص می‌کند.



شکل۱. حجم نمونه m^*، با روش فاصله اطمینان و روش رگرسیون لجستیک

۳ مسئله انتخاب ویژگی

A زیرمجموعه‌ای از فهرست ویژگی‌ها $\{1, \dots, n\}$ و $A \subseteq j \subset$ و $\hat{A}$ زیرمجموعه بهینه از فهرست می‌باشد. X_A ماتریسی است که ترکیبی از ستون‌های ماتریس X با فهرست‌هایی در A می‌باشد و B_A بردار مربوط به پارامترها است. مسئله انتخاب ویژگی مقدار بیشینه‌ای است که از رابطه (٦) به دست می‌آید.

$$\hat{A} = \arg\max \mathrm{AUC}(A), \quad \text{subject to } |A| = \text{const} \qquad (٦)$$

٤ تخمین حجم نمونه

داده مورد بررسی بیماران دو کلاس را توصیف می‌کند: ۱- کسانی که قبلاً مورد حمله قلبی قرار گرفته‌اند ۲- بیمارانی که ممکن است در آینده آن را تجربه کنند. از غلظت پروتئین در سلول‌های خونی به عنوان ۲۰ ویژگی استفاده شده است. در این بخش ٤ روش

[٤] maximum likelihood estimation

برای تخمین حجم نمونه معرفی می‌شوند. که آخرین روش جزء متدهای نوین می‌باشد.

۱.٤ روش قابلیت اطمینان

$D = \{(x_i, y_i) : i \in l = \{1, \ldots, m\}\}$ که هر متغیر پاسخ y_i وابسته به متغیر مستقل منحصر به فرد x_i است که $x_i \sim N(\mu, \sigma^2)$ و Δ تفاضل میانگین و مقدار مورد انتظار شناخته شده μ از متغیر تصادفی x_i است. با معلوم بودن σ^2 متغیر توزیع نرمال استاندارد به صورت رابطه (۷) به دست می‌آید.

$$Z = \frac{\bar{x} - \mu}{\sigma}\sqrt{m} = \frac{\Delta}{\sigma}\sqrt{m} \sim N(0,1) \qquad (۷)$$

در نهایت کمترین حجم نمونه یا m^* با سطح معنی داری α به صورت رابطه (۸) به دست می‌آید.

$$m^* = \left(\frac{Z_{\alpha/2}\sigma}{\Delta}\right)^2 \qquad (۸)$$

این روش فقط به دست‌یابی به یک تخمین حدسی از m^* کمک می‌کند و دلیل آن این است که هیچ یک از مقادیر μ و σ^2 معلوم نیستند. x_i به صورت ترکیب توزیع‌ها به صورت رابطه (۹) توزیع شده.

$$x_i = \begin{cases} N(\mu_1, \sigma_1^2) & \text{با احتمال } \theta_i \\ N(\mu_2, \sigma_2^2) & \text{با احتمال } 1-\theta_i \end{cases} \qquad (۹)$$

۲.٤ روش ارزیابی حجم نمونه در رگرسیون لجستیک

فرضیه $H_0 : B_j = 0, j \notin A$ را در نظر می‌گیریم که B_j، j امین عنصر بردار B از پارامترهای رگرسیون لجستیک است. در این روش فرض بر آن است که j امین ویژگی در مدل وجود ندارد سپس تخمینی از بردار پارامترهای تحت H_0 صورت می‌گیرد. سپس بردار B_A را به دست آورده و تحت تناوب $H_1 : B_j \neq 0$، B_{A^*} به دست می‌آید. مجموعه فهرست A^* از A و فهرست j درست می‌شود سپس H_0 و H_1 می‌تواند با قوانین پارامترهای θ_i از توزیع برنولی اصلاح شده و به صورت $H_0 : \theta = \theta_A$ و $H_1 : \theta = \theta_{A^*}$ نوشته شود. ضمناً مقدار دقیق θ_i در هر موضوع مهم نبوده و ما فقط به مقدار جدا شده C_0 اهمیت می‌دهیم سر انجام داریم:

$$H_0 : 1 - C_0 = p_0, \quad H_1 : 1 - C_0 = p_1.$$

برای آزمایش فرضیه H_0 آماره $Z = \frac{\hat{p} - p_0}{\sqrt{p_0 c_0 / m}}$ و $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} y_i$ محاسبه می‌شود که $\hat{p}$ برآوردگر حداکثر درست نمایی برای θ می باشد. سرانجام فرمول (۱۰) برای m^* به دست می‌آید.

$$m^* = \frac{p_0 c_0 \left(Z_{1-\frac{\alpha}{2}} + Z_{1-B}\sqrt{\frac{p_1 c_1}{p_0 c_0}}\right)^2}{(p_1 - p_0)^2} \qquad (۱۰)$$

۳.٤ وارسی اعتبار (اعتبار سنجی متقابل)

اعتبارسنجی متقابل که گاهی تخمین گردشی نیز نامیده می‌شود، در امر پیش‌بینی مورد استفاده قرار می‌گیرد. استفاده از این روش باعث تعمیم نتایج حاصل از تحلیل آماری بر روی مجموعه داده، صرف نظر از داده‌های آموزشی خواهد شد. همچنین در صورت استفاده از اعتبارسنجی متقابل، به مفید بودن مدل مورد نظر در عمل پی برده می‌شود. به‌طور کلی یک دور از اعتبارسنجی متقابل شامل افراز داده‌ها به دو زیرمجموعه مکمل، انجام تحلیل بر روی یکی از آن زیرمجموعه‌ها (داده‌های آموزشی) و اعتبارسنجی تحلیلی بر روی داده‌های مجموعه دیگر (داده‌های تست) است. در پژوهش حاضر، به منظور کاهش پراکندگی، عمل اعتبارسنجی با افرازهای مختلف انجام و از نتایج اعتبارسنجی‌ها میانگین گرفته شده است. از جمله روش‌های اعتبار سنجی متقابل میتوان اعتبارسنجی متقابل $K-\text{Fold}$ و نمونه‌گیری تصادفی چند مرتب‌ها را نام برد [۱۱، ۱۲].

در این روش نمونه داده شده به دو دسته آموزشی و آزمایشی تقسیم می‌شود $D_L = \{x_i, y_i : i \in l\}$ و $D_T = \{x_i, y_i : i \in T\}$ که $l = L \sqcup T$.

حداکثر سازگاری با نرخ رابطه (۱۱) محاسبه می‌شود.

$$RS(m) = \frac{\text{AUC}(A, D_{T(m)}}{\text{AUC}(A, D_{L(m)}} \qquad (۱۱)$$

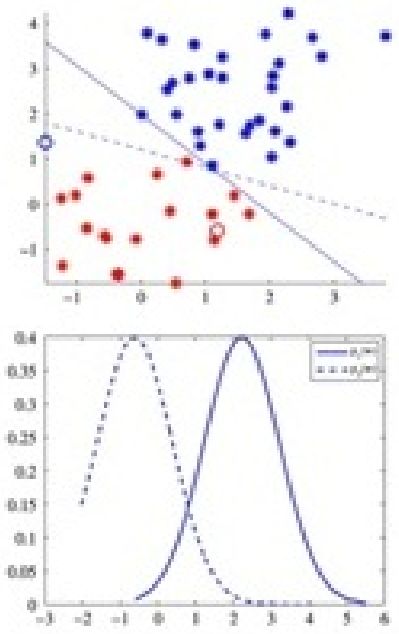
در این موضوع مدل f مجموعه آموزشی را تخمین می‌زند ولی برای توصیف مجموعه آزمایشی نمی‌تواند استفاده شود. حداکثر سازگاری ممکن است زمانی که حجم نمونه خیلی کم است اتفاق بیفتد. سرانجام برای برآورد m^* زمانی که مجموعه داده به دو دسته آموزشی و آزمایشی تحت یک نرخ معین تقسیم می‌شود باید حجم نمونه m افزایش داده شود.

با افزایش m مقدار $RS(m)$ به یک نزدیک می‌شود سرانجام m^* کافی اگر برای هر $m \geq m^*$ نرخ $RS(m)$ از یک مقدار معین $1 - \varepsilon_1$ بیشتر شود به دست می‌آید.

٤.٤ واگرایی کالبک لایبرر برای تخمین حجم نمونه

این روش بر اساس مقایسه توابع چگالی احتمالی پارامترهای مدل است [۱۳]. فرض کنید دو دسته مشابه از فهرست ویژگی‌ها $B_1 \in j$ و $B_2 \in j$. مجموعه‌های B_1 و B_2 مشابه خواهند بود اگر $|(B_1 \setminus B_2) \cup (B_2 \setminus B_1)| = 1$.

در این روش B_2 می‌تواند با حذف، جایگزین یا اضافه کردن یک عنصر به B_1 به دست آید. پارامترها در نمونه‌های مختلف $B_1 \neq B_2$ ارزیابی می‌شوند. شکل ۲ نشان می‌دهد که چه‌طور صفحه جداکننده عوض می‌شود زمانی که دو عنصر به نمونه اضافه می‌شود. اگر نمونه D_{B_1} به اندازه کافی بزرگ باشد بردار پارامتر B_1 بر اساس D_{B_1} که نباید به‌طور قابل توجهی متفاوت از B_2 به دست آمده از یک نمونه مشابه نمونه D_{B_2} باشد ارزیابی می‌شود. ساده‌ترین روش برای مقایسه آنها محاسبه فاصله اقلیدسی بین B_1 و B_2 است.



شکل۲. دو کلاس به وسیله صفحه‌ای جدا شده‌اند. یک خط چین موقعیت صفحه را بعد از اضافه شدن دو موضوع که داخل دایره مشخص‌اند نشان می‌دهد

با ثابت کردن مجموعه داده D و مدل $f_A = f(x_A^T B)$ رابطه (۱) به صورت رابطه (۱۲) نوشته می‌شود.

$$p(y|X, f_A) \equiv p(D|B, f_A) = \prod_{i=1}^{m} \theta_i^{y_i}(1-\theta_i)^{1-y_i} \quad (۱۲)$$

فرض می کنیم بردار پارامترهای رگرسیون B دارای توزیع نرمال با تابع چگالی رابطه (۱۳) است.

$$p(B|f_A, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{\frac{|A|}{2}} \exp\left(-\frac{\alpha}{2}\|B - B_0\|^2\right) \quad (۱۳)$$

برای پیدا کردن تابع چگالی احتمالی $P(B|D, \alpha, f_A)$ از پارامترهای رگرسیون از قضیه بیز رابطه (۱٤) استفاده شده است که $P(D|B, f_A)$ داده درست نمایی و $P(B|\alpha, f_A)$ یک تابع چگالی احتمالی پیش‌بینی معین است.

$$p(B|D, \alpha, f_A) = \frac{p(D|B, f_A)p(B|\alpha, f_A)}{p(D|\alpha, f_A)} \quad (۱٤)$$

با تعویض رابطه(۱۲) و (۱۳) با رابطه (۱٤) و با $Z(\alpha) = P(D|\alpha, f_A)$ رابطه (۱۵) به دست می‌آید.

$$p(B|D, f_A) = \frac{p(y|x, B, f_A)p(B|f_A, \alpha)}{Z(\alpha)} = \frac{\alpha^{\frac{|A|}{2}}}{(2\pi)^{\frac{|A|}{2}} Z(\alpha)} \exp\left(-\frac{\alpha}{2}\|B - B_0\|^2\right) \prod_{i=1}^{m} \theta_i^{y_i}(1-\theta_i)^{1-y_i} \quad (۱۵)$$

با فرض دو نمونه مشابه D_{B_1} و D_{B_2} و هم‌چنین روابط توزیع‌های مؤخر $P_1(B)$ و $P_2(B)$ تشابه این توزیع‌ها از رابطه (۱٦) به دست می‌آید.

$$D_{KL}(P_1, P_2) = \int_{B \in W} P_1(B) \ln \frac{P_1(B)}{P_2(B)} dB \quad (۱٦)$$

$$D_{KL}(p_1, p_2) = \int_{\beta \in w} p_1(\beta) \ln \frac{p_1(\beta)}{p_2(\beta)} d\beta \quad (۱۷)$$

برای تخمین کمترین حجم نمونه m^* به‌طور تصادفی ویژگی‌هایی از مجموعه داده یکی یکی حذف می شود و در نتیجه کاهش حجم نمونه m و محاسبه توزیع موخر بردار B با رابطه (۱٤) و سپس رابطه (۱۷) بین توابع چگالی احتمالی از پارامترهای مورد ارزیابی در مجموعه‌های داده مشابه تخمین زده می‌شود. این عمل N بار انجام شده و سپس میانگین نتایج گرفته می‌شود حجم نمونه m^* کافی به نظر می‌رسد اگر رابطه (۱۷) کمتر با برخی ε_2 معین برای $m \geq m^*$ تغییر یابد.

با یک عمل جمع انتگرال رابطه (۱۷) تقریب زده می‌شود. در این قسمت از بردار پارامترهای رگرسیون لجستیک $\hat{B}$ به دست آمده از حل مسئله (٤) به عنوان بردار میانگین یعنی B_0 در رابطه (۱٦) استفاده می‌شود. سپس نمونه‌ای از ۵۰۰ بردار B بر اساس توزیع نرمال $N(\hat{B}, I_A)$ تولید می‌شود برای هر کدام از آنها $P_1(B)$

و $P_2(B)$ محاسبه می‌شود و از حال به بعد به عنوان یک جمع رفتار خواهیم کرد.

۵ محاسبه آزمایشات

آزمایش روی داده‌های واقعی

مجموعه داده شامل مشاهدات غلظت ۲۰ پروتئین در سلول خونی برای مریض‌های دو کلاس است. شامل به ترتیب ۳۱ و ۱٤ موضوع. همه ویژگی یا نشانگرهای زیستی در سطرهای اول و دوم جدول ۱ لیست شده‌اند.

جدول ۲ مجموعه بهینه از ویژگی‌ها را مشابه مقادیر حداکثر AUC و مقادیر دقیق AUC نشان می‌دهد. در اینجا $k = 5$ مجموعه‌های بهینه برای بررسی انتخاب شده بود.

به دلیل هزینه بالای بررسی‌های پزشکی هر مریض کاهش تعداد آزمایشات ضروری است بنابراین پیشنهاد شده که فقط ویژگی‌های خیلی مهم اندازه‌گیری شود. با داشتن فهرست‌های جمعی از همه ویژگی‌های جدول ۲ مجموعه‌ای از شاخصه‌های ویژگی‌های بسیار مهم به دست می آوریم. $S = U$ برای هر ویژگی تعداد دفعاتی که باید شامل S باشد محاسبه می‌شود. جدول ۱ این تعداد را برای هر ویژگی نشان می‌دهد.

در نمودار فراوانی شکل ۳ مقدار حجم نمونه m^* برای هر ویژگی به طور جداگانه با رابطه‌های (۹) و (۱۱) نشان داده شده است. حجم نمونه m^* تنها برای ویژگی‌هایی که شامل مدل هستند، نتیجه آنها مهم نیست و نباید در نظر گرفته شوند محاسبه شده است.

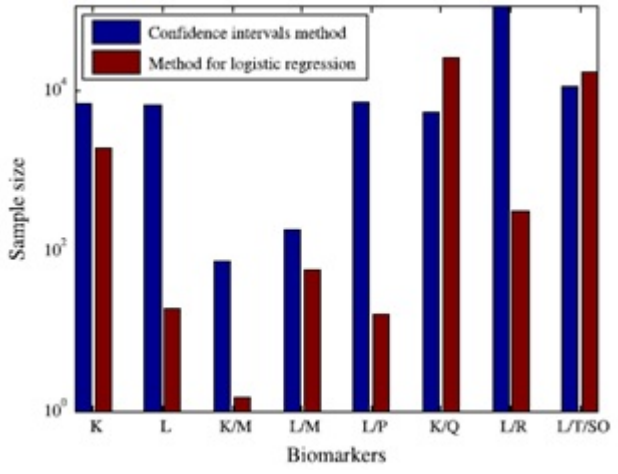
تخمین‌های حجم نمونه و به‌دست آمده از روابط (۹) و (۱۱) وابستگی مشابهی روی یک فهرست ویژگی دارد. به دلیل اینکه در هر دو روش تخمین حجم نمونه j امین ویژگی وابسته به این است که ویژگی مهم چه‌طور است در رگرسیون لجستیک ویژگی‌های مهم، یک مقدار قابل توجهی از عنصر مشابه B_j از بردار پارامترها دارند.

جدول ۱: تعداد ورودی به مجموعه بهینه K برای هر ویژگی

K	L	K/M	L/M	K/N	K/O	L/O	K/P	L/P	K/Q
5	4	3	1	0	0	0	0	2	1
K/R	L/R	$L/R/SA$	$L/T/SA$	$L/T/SO$	U/V	U/W	U/X	U/Y	U/Z
0	1	0	0	1	0	0	0	0	0

جدول ۲: نتایج حاصل از انتخاب ویژگی

$\boldsymbol{A}$	$\boldsymbol{S(A)}$
$K, L, L/P$	0.9750
$K, L, K/M, K/Q$	0.9671
$K, L, L/M, L/T/SO$	0.9933
$K, L, K/M, L/R$	0.9867
$K, K/M, L/P$	0.9742



شکل۳. برآورد حجم نمونه با استفاده از روش فواصل اعتماد و روش رگرسیون لجستیک محاسبه برای ویژگی‌های آموزنده‌ترین

٦ نتیجه‌گیری

این مقاله یک الگوریتم که بیماران دارای اختلال در سیستم قلبی-عروقی را طبقه‌بندی می‌کرد را نشان داد. برای انتخاب مدل رگرسیونی، یک الگوریتم تحقیقی کامل استفاده شد. محققین برای تخمین حجم نمونه یک متد نوین فراهم کرده‌اند که بر اساس تکنیک تقاطع اعتبار می‌باشد و از واگرایی کالبک لایبرر بین دو توزیع از پارامترهای مدل، مورد ارزیابی در زیر مجموعه داده‌های مشابه استفاده می‌کند. چهار الگوریتم متفاوت تخمین حجم نمونه مقایسه شده‌اند.

با این مقاله محققین روش‌های احتمالی و تحلیلی برای داده‌کاوی و آموزش آماری را بیشتر معرفی می‌کنند. در آینده بایستی با این زمینه در ریاضیات کاربردی پیشرفته به سوی اصلاح در پیش‌بینی‌ها در همه زمینه‌های علمی، مهندسی و زندگی واقعی حرکت کرد.

مراجع

[1] D. Hosmer, S. Lemeshow, Applied Logistic Regression, Wiley, NY, 2000.

[2] B. Rosner, Fundamentals of Biostatistics, Duxbury Press, 1999.

[3] E. Demidenko, Sample size determination for logistic regression revisited, Statistics in Medicine 26 (2007), pp. 3385–3397.

[4] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[5] D. J. C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003.

[6] J. Friedman, T. Hastie, R. Tibshirani, Additve logistic regression: a statistical way of boosting, The Annals of Statistics 28 (2) (2000), pp. 337–407.

[7] B. Efron, et al., Discussion of least square regression, least angle regression, The Annals of Statistics 32 (2) (2004), pp. 465–469.

[8] T. Fawcet, ROC Graphs: Notes and Practical Considerations for Researchers, Kluwer Academic Publishers, 2004.

[9] E. Kurum, K. Yildirak, G.-W. Weber, A classification problem of credit risk rating investigated and solved by optimisation of the ROC curve, Central European Journal of Operations Research (CEJOR) (2012) (special issue at the occasion of EURO XXIV 2010 in Lisbon).

[10] http://www.medicalbiostatistics.com.

[11] J. O. Berger, L. R. Pericchi, Training samples in objective Bayesian model selection, The Annals of Statistics 32 (3) (2004), pp. 841–869.

[12] S. Amari, N. Murata, K.-R. Muller, M. Finke, H.H. Yang, Asymptotic statistical theory of overtraining and cross-validation, IEEE Transactions on Neural Networks 8 (5) (1997), pp. 985–996.

[13] F. Perez-Cruz, Kullback–Leibler divergence estimation of continuous distributions, in: IEEE International Symposium on Information Theory, 2008.

[14] S. Bush, (2015), Sample Size Determination for Logistic Regression: A Simulation Study, Communications in Statistics - Simulation and Computation. Volume 44, Issue 2, pp. 360-373.

[15] E. Frank, Jr. Harrell, (2015), Binary Logistic Regression. Part of the series Springer Series in Statistics, pp. 219-274.