تقریب‌های قوی برای فرآیندهای چندکی بازنمونه‌گیری شده و کاربرد آن در روش‌شناسی ROC

سارا خزاعی[۱] - وحید فکور[۲]

[۱] دانشگاه آزاد اسلامی، واحد مشهد، دانشکده علوم پایه
[۲] دانشگاه فردوسی مشهد، دانشکده علوم ریاضی

چکیده: خم ROC[۱] از رسم مقادیر نرخ مثبت واقعی در مقابل نرخ مثبت کاذب حاصل می‌شود که به طور گسترده‌ای در علوم پزشکی برای توانایی در اندازهٔ درستی آزمون‌های تشخیصی استفاده می‌شود، از نظر ریاضی خم ROC ترکیب تابع بقای یک جامعه با تابع چندکی از جامعه دیگراست. در این مقاله ابتدا تقریب قوی برای فرآیندهای چندکی توزیع‌های بازنمونه‌گیری خودگردان‌ساز بیزی[۲] (BB) را مورد مطالعه قرار می‌دهیم و در ادامه با استفاده از این نتایج، تقریب‌های قوی برای BB فرآیند ROC در عبارت‌های با دو فرآیند کی‌فر مستقل را بررسی می‌کنیم (گو و گوسال، ۲۰۰۸).

واژه‌های کلیدی: تقریب قوی، خم ROC، خودگردان‌ساز بیزی، فرآیند چندکی، فرآیند کی‌فر

۱ مقدمه

خم ROC، نخستین بار در جنگ جهانی دوم برای تحلیل سیگنال‌های رادار استفاده شد و سپس در سال ۱۹۵۰ در رشته روانشناسی مورد استفاده قرار گرفت (گرین، ۱۹۶۶) و از سال ۱۹۷۰ به عنوان یک ابزار ضروری برای اندازه‌ی درستی آزمون‌های تشخیصی برای جداسازی افراد بیمار از سالم مطرح شد (متز، ۱۹۷۸). در مطالعات پزشکی که افراد به دو گروه سالم و بیمار طبقه‌بندی می‌شوند، ابتدا نقطه‌ی برش[۳] (c) را انتخاب کرده، افراد با نتایج آزمون بیشتر از c به عنوان بیمار در نظر گرفته می‌شوند، در غیر این صورت سالم هستند.
فرض می‌کنیم X دارای تابع توزیع تجمعی F مربوط به جامعه‌ی افراد سالم و Y دارای

[۱] Receiver operating characteristic
[۲] Bayesian Bootstrap
[۳] Cut off point

تابع توزیع تجمعی G مربوط به جامعه‌ی افراد بیمار است. حساسیت[۴] یک آزمون را به صورت $SE(c) = ۱ - G(c)$ تعریف می‌کنیم که احتمال درست طبقه‌بندی شدن یک فرد بیمار است و ویژگی[۵] یک آزمون به صورت $SP(c) = F(c)$ است که احتمال درست طبقه‌بندی شدن یک فرد سالم است (سیه و تارنبال، ۱۹۹۶).

خم ROC نموداری از $SE(c)$ یا نرخ مثبت واقعی روی محور عمودی در مقابل $۱ - SP(c)$ یا نرخ مثبت کاذب روی محور افقی است.

خم ROC را به صورت $\{(P(X > c), P(Y > c)) : X \sim F, Y \sim G, c \in \mathbb{R}\}$ تعریف می‌کنیم، با فرض این که $t = P(X > c)$، می‌توان تابع خم ROC را به صورت $\{(t, R(t))\}$ در نظر گرفت که $۰ < t < ۱$، $R(t) = \overline{G}(\overline{F}^{-۱}(t))$ است، $\overline{F}(x) = ۱ - F(x)$ و $\overline{G}(y) = ۱ - G(y)$، تابع بقای متغیرهای مستقل X و Y هستند و مشتق آن به صورت $R'(t) = g(\overline{F}^{-۱}(t))/f(\overline{F}^{-۱}(t))$ است.

از ویژگی‌های خم ROC می‌توان به موارد زیر اشاره کرد:

۱) یک خم ROC می‌تواند تمام اطلاعات مشخصه یک آزمون تشخیصی را در یک نمودار واحد نشان دهد که برای هدف‌های طبقه‌بندی شده مورد استفاده قرار می‌گیرد.

۲) آزمون‌های تشخیصی متفاوت توسط خم‌های ROC متناظرشان می‌توانند مقایسه شوند، بدین معنی که اگر یک خم ROC بالاتر از خم دیگر قرار گیرد، خم بالاتر را به عنوان خم ROC بهتر در نظر می‌گیریم (ویاند و همکاران، ۱۹۸۹). مقایسه‌ی خم‌های ROC در شکل (۱) نشان داده شده است. سطح زیر خم[۶] (AUC) را می‌توان به صورت $P(Y > X)$ تعریف کرد که بیانگر دقت تشخیص آزمون می‌باشد (بامبر، ۱۹۷۵).

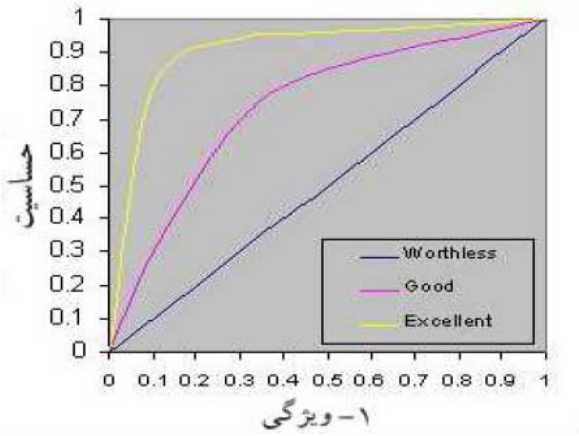
از جمله روش‌های محاسبه‌ی AUC می‌توان به استفاده از قاعده ذوزنقه‌ای، برازاندن داده‌ها به مدل تبدیل دو نمونه‌ای با استفاده از برآورد ماکزیمم درستنمایی و استفاده از آماره‌ی من–ویتنی نیز اشاره کرد.

برآوردگر ناپارامتری ROC توسط جایگزینی در همتای تجربی آن به دست می‌آید و تغییرپذیری آن با استفاده از خودگردان‌ساز[۷] برآورد می‌شود (پپ، ۲۰۰۳)، که در ادامه به آن می‌پردازیم.

در این مقاله در بخش دوم مروری بر فرآیند چندکی خواهیم داشت، در بخش سوم خودگردان‌ساز بیزی را معرفی می‌کنیم و در بخش پایانی تقریب‌های قوی گوسی برای

---

۴ Sensitivity
۵ Specificity
۶ Area under curve
۷ Bootstrap



شکل ۱.: مقایسه خم‌های ROC

فرآیندهای چندکی BB و خم ROC را مورد مطالعه قرار می‌دهیم.

۲ فرآیند چندکی

ابتدا فرآیند تجربی را تعریف کرده و سپس فرآیند چندکی را معرفی می‌کنیم.

تعریف **۱۳** فرض کنید $\{X_i, i \geq ۱\}$ دنباله‌ای از متغیرهای تصادفی مستقل و هم‌توزیع، با توزیع مشترک F باشند، تابع توزیع تجربی این متغیرها عبارت است از:

$$\mathbb{F}_n(x) = \frac{۱}{n}\sum_{i=۱}^{n} I(X_i \leq x)$$

که در آن $I(\cdot)$ تابع نشانگر است.

اگر $X_{۱:n} \leq X_{۲:n} \leq \cdots \leq X_{n:n}$ آماره‌های ترتیبی نمونه‌ی تصادفی $X_۱$، ...، X_n باشد، آن‌گاه تابع توزیع تجربی را به صورت زیر نشان می دهیم

$$\mathbb{F}_n(x) = \begin{cases} ۰ & x < X_{۱:n} \\ \frac{j}{n} & X_{j:n} \leq x < X_{j+۱:n},\ j = ۱, \cdots, n-۱ \\ ۱ & X_{n:n} \leq x \end{cases}$$

فرآیند تجربی را به شکل زیر تعریف می‌کنیم:

$$\mathbb{J}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x))$$

تعریف ۱۴ فرض کنید X_1، ...، X_n نمونه‌ی تصادفی از توزیع F باشند، آن‌گاه تابع چندک به صورت زیر تعریف می‌شود:

$$F^{-1}(y) = \inf\{x\ ;\ F(x) \geq y\}\ ; \qquad 0 < y < 1$$

و برآوردگر نمونه‌ای آن برابر است با:

$$\mathbb{F}_n^{-1}(y) = \inf\{x\ ;\ \mathbb{F}_n(x) \geq y\}\ ; \qquad 0 < y < 1$$

و یا

$$\mathbb{F}_n^{-1}(y) = X_{K:n}, \qquad \frac{K-1}{n} < y \leq \frac{K}{n}, \qquad K = 1, 2, \cdots, n$$

فرآیند چندکی را به صورت زیر تعریف می‌کنیم:

$$\mathbb{Q}_n(y) = \sqrt{n}(\mathbb{F}_n^{-1}(y) - F^{-1}(y))$$

۳ خودگردان‌ساز بیزی (BB)

روش خودگران‌ساز توسط افران در سال ۱۹۷۹ ارائه شد. خودگردان‌ساز یک روش برای برآورد واریانس و نیز یافتن توزیع آماره‌ها می‌باشد. همچنین می‌توانیم از خودگردان‌ساز برای ساختن بازه‌های اطمینان استفاده کنیم (واسرمن، ۲۰۰۶). پس از روش خودگردان‌ساز، رابین در سال ۱۹۸۱ روش خودگردان‌ساز بیزی را معرفی کرد.
فرض کنید X_1، ...، X_n نمونه‌ی تصادفی از توزیع F باشند، در روش خودگردان‌ساز نمونه‌ی تصادفی X_1^*، ...، X_n^* را به روش نمونه‌گیری با جایگذاری از نمونه‌ی X_1، ...، X_n انتخاب می‌کنیم که هر مشاهده‌ی X_i^* با احتمال $\frac{1}{n}$ انتخاب می‌شود. اما در روش BB به جای انتخاب با احتمال $\frac{1}{n}$، یک توزیع احتمال پسین برای X_i ها به صورت زیر در نظر می‌گیریم:
ابتدا $n-1$ متغیر تصادفی یکنواخت از بازه‌ی $[0, 1]$ را در نظر می‌گیریم.
$V_{1:n-1}$، ...، $V_{n-1:n-1}$ را به عنوان آماره‌های ترتیبی V_1، ...، V_{n-1} فرض می‌کنیم که $V_{0:n-1} = 0$ و $V_{n:n-1} = 1$ هستند.
آن‌گاه $\Delta_{j:n} = V_{j:n-1} - V_{j-1:n-1}$ را نمو بین آماره‌های ترتیبی یکنواخت معرفی

می‌کنیم.

بردار $\Delta = \begin{pmatrix} \Delta_{1:n} \\ \vdots \\ \Delta_{n:n} \end{pmatrix}$ برای تعیین احتمالات نمونه‌ی BB استفاده می‌شود (رابین، ۱۹۸۱).

با توجه به نمونه‌ی X_1، ...، X_n، تابع توزیع تجمعی BB به صورت $\mathbb{F}_n^{\#}(x) = \sum_{1\leq j\leq n} \Delta_{j:n} I(X_{j:n} \leq x)$ است.

به سادگی تابع چندکی BB به صورت زیر به دست می‌آید:

$$\mathbb{F}_n^{\#-1}(y) = \begin{cases} X_{j:n} & V_{j-1:n-1} < y \leq V_{j:n-1}, j = 1, \cdots, n \\ X_{0:n} & y = 0 \end{cases}$$

فرآیند تجربی BB و فرآیند چندکی BB را به صورت زیر تعریف می‌کنیم:

فرآیند تجربی BB:

$$\mathbb{J}_n^{\#}(x) = \sqrt{n}(\mathbb{F}_n^{\#}(x) - \mathbb{F}_n(x))$$

فرآیند چندکی BB:

$$\mathbb{Q}_n^{\#}(y) = \sqrt{n}(\mathbb{F}_n^{\#-1}(y) - \mathbb{F}_n^{-1}(y))$$

۴ تقریب‌های قوی گوسی برای فرآیندهای چندکی BB و خم ROC

فرض می‌کنیم X_1، ...، X_n نمونه‌ی تصادفی از توزیع F باشند و دامنه‌ی X را به صورت $[a, b]$ تعریف می‌کنیم که

$$b = \inf\{x \ : \ F(x) = 1\} \ , \ a = \sup\{x \ : \ F(x) = 0\}$$

است و $l_n = n^{-\frac{1}{4}}(\log\log n)^{\frac{1}{4}}(\log n)^{\frac{1}{2}}$ را در نظر می‌گیریم.

شرایط A (چورگو و روز، ۱۹۷۸) و B را فرض می‌کنیم که برای یک کلاس بزرگ از تابع‌های توزیع برقرار هستند.

شرط A: تابع توزیع پیوسته F را در نظر می‌گیریم که مشتق مرتبه دوم آن روی (a, b) وجود دارد و $F' = f \neq 0$ روی (a, b) است و برای بعضی $\gamma > 0$ داریم:

$$\sup_{a<x<b} F(x)(1 - F(x))|f'(x)/f^2(x)| \leq \gamma.$$

شرط B: F و G تحت شرط A برقرارند و شرایط زیر را نیز در نظر می‌گیریم:

$$\sup_{a<x<b} F(x)(1-F(x))\left|\frac{g'(x)}{f^2(x)}\right| < \infty$$

و

$$\sup_{a<x<b} F(x)(1-F(x))\left|\frac{g(x)}{f(x)}\right| < \infty$$

تقریب قوی دارای بیشینه‌ای طولانی می‌باشد، دستاوردهای بیشماری در این زمینه موجود است. قضایای زیر برخی از نتایج در این زمینه را نشان می‌دهند.
کومولوش و همکاران (۱۹۷۵) نشان دادند که می‌توان $\mathbb{J}_n$ را توسط دنباله‌ای از پل‌های براونی[۸] تقریب زد. چورگو و رِوِز (۱۹۷۸) ثابت کردند که $\mathbb{Q}_n$ را می‌توان توسط یک فرآیند کی‌فر[۹] تقریب زد. لو (۱۹۸۷) قضیه‌ی تقریب قوی برای تابع توزیع تجمعی فرآیندهای خودگردان‌ساز و BB را مورد مطالعه قرار داد. در ادامه ابتدا فرآیند کی‌فر را تعریف کرده و در ادامه به برررسی قضایای تقریب قوی از (گو و گوسال، ۲۰۰۸) می‌پردازیم.

تعریف ۱۵ فرآیند $\{K(x,y) : 0 \le x \le 1 \ , \ 0 \le y < \infty\}$ یک فرآیند کی‌فر نامیده می‌شود هرگاه بتوان آن را به صورت $K(x,y) = W(x,y) - xW(1,y)$ نوشت که $W(x,y)$ یک فرآیند وینر دوپارامتری است.

لم ۱ فرض کنید X_1، ...، X_n نمونه‌ی تصادفی از توزیع F، تحت شرط A باشند، V_1، ...، V_{n-1} نمونه‌ی تصادفی از توزیع $U(0,1)$ و مستقل از X_i ها هستند و $\tilde{\mathbb{U}}_{n-1}(x)$ را به عنوان تابع تجربی V_1، ...، V_{n-1} در نظر می‌گیریم. آن‌گاه:

$$\sup_{0<y<1} \sqrt{n}\left|\mathbb{F}_n^{\#-1}(y) - \mathbb{F}_n^{-1}\left(\tilde{\mathbb{U}}_{n-1}(y)\right)\right| = O\left(n^{-\frac{1}{2}}\log n\right), \ a.s. \qquad (1)$$

علاوه بر آن، یک فرآیند کی‌فر K وجود دارد به طوری که:

$$\sup_{\delta_n^{\#} \le y \le 1-\delta_n^{\#}} \left|\sqrt{n} f(F^{-1}(y))(\mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) - \mathbb{F}_n^{-1}(y)) - n^{-\frac{1}{2}} K(y,n)\right| = O(l_n), \ a.s. \qquad (2)$$

□

[۸] Brownian Bridge
[۹] Kiefer Process

قضیه **۱۲** فرض کنید X_1، ...، X_n نمونه‌ی تصادفی از توزیع F تحت شرط A باشند. آن‌گاه $\mathbb{Q}_n^{\#}(y)$ را می‌توان توسط یک فرآیند کی‌فر K تقریب زد،

$$\sup_{\delta_n^{\#} \le y \le 1-\delta_n^{\#}} \left| f(F^{-1}(y))\mathbb{Q}_n^{\#}(y) - n^{-\frac{1}{2}} K(y,n) \right| = O(l_n),\ a.s. \qquad (3)$$

که $\delta_n^{\#} = \delta_n + n^{-\frac{1}{2}}(\log\log n)^{\frac{1}{2}}$ و $\delta_n = 25 n^{-1}\log\log n$.

□

قضیه **۱۳** فرض کنید X_1، ...، X_m نمونه‌ی تصادفی از توزیع F و Y_1، ...، Y_n نمونه‌ی تصادفی از توزیع G، تحت شرایط A و B باشند، آن‌گاه

$$\mathbb{R}(t) = R(t) + R'(t)\frac{K_1(t,m)}{m} + \frac{K_2(R(t),n)}{n} + O(\alpha_m^{-1}\tau_m), \qquad t \in (\alpha_m, 1-\alpha_m) \qquad (4)$$

$$\mathbb{R}^{\#}(t) = \mathbb{R}(t) + R'(t)\frac{K_1(t,m)}{m} + \frac{K_2(R(t),n)}{n} + O(\alpha_m^{-1}\tau_m), \qquad t \in (\alpha_m^{\#}, 1-\alpha_m^{\#}) \qquad (5)$$

که K_1 و K_2 فرآیندهای کی‌فر مستقل هستند و

$$\mathbb{R}(t) = \mathbb{R}_{m,n}(t) = \overline{\mathbb{G}}_n(\overline{\mathbb{F}}_m^{-1}(t)),$$

$$\mathbb{R}^{\#}(t) = \mathbb{R}_{m,n}^{\#}(t) = \overline{\mathbb{G}}_n^{\#}(\overline{\mathbb{F}}_m^{\#-1}(t)),$$

$\tau_m = \frac{l_m}{\sqrt{m}}$، $\alpha_m^{\#} = \alpha_m + (m-1)^{-\frac{1}{2}}(\log\log(m-1))^{\frac{1}{2}}$ و $\alpha_m^{-1} = O\left(m^{\frac{1}{4-\epsilon}}\right)$ که $\epsilon \in \left(0, \frac{1}{4}\right)$ را در نظر می‌گیریم.

□

تذکر ۱. نتیجه‌ی مشابه (۴) در (سیه و تارنبال، ۱۹۹۶) اثبات شده است. تفاوت بین این دو یک موازنه‌ی متفاوت بین بازه‌ی اعتبار و نرخ خطای تقریب است.

تذکر۲. تقریب قوی برای فرآیند چندکی خودگردان‌ساز و تقریب قوی برای برآوردگر خودگردان‌ساز ROC را می‌توان مشابهاً با همین نرخ خطا به دست آورد امّا با دامنه تغییرات متمایز t.

بحث و نتیجه‌گیری

در این مقاله، ابتدا به بررسی خم ROC پرداخته که عملکرد یک آزمون تشخیصی را توصیف می کند که نموداری از حساسیت درمقابل ۱ − ویژگی است. سپس فرآیندهای چندکی BB را توسط یک فرآیند کی فر با نرخ همگرایی $O(l_n)$ تقریب زده و در ادامه با استفاده از این نتایج، تقریب های قوی برای BB فرآیند ROC ، مرتبط با دو فرآیند کی فر مستقل را مور بررسی قرار دادیم.

مراجع

Bamber, D. (1975), The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, J. Math. Psychol. 12, pp. 387-415.

Csőrgö ,M. and Révész, P. (1978) , Strong approximations of the quantile process, Ann. Statist. 6, pp. 882-894.

Green, D.M. and Swets. J.A.(1966), Signal Detection Theory and Psychophysics, JohnWiley & Sons, NewYork.

Gu, J. and Ghosal ,S.(2008), Strong approximations for resample quantile processes and application to ROC methodology. Journal of Nonparametric Statistics, 20:3, 229-240.

Hsieh, F.S. and Turnbull, B.W.(1996) , Nonparametric and semiparametric estimation of the receiver operating characteristic curve, Ann. Statist. 24 ,pp. 25-40.

Komlós, J. and Major,.P. and Tusnády ,G .(1975) , An approximation of partial sums of independent RV's and the sample DF. I., Z.Wahrsch. Verw. Gebiete, 32, pp. 111-131.

Lo, A.Y ,(1987) A large sample study of the Bayesian bootstrap, Ann. Statist. 15, pp. 360-375.

Metz, C.E.(1978) , Basic principles of ROC analysis, Seminars Nuc. Med. 8, pp. 283-298.

Pepe, M.S.(2003), The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford University Press, Oxford.

Rubin, DB.(1981), The Bayesian bootstrap. The Annals of Statistics, 9:130-134.

Wasserman, L.(2006), All of Nonparametric Statistics,springer.

Wieand ,S.and Gail, M.H. and James, B.R. and James, K.L.(1989), A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data, Biometrika 76 ,pp. 585-592.