한국통계학회논문집
2011, 18권, 3호, 343–355

DOI: 10.5351/CKSS.2011.18.3.343

정규혼합에서 분류정확도 측도들의 최적기준

유현상[a], 홍종선[1,b]

[a]성균관대학교 응용통계연구소, [b]성균관대학교 통계학과

요 약

두 분포함수의 혼합모형을 가정한 자료에서 적절한 분류점을 찾고 평가하는 것은 중요한 문제이다. 분류정확도 측도로 많이 사용하는 아홉 종류의 MVD, Youden지수, $(0,1)$까지 최단기준, 수정된$(0,1)$까지 최단기준, SSS, 대칭점, 정확도면적, TA, TR에 대하여 설명하고, 이 측도들의 관계를 발견하면서 정확도 측도들의 조건을 몇 개의 범주로 군집화한다. 정규혼합분포를 가정하여 군집된 측도들에 기반하는 분류점들을 구하고, 그 분류점에 대응하는 제I종 오류율과 제II종 오류율 그리고 두 종류의 오류율합을 구하여 크기를 비교하고 토론한다. 추정된 혼합분포에 대하여 어떤 분류정확도 측도의 제I종과 II종 오류율 또는 오류율합이 최소인지를 탐색할 수 있으며 자주 인용하는 정확도 측도의 장점과 단점을 파악할 수 있다.

주요용어: 민감도, 분류, 오류율, 정확도, 특이도, 판별.

1. 서론

두 분포함수의 혼합분포로부터 판별력을 극대화하는 분류점(절단점; threshold, cut-off)을 추정하는 최근의 연구는 차주(borrower)의 미래상태인 부도(default; d) 혹은 정상(non-default; n) 상태에 대한 예측력을 최대화하는 신용평가(credit evaluation) 분야와 환자가 질병(disease) 또는 정상(non-disease) 상태인지를 진단하는 의학통계분야 등에서 많이 활용되고 있다. 본 연구에서는 신용평가에서 차주의 신용가치를 기준으로 대출상환능력에 따라 부도와 정상상태를 판별하는 문제를 고려하자. 확률변수 X는 스코어 변수로 연속형 실수값이다. 모수공간은 부도와 정상상태로 가정하여 $\Theta = \{\theta_d, \theta_n\}$로 정의한다. $F_d(x)$와 $F_n(x)$를 각각 차주의 부도와 정상상태에서 스코어의 조건부 누적분포함수 $P(X \le x|\theta_d)$와 $P(X \le x|\theta_n)$로 정의하며, 스코어 확률변수 X의 누적분포함수 $F(x)$는 다음과 같이 가정한다.

$$F(x) = \gamma F_d(x) + (1-\gamma)F_n(x), \tag{1.1}$$

여기서 γ는 전체부도율이다. 즉 $\gamma = P(\Theta = \theta_d)$.

ROC(Receiver Operating Characteristic) 곡선은 성과(performance)를 기반으로 분류모형(classification model) 또는 분류자(classifiers)를 시각화하며 평가할 수 있는 유용한 방법이다. ROC 곡선은 분류자의 'sensitivity'(민감도)와 '1 − specifity'(1 − 특이도) 사이에 교환(trade-off)을 나타내는 신호탐지이론에서 오랫동안 사용되었으며, 의사결정과 의학진단의 체계에서 폭넓게 사용되었다. ROC 곡선의 특성에 관한 연구와 실증분석에서 ROC 분석을 응용하는데 관련된 정보는 많이 있으나 2000년 이후의 연구인 Provost와 Fawcett (2001), Sobehart와 Keenan (2001), Zhou 등 (2002), Engelmann 등 (2003), Fawcett (2003), 홍종선과 최진수 (2009), 홍종선 등 (2010) 이외의 많은 문헌에서 발견할 수 있다.

부도 예측모형의 분류성과를 평가하는 가장 기본적인 접근 방법은 예측된 부도(또는 정상) 차주 수를 고려하고 이를 실제로 발생한 부도(또는 정상) 차주 수와 비교한다. 이를 표현하는 기본적인 방법은 표 1과 같이 분할표 또는 혼동행렬(confusion matrix) 형태로 표현된다.

[1] 교신저자: (110-745) 서울시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수.
E-mail: cshong@skku.ac.kr

표 1: 혼동 행렬

		실제	
		부도	정상
예측	부도	TP	FP
	정상	FN	TN
합		p	n

ROC 곡선은 이항분류자(binary classifiers)를 사용하여 각 분류점에서 얻는 비율들로 구성되어 있으며, 실제부도를 부도로 정확히 예측하는 비율 TPR(true positive rate 또는 hit rate, recall, sensitivity)과 실제정상을 부도로 잘못 예측하는 비율 FPR(false positive rate 또는 false alarm rate, 1 − specificity)을 각각 수직축과 수평축 좌표에 대응시킨 그래프로 다음과 같은 좌표로 표현된다 (상세한 정보는 Pepe (2003)와 Tasche (2006) 참조).

$$(\text{FPR}, \text{TPR}) = (F_n(x), F_d(x)) = (u, \text{ROC}(u)),$$

여기서 TPR = TP/p, FPR = FP/n이며 $\text{ROC}(u) = F_d(F_n^{-1}(u))$, $u \in [0, 1]$.

혼합분포, 혼동행렬 그리고 ROC 곡선에서 정의된 분류정확도 측도들이 많이 존재한다. 본 연구에서는 Cantor 등 (1999), Greiner와 Gardner (2000), Freeman과 Moisen (2008) 그리고 Liu 등 (2009) 이외의 많은 문헌에서 논의한 분류정확도를 측정하는 다양한 측도들의 성격을 파악하고, 측도들 사이의 관계를 정립한다. 그리고 여러 측도들을 기반으로 하는 분류점들을 구한다. 분류점을 비교 탐색하기 위하여 정규혼합(normal mixture) 분포를 가정하여 정확도 측도에 대응하는 분류점을 구하고, 그 분류점에 대응하는 오류율을 구하여 비교 분석한다. 본 연구의 2절에서는 많은 종류의 분류정확도 측도들 중에서 대표적인 아홉 종류의 측도를 간략히 설명하고, 3절에서는 2절에서 언급한 측도들의 관계를 정리한다. 그리고 측도가 포함하는 분류정확도의 조건함수를 유도하여 성격이 유사한 것들을 모아 일곱 종류의 범주로 구분한다. 4절에서는 다양한 정규혼합분포를 고려해서 최소 오류율을 나타내는 분류정확도의 조건함수와 전체부도율 γ와의 관계를 탐색한다. 즉 주어진 전체부도율 γ에 의존하는 특정한 정규혼합분포의 경우에 제I종 오류율이 최소일 때의 정확도 측도는 무엇이며 어떠한 정확도 측도가 최소의 제II종 오류율을 나타내는지 파악한다. 마지막으로 결론은 5절에서 유도한다.

2. 정확도 측도들

(1) MVD

ROC 분석에서 많이 사용하는 분류정확도 측도인 MVD(maximum vertical distance)는 ROC 곡선과 대각선(chance line)의 최대 수직 거리로서 다음과 같이 정의된다 (Krzanowski와 Hand, 2009).

$$\text{MVD} = \max |\text{ROC}(u) - u|,$$

여기서 $\text{ROC}(u) = F_d(F_n^{-1}(u))$, $u \in [0, 1]$이며, $\text{MVD} = \max |F_d(x) - F_n(x)|$로 표현된다. Ward (1986)가 제안한 최대 DPR(differential positive rate)은 $\text{DPR} = \max |\text{TPR} - \text{FPR}|$으로 정의되며 $\max |F_d(x) - F_n(x)|$으로 표현할 수 있으므로 DPR은 MVD와 동일하다.

(2) Youden지수

최적분류점을 찾을 때 가장 많이 이용하는 정확도 측도 중의 하나는 Youden (1950)에 의해 제안된 J 지수이며, 두 분포의 최대 수직 차이를 의미한다.

$$J = \max[F_d(x) - F_n(x)].$$

(3) $(0, 1)$까지최단거리기준(The closest-to-$(0, 1)$ criterion)

$(0, 1)$까지최단거리기준(이하 $(0, 1)$기준)은 ROC 곡선으로부터 완벽하게 분류될 때의 완벽점인 $(0, 1)$까지 가장 거리가 가까운 점에 대응하는 스코어를 분류 기준으로 설정하며 다음과 같이 정의한다 (Perkins와 Schisterman, 2006).

$$(0,1)\text{기준} = \min\left\{(1 - F_d(x))^2 + F_n(x)^2\right\} \quad \text{또는} \quad \min\sqrt{(1 - F_d(x))^2 + F_n(x)^2}.$$

(4) 수정된 $(0, 1)$까지최단거리기준(The amended closest-to-$(0, 1)$ criterion)

수정된 $(0, 1)$까지최단거리기준(이하 수정된$(0, 1)$기준)은 $(0, 1)$점에서 ROC 곡선까지의 거리(반지름의 제곱)가 대각선까지의 거리(반지름의 제곱)에 대한 비율을 최소화하는 기준으로 다음과 같이 요약된다 (Perkins와 Schisterman, 2006).

$$\text{수정된}(0,1)\text{기준} = \min\sqrt{\frac{F_n(x)^2 + (1 - F_d(x))^2}{\left(\frac{F_n(x)}{1-F_d(x)+F_n(x)}\right)^2 + \left(\frac{1-F_d(x)}{1-F_d(x)+F_n(x)}\right)^2}} = \min\left\{1 - F_d(x) + F_n(x)\right\}.$$

$(0, 1)$기준과 Youden지수로 구한 최적분류점은 서로 다를 수 있으나, 수정된$(0, 1)$기준으로 구한 최적분류점은 Youden지수로 구한 최적분류점과 동일하다.

(5) SSS

Connell과 Koepsell (1985)이 제안한 SSS(sum of sensitivity and specificity 또는 gain in certainty) 측도는 정확도를 나타내는 민감도(sensitivity)와 특이도(specificity)의 합을 최대화하는 기준으로서 다음과 같이 정의한다.

$$\text{SSS} = \max\{F_d(x) + (1 - F_n(x))\}.$$

(6) 대칭점

대칭점(symmetry point)은 ROC 곡선에서 민감도와 특이도가 일치하는 점 또는 ROC 그림에서 $(0, 1)$점과 $(1, 0)$점을 지나는 직선(기울기가 -45°인 직선)과 ROC 곡선과의 교차점에 대응하는 스코어를 분류 기준으로 설정한다 (자세한 설명과 특징은 Moses 등 (1993)과 Pepe (2003) 참조).

$$\text{ROC}(u) = 1 - u.$$

이 식은 $F_d(x) = 1 - F_n(x)$ 또는 $F_d(x) + F_n(x) = 1$로 표현된다.

(7) 정확도 면적

R 프로그램의 DiagnosisMed 팩키지에 정확도를 계산하는 측도 중의 하나로 분류정확도를 나타내는 민감도와 특이도의 곱을 최대로 하는 정확도면적(accuracy area; AA)이란 측도는 다음과 같다 (Brasil, 2010).

$$\text{AA} = \max\{F_d(x)(1 - F_n(x))\}.$$

(8) 전체정확도(Total Accuracy; TA)

전체정확도 TA는 가장 많이 알려진 측도로 다음과 같이 정의한다 (Lambert와 Lipkovich, 2008).

$$\text{TA} = \max\left\{\frac{\text{TP} + \text{TN}}{p + n}\right\} = \max\left\{\gamma F_d(x) + (1 - \gamma)(1 - F_n(x))\right\}.$$

Finley (1884)가 기상예측분야에서 처음 발표한 측도로서 TA란 명칭 이외 efficiency (Greiner와 Gardner, 2000), index of validity (Feinstein, 2002), percent correctly classified (Freeman와 Moisen, 2008), overall accuracy (Liu 등, 2009) 등의 이름으로도 정의된다.

(9) 진실율(True Rate; TR)

전체정확도 TA는 민감도와 특이도의 가중평균으로 정의되지만, Velez 등 (2007)은 민감도와 특이도의 산술평균으로 balanced accuracy(BA) 그리고 홍종선 등 (2010)은 true positive rate와 true negative rate의 산술평균이란 의미로 진실율 TR로 정의하였다.

$$\mathrm{TR} = \max\left\{\frac{1}{2}\left(\frac{\mathrm{TP}}{p} + \frac{\mathrm{TN}}{n}\right)\right\} = \max\left\{\frac{1}{2}\left[F_d(x) + (1 - F_n(x))\right]\right\}.$$

홍종선 등 (2010)이 제안한 TR은 비용측면에서 TA보다 장점을 보였으며, 적은 전체부도율 γ가 작은 경우에 정확도가 과소평가되는 TA의 단점을 보완한다. 그리고 Cantor와 Kattan (2000)은 볼록한 ROC 곡선 아래 면적의 하한값이 TR임을 보였다.

3. 최적분류조건

2절에서 토론한 정확도 측도들의 성격과 관계를 정리1에서 설명한다.

정리 1. MVD, J, 수정된$(0,1)$기준, SSS, TR은 콜모고로프-스미르노프(Kolmogorov-Smirnov; KS) 통계량의 일차함수로 표현된다.

증명: Krzanowski와 Hand (2009)는 MVD와 J 통계량이 두 분포함수가 동일함을 검정하는 가설 (H_0 : $F_d(x) = F_n(x)$ *vs.* H_1 : $F_d(x) > F_n(x)$)에 대한 KS 검정통계량인 $\max\{F_d(x) - F_n(x)\}$와 일치함을 보였다. 즉 MVD, DPR 그리고 J는 KS 통계량과 일치하고, Perkins와 Schisterman (2006)이 제안한 수정된$(0,1)$기준은 J 통계량과 일차적인 관계를 설명하였다. 수정된$(0,1)$기준뿐만 아니라 SSS, TR 통계량은 KS 통계량과 다음과 같은 일차함수이다.

$$\begin{aligned}
\text{수정된}(0,1)\text{기준} &= \min\{1 - F_d(x) + F_n(x)\} = 1 - \mathrm{KS},\\
\mathrm{SSS} &= \max\{F_d(x) + (1 - F_n(x))\} = \mathrm{KS} + 1,\\
\mathrm{TR} &= \max\left\{\frac{1}{2}[F_d(x) + (1 - F_n(x))]\right\} = \frac{1}{2}(\mathrm{KS} - 1).
\end{aligned}$$

□

본 연구에서 토론한 정확도 통계량에 대응하는 최적분류점의 조건을 살펴보기 위하여 식 (1.1)의 모형에서 분포함수는 연속형이며 적절한 구간에서 미분 가능하다고 가정한다. 즉 $f_d(x)$, $f_n(x)$는 각각 $F_d(x)$, $F_n(x)$의 미분함수인 확률밀도함수이다. 최적분류점의 조건을 만족하는 국소 최대값 또는 최소값(local maximum or minimum)을 구하기 위해 2절에서 언급한 통계량을 미분한 조건함수식을 다음의 일곱 종류의 범주로 구분하여 살펴본다.

제1범주: $(0,1)$기준을 만족하는 분류점은 x에 대하여 미분하고 0으로 설정하여 구한다. 즉 $d\{(1 - F_d(x))^2 + F_n(x)^2\}/dx = 0$. 이 조건을 제1범주라고 정하고 다음과 같다.

$$\frac{f_d(x)}{f_n(x)} = \frac{F_n(x)}{1 - F_d(x)}. \tag{3.1}$$

표 2: 최적분류점에 대한 조건

범주	조건	측도
1	$(1-F_d(x_o))f_d(x_o) = F_n(x_o)f_n(x_o)$	$(0,1)$기준
2	$f_d(x_o) = f_n(x_o)$	MVD, J, 수정된$(0,1)$기준, SSS, TR
3	$1-F_d(x_o) = F_n(x_o)$	대칭점
4	$\gamma f_d(x_o) = (1-\gamma)f_n(x_o)$	TA
5	$(1-\gamma)(1-F_d(x_o)) = \gamma F_n(x_o)$	-
6	$(1-F_n(x_o))f_d(x_o) = F_d(x_o)f_n(x_o)$	AA
7	$\gamma F_d(x_o) = (1-\gamma)(1-F_n(x_o))$	-

제2범주: 제1범주의 조건식 (3.1)의 왼쪽 항을 1로 설정하면 다음과 같으며, 2절 정리 1에서 언급한 통계량인 MVD, J, 수정된$(0,1)$기준, SSS, TR을 미분하여 0으로 설정한 조건과 일치한다.

$$f_d(x) = f_n(x).$$

제3범주: 식 (3.1)의 오른쪽 항을 1로 설정하면 다음과 같으며, 2절에서 언급한 대칭점 통계량과 일치한다. 이 조건을 제3범주라 한다.

$$F_n(x) = 1 - F_d(x).$$

제4범주: 제1범주의 조건식 (3.1)의 왼쪽 항을 다음과 같이 설정하면, 2절에서 언급한 TA 통계량을 미분하여 0으로 설정한 조건과 일치한다. 이 조건을 제4범주라고 정한다.

$$\frac{f_d(x)}{f_n(x)} = \frac{1-\gamma}{\gamma}.$$

제5범주: 식 (3.1)의 오른쪽 항을 다음과 같은 조건으로 설정한다.

$$\frac{F_n(x)}{1-F_d(x)} = \frac{1-\gamma}{\gamma}.$$

제6범주: 2절에서 언급한 정확도 면적 AA 측도를 미분하여 0으로 설정한 조건은 다음과 같다.

$$\frac{f_d(x)}{f_n(x)} = \frac{F_d(x)}{1-F_n(x)}. \tag{3.2}$$

제7범주: 제6범주의 조건식 (3.2)의 왼쪽 항을 1과 $(1-\gamma)/\gamma$로 설정하면, 각각 제2범주와 제 4범주와 동일하다. 식 (3.2)의 오른쪽 항을 1로 설정하면 제3범주와 동일하고, $(1-\gamma)/\gamma$으로 설정하면 다음과 같다. 이 조건을 제7범주라 한다.

$$\frac{F_d(x)}{1-F_n(x)} = \frac{1-\gamma}{\gamma}.$$

일곱 종류의 범주와 조건식 그리고 이에 대응하는 측도들을 요약하면 표 2와 같다. 여기서 주목할 점은 2절에서 언급한 정확도 측도들은 모두 누적분포함수인 $F_d(x), F_n(x)$로 정의되었으며, 3절에서 논의한 조건함수식들은 누적분포함수뿐만 아니라 확률밀도함수인 $f_d(x), f_n(x)$로 표현된다.

표 3: 각 범주에 해당하는 오류율의 크기

범주	$\sigma_n^2 = 2,\ \gamma = 0.3$							
	$\mu_n = 1$				$\mu_n = 2$			
	분류점	α	β	$\alpha+\beta$	분류점	α	β	$\alpha+\beta$
1	0.5472	0.2921	0.3744	0.6665	0.9190	0.1790	0.2223	0.4041
2	0.8402	0.2004	0.4550	0.6554	1.0637	0.1437	0.2540	0.3977
3	0.4142	0.3394	0.3394	0.6787	0.8284	0.2037	0.2037	0.4074
4	-	-	-	-	0.4489	0.3268	0.1364	0.4631
5	0.8537	0.1966	0.4588	0.6554	1.1752	0.1200	0.2799	0.3998
6	0.6205	0.2675	0.3942	0.6617	0.9977	0.1592	0.2392	0.3985
7	-	-	-	-	-	-	-	-

4. 정규혼합분포

4.1. 평가기준

3절에서 논의한 일곱종류의 조건함수들에 대한 평가기준으로 제I종 오류율(α), 제II종 오류율(β) 그리고 오류율합($\alpha+\beta$)의 크기를 비교판단한다. 식 (1.1)에서 $F_d(x)$와 $F_n(x)$를 각각 정규분포의 누적분포함수로 설정하여 다음과 같이 가정하자.

$$F(x) = \gamma\Phi\left(x\,\middle|\,\mu_d, \sigma_d^2\right) + (1-\gamma)\Phi\left(x\,\middle|\,\mu_n, \sigma_n^2\right). \tag{4.1}$$

$F_d(x)$가 $\mu_d = 0$, $\sigma_d^2 = 1$인 표준정규분포 $\Phi(x|0,1)$와 $F_n(x)$이 $\mu_n = 1$과 2 그리고 $\sigma_n^2 = 2$인 정규분포 $\Phi(x|1,2)$와 $\Phi(x|2,2)$ 그리고 전체부도율 $\gamma = 0.3$인 경우에 각 범주에 해당하는 분류점과 이에 대응하는 제I종 오류율(α)과 II종 오류율(β) 그리고 오류율합($\alpha+\beta$)을 구하여 표 3에 구현하였다. 여기서 제I종 오류율과 II종 오류율은 임의의 분류점 x_o에 대하여 $\alpha = 1-\Phi(x_o|\mu_d, \sigma_d^2)$, $\beta = \Phi(x_o|\mu_n, \sigma_n^2)$로 정의한다. 정규혼합분포인 경우 제2범주와 4범주의 최적분류점은 홍종선 등 (2010)의 식 (4.3)과 (4.4) 그리고 (4.1)과 (4.2)에 의하여 구하였으며, 그외의 범주에서는 조건식을 만족하는 분류점이 두 모평균 사이인 구간 $[\mu_d, \mu_n]$에 존재하는 조건에서 수치계산 방법을 이용하여 구한다.

표 3에서 $\mu_n = 1$일 때 제4와 7범주인 경우 그리고 $\mu_n = 2$일 때 제7범주인 경우에는 적절한 분류점이 존재하지 않는다. 제I종 오류율이 최소일 때는 μ_n에 관계없이 제5범주인 경우이며, 제II종 오류율이 최소일 때는 $\mu_n = 1$에서는 제3범주 그리고 $\mu_n = 2$에서는 제4범주 경우이다. 오류율합이 최소일 때는 제2범주인 경우이나 $\mu_n = 1$에서는 제5범주의 오류율합도 최소값에 근사한다.

오류율의 크기가 최소인 경우가 일정하지 않으므로 다양한 정규혼합분포의 경우로 확대하여 살펴본다. $F_d(x)$는 표준정규분포 $\Phi(x|0,1)$이며 $F_n(x)$는 $\mu_n = 1, 2$와 $\sigma_n^2 = 0.5, 1, 1.5, 2, 2.5$를 따르는 정규분포 $\Phi(x|\mu_n, \sigma_n^2)$ 그리고 전체부도율 γ는 0.5 이하의 값의 정규혼합분포에서 제I종과 II종 오류율이 최소인 경우를 탐색한다. 제I종 오류율이 최소일 때는 제2범주인 경우이나 때로는 제5와 7범주인 경우이며, 제II종 오류율이 최소일 때는 제4범주인 경우이나 때로는 제3범주인 경우임을 알았다. 따라서 제I종과 II종 오류율이 최소인 경우의 조건함수식이 일정하지 않다는 것을 발견하였다.

4.2. 최적기준

정확도 측도들에 따라 최적분류점이 다르며 이에 대응하는 오류율의 크기가 혼합분포에 따라 일정하지 않아 어떤 경우가 최적인지 판단하기 위하여 그림 1과 2에 정확도 측도들의 성격에 따라 범주화한 조건함수식들 $f_d(x)/f_n(x)$, $F_d(x)/(1-F_n(x))$ 그리고 $F_n(x)/(1-F_d(x))$를 구현하였다. 만약 $f_d(x)/f_n(x) = 1$을 만족하는 스코어 x는 제2범주에 속하는 정확도 측도들에 대응하는 최적분류점을 의미하며 그림의 수평축에 '2'라고 나타낸다. $F_n(x)/(1-F_d(x)) = 1$을 만족하는 스코어는 제3범주에 해당하며 '3'으

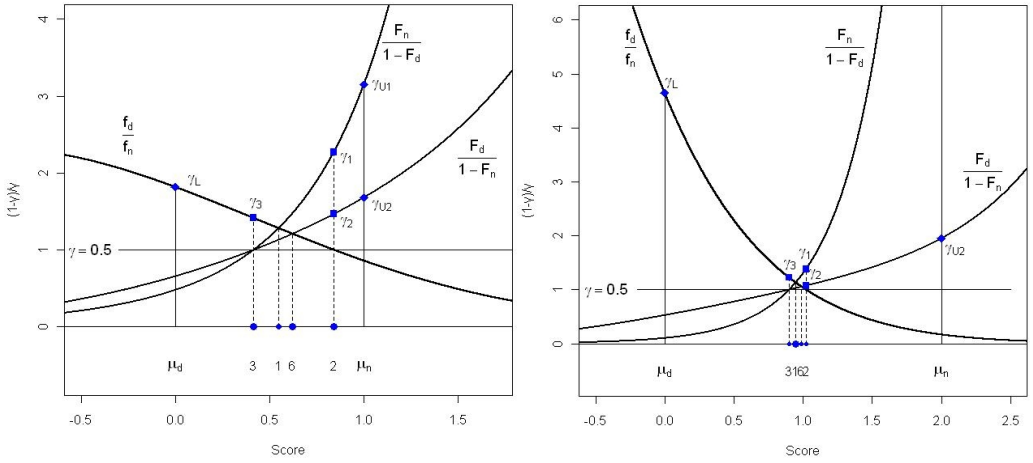

그림 1: *조건함수에서 분류점: d:* $N(0,1)$*과 n:* $N(1,2)$ *그리고 d:* $N(0,1)$*과 n:* $N(2,1.5)$*의 혼합분포*

로 표현한다. 그리고 $f_d(x)/f_n(x)$와 $F_n(x)/(1-F_d(x))$가 서로 교차하는 x에서는 제1범주에 속하는 측도의 최적분류점을 나타내며 '1'로 표현한다. $f_d(x)/f_n(x)$와 $F_d(x)/(1-F_n(x))$가 교차하는 스코어에서는 제6범주에 해당하며 '6'으로 나타낸다. 수직축은 $(1-\gamma)/\gamma$의 값으로 나타내는데 $\gamma = 0.5$이면 $(1-\gamma)/\gamma = 1$이므로 수직축의 값이 1일 때 참조선(reference line)으로 표현한다.

조건함수식에 속한 측도의 최적분류점에 대하여 다른 조건함수식을 만족하는 전체부도율을 정의하고 그림에 표기한다. 우선 분류점의 상한값인 μ_n에서 제5범주의 조건식 $(1-\gamma)(1-F_d(\mu_n)) = \gamma F_n(\mu_n)$을 만족하는 γ_{U1}과 제7범주의 조건식 $\gamma F_d(\mu_n) = (1-\gamma)(1-F_n(\mu_n))$을 만족하는 γ_{U2}는 다음과 같이 각각 정의하며,

$$\gamma_{U1} = \frac{1-F_d(\mu_n)}{F_n(\mu_n)+1-F_d(\mu_n)}, \qquad \gamma_{U2} = \frac{1-F_n(\mu_n)}{F_d(\mu_n)+1-F_n(\mu_n)},$$

분류점의 하한값인 μ_d에서 제4범주의 조건식 $\gamma f_d(\mu_d) = (1-\gamma) f_n(\mu_d)$을 만족하는 γ_L은 다음과 같이 설정한다.

$$\gamma_L = \frac{f_n(\mu_d)}{f_d(\mu_d)+f_n(\mu_d)}.$$

제2범주의 조건식 $f_d(x_o) = f_n(x_o)$이 성립하는 분류점 x_o에 대하여, $(1-\gamma)(1-F_d(x_o)) = \gamma F_n(x_o)$을 만족하는 γ_1과 $\gamma F_d(x_o) = (1-\gamma)(1-F_n(x_o))$을 만족하는 γ_2는 다음과 같이 각각 정의하며,

$$\gamma_1 = \frac{1-F_d(x_o)}{F_n(x_o)+1-F_d(x_o)}, \qquad \gamma_2 = \frac{1-F_n(x_o)}{F_d(x_o)+1-F_n(x_o)},$$

제3범주의 조건식 $1-F_d(x_o) = F_n(x_o)$이 성립하는 x_o에 대하여, $\gamma f_d(x_o) = (1-\gamma) f_n(x_o)$을 만족하는 γ_3는 다음과 같이 설정한다.

$$\gamma_3 = \frac{f_n(x_o)}{f_d(x_o)+f_n(x_o)}.$$

그림 1의 왼쪽 그래프는 혼합분포에서 조건함수식과 $(1-\gamma)/\gamma$와의 관계에 따라 분류점이 $\gamma_{U1} < \gamma_1 < \gamma_{U2} < \gamma_2$ 경우이며, 오른쪽 그래프는 $\gamma_{U1} < \gamma_{U2} < \gamma_1 < \gamma_2$ 경우이다. 그림 1의 왼쪽 그래프

표 4: $\gamma_{U1}, \gamma_{U2}, \gamma_L, \gamma_1, \gamma_2, \gamma_3$의 값

σ_n^2	μ_n=1						μ_n=2					
	γ_{U1}	γ_1	γ_{U2}	γ_2	γ_L	γ_3	γ_{U1}	γ_1	γ_{U2}	γ_2	γ_L	γ_3
1	0.2409	0.5000	0.3728	0.5000	0.3775	0.5000	0.0435	0.5000	0.3385	0.5000	0.1192	0.5000
1.5		0.3817		0.4437	0.3691	0.4495		0.4198		0.4820	0.1771	0.4495
2		0.3058		0.4053	0.3551	0.4142		0.3614		0.4656	0.2064	0.4142
2.5		0.2569		0.3807	0.3412	0.3874		0.3175		0.4514	0.2213	0.3874

표 5: γ 범위와 최적기준

γ의 범위		α 최소범주
$\gamma_{U1} < \gamma_1 < \gamma_{U2} < \gamma_2$	$\gamma_{U1} \le \gamma < \gamma_1$	5
	$\gamma_{U2} \le \gamma < \gamma_2$	7
	그외	2
$\gamma_{U1} < \gamma_{U2} < \gamma_1 < \gamma_2$	$\gamma_{U1} \le \gamma < \gamma_{U2}$	5
	$\gamma_{U2} \le \gamma < \gamma_2$	7
	그외	2
γ의 범위		β 최소범주
$\gamma_L \le \gamma < \gamma_3$		4
그외		3

는 d: $N(0,1)$과 n: $N(1,2)$의 혼합분포를 표현한 그래프이다. 제I종 오류율이 최소인 경우를 살펴보면, $\gamma_{U1}(= 0.2409) \le \gamma < \gamma_1(= 0.3058)$에서는 $(1-\gamma)(1-F_d(x_o)) = \gamma F_n(x_o)$을 만족하므로 제5범주이며, $\gamma = \gamma_1(= 0.3058)$일 때는 전체부도율 γ에 의존하지 않는 $f_d(x_o) = f_n(x_o)$의 조건을 만족하는 제2범주에 대응하는 분류점을 만나고 전체부도율 γ가 구간 (γ_1, γ_{U2})에 속할 때 제2범주가 최소의 제I종 오류율을 나타낸다. 그리고 $\gamma \ge \gamma_{U2}(= 0.3728)$에서는 $\gamma F_d(x_o) = (1-\gamma)(1-F_n(x_o))$의 조건을 만족하는 제7범주에 대응하는 분류점이 존재하기 시작하여 $\gamma_{U2}(= 0.3728) \le \gamma < \gamma_2(= 0.4053)$에서는 제7범주가 최소의 제I종 오류율을 나타내고 $\gamma \ge \gamma_2(= 0.4053)$에서는 다시 $f_d(x_o) = f_n(x_o)$의 조건을 만족하는 제2범주가 제I종 오류율이 최소임을 나타낸다. 다음으로 제II종 오류율이 최소인 경우를 살펴보면, $\gamma_L(= 0.3551) \le \gamma < \gamma_3(= 0.4142)$에서는 $\gamma f_d(x_o) = (1-\gamma) f_n(x_o)$의 조건을 만족하는 제4범주이며, $\gamma = \gamma_3(= 0.4142)$일 때는 전체부도율에 의존하지 않는 $1 - F_d(x_o) = F_n(x_o)$의 조건을 만족하는 제3범주에 대응하는 분류점을 만난다. 따라서 $\gamma \ge \gamma_3(= 0.4142)$에서는 제3범주가 최소의 제II종 오류율을 나타낸다 (표 4와 5 참조).

그림 1의 오른쪽 그래프는 d: $N(0,1)$과 n: $N(2,1.5)$의 혼합분포인 경우이며, $\gamma_{U1}(= 0.0435) \le \gamma < \gamma_{U2}(= 0.3385)$에서는 $(1-\gamma)(1-F_d(x_o)) = \gamma F_n(x_o)$조건을 만족하므로 제5범주이며, $\gamma_{U2}(= 0.3385) \le \gamma < \gamma_2(= 0.4820)$에서는 $\gamma F_d(x_o) = (1-\gamma)(1-F_n(x_o))$조건을 만족하기 때문에 제7범주가 제I종 오류율이 최소임을 나타낸다. 여기서 γ_1은 구간 (γ_{U2}, γ_2)에 존재한다. 그리고 $\gamma = \gamma_2(= 0.4820)$일 때는 전체부도율에 의존하지 않는 $f_d(x_o) = f_n(x_o)$의 조건을 만족하는 제2범주에 대응하는 분류점을 만나서 $\gamma \ge \gamma_2(= 0.4820)$에서는 제2범주가 제I종 오류율이 최소임을 나타낸다. 다음으로 제II종 오류율이 최소인 경우를 살펴보면, $\gamma_L(= 0.1771) \le \gamma < \gamma_3(= 0.4495)$에서는 $\gamma f_d(x_o) = (1-\gamma) f_n(x_o)$의 조건을 만족하는 제4범주이며, $\gamma \ge \gamma_3(= 0.4495)$에서는 전체부도율에 의존하지 않는 $1 - F_d(x_o) = F_n(x_o)$의 조건을 만족하는 제3범주가 최소의 제II종 오류율을 나타낸다. 이에 대하여는 표 4와 5에서 자세히 설명된다.

표준정규분포 $\Phi(x|0,1)$과 $\mu_n = 1, 2$와 $\sigma_n^2 = 1, 1.5, 2, 2.5$를 따르는 정규분포 $\Phi(x|\mu_n, \sigma_n^2)$ 그리고 전체부도율 γ은 0.5 이하의 값의 정규혼합분포에서 $\gamma_{U1}, \gamma_{U2}, \gamma_L, \gamma_1, \gamma_2, \gamma_3$의 값을 표 4에 구하였다. 그리고 표 4를 바탕으로 전체부도율의 크기가 어떤 범위에서 α 또는 β가 최소일 때에 해당하는 조건함수의 범

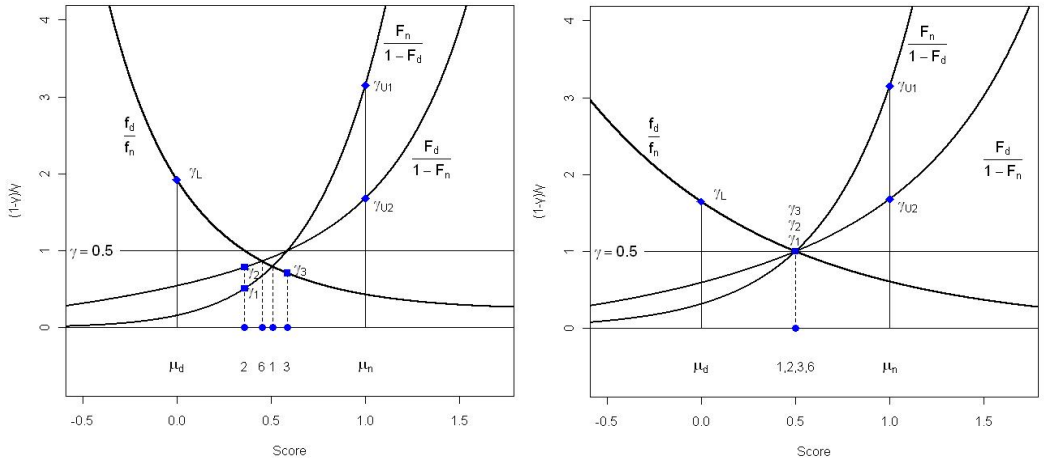

그림 2: *조건함수에서 분류점: d: $N(0,1)$과 n: $N(1,1.5)$ 그리고 d: $N(0,1)$과 n: $N(1,1)$의 혼합분포*

주가 무엇인지를 전체적으로 탐색한 결과를 표 5에 구현하였다.

표 5를 바탕으로 제I종 오류율과 제II종 오류율 그리고 오류율합이 최소일 때의 상황을 다음과 같이 정리할 수 있다.

- 제I종 오류율이 최소인 경우를 살펴보자. $\gamma_{U1} < \gamma_1 < \gamma_{U2} < \gamma_2$와 $\gamma_{U1} < \gamma_{U2} < \gamma_1 < \gamma_2$인 두가지 경우로 구분하여 설명할 수 있으나 두 경우 모두 $\gamma_{U1} \le \gamma < \min(\gamma_1, \gamma_{U2})$에서 제I종 오류율이 최소일 때는 제5범주이며, $\gamma_{U2} \le \gamma < \gamma_2$에서는 제7범주가 최소의 제I종 오류율을 나타낸다. 두가지 범주에 속하지 않는 전체부도율의 구간에서는 제2범주에 속하는 MVD, J, 수정된$(0,1)$기준, SSS, TR 기준의 제I종 오류율이 최소임을 나타낸다. 전체부도율 γ가 특정한 구간에서는 제5와 7범주에 대한 조건에서 제I종 오류율이 최소임에도 불구하고 제5와 7범주에 속하는 정확도 측도는 존재하지 않는다. 그러므로 2절에서 고려한 정확도 측도들 중에서 제I종 오류율이 최소일 때의 측도는 MVD, J, 수정된$(0,1)$기준, SSS, TR 측도이다.

- 제II종 오류율이 최소인 경우를 살펴보자. $\gamma_L \le \gamma < \gamma_3$에서는 제4범주가 최소의 제II종 오류율을 나타내고, $\gamma = \gamma_3$일 때는 제3범주에 대응하는 분류점이 존재하므로 $\gamma \ge \gamma_3$에서는 제3범주에 속하는 대칭점 통계량의 제II종 오류율이 최소이다.

- 제I종과 제II종 오류율합이 최소인 경우는 전체부도율에 의존하는 어떠한 정규혼합분포의 종류에도 관계없이 제2범주에 속하는 MVD, J, 수정된$(0,1)$기준, SSS, TR 기준의 오류율합이 최소이다.

표본수가 많은 정상 차주의 분산은 표본수가 적은 부도 차주의 분산보다 일반적으로 크기 때문에 $\sigma_n^2 > \sigma_d^2$인 경우만을 살펴보았다. 만약 정상차주의 분산이 부도차주보다 작은 즉 $\sigma_n^2 < \sigma_d^2$인 경우 예를 들어 $\sigma_n^2 = 0.5$일 때 그림 1과 유사한 조건함수와 분류점을 작성한 그림 2의 왼쪽을 살펴보자. 그림 1에서 제3범주에 대응하는 분류점이 제일 작은 스코어에서 등장하고 제2범주에 대응하는 분류점이 마지막으로 등장한다. 그러나 그림 2의 왼쪽에서는 스코어가 증가할수록 분류점이 역순으로 등장하는 것을 파악할 수 있다. 앞에서 정의한 $\gamma_1, \gamma_2, \gamma_3$의 값이 모두 0.5이상이며 $\gamma_{U1} < \gamma_{U2} < 0.5 < \gamma_2 < \gamma_1$과 $\gamma_L < 0.5 < \gamma_3$의 관계를 갖는다. 그러므로 $\sigma_n^2 < \sigma_d^2$인 경우에 전체부도율 γ값이 구간 $(\gamma_{U1}, \gamma_{U2})$에 존재하면 제5범주, 구간 $(\gamma_{U2}, 0.5)$에 존재하면 제7범주 그리고 $\gamma \ge 0.5$이면 제3범주에 대응하는 분류점의

제I종 오류율이 최소이다. 또한 γ값이 구간 $(\gamma_L, 0.5)$에 존재하면 제4범주 그리고 $\gamma \geq 0.5$이면 제2범주에 대응하는 분류점의 제II종 오류율이 최소임을 파악할 수 있다.

정상차주와 부도차주의 분산이 동일한 $\sigma_n^2 = \sigma_d^2$인 경우에는 그림 2의 오른쪽과 같이 나타나는데 전체부도율 γ값과 독립적인 제1, 2, 3, 6범주에 속하는 정확도 측도에 대응하는 분류점이 모두 동일하다. 그러므로 두 분포함수의 모분산이 동일하면, 제4범주에 속하는 TA 측도의 제II종 오류율이 최소이며 제I종 오류율이 최소일 때의 정확도 측도는 제1, 2, 3, 6범주에 속하는 MVD, J, $(0, 1)$기준, 수정된$(0, 1)$기준, SSS, TR, AA, 대칭점이며 TA 측도만 제외한 측도들이다.

5. 결론

본 연구는 두 분포함수의 혼합분포에서 분류정확도를 측정하는 아홉 종류의 측도들의 성격을 바탕으로 측도들 중에서 MVD, J, 수정된$(0, 1)$기준, SSS, TR은 콜모고로프-스미르노프 통계량과 일차함수 관계로 나타남을 발견하였다.

토론한 아홉 종류의 정확도 측도를 일곱 범주의 조건함수식으로 범주화하였다. 그리고 범주화된 여러 측도들에 기반하는 분류점들을 구하고, 그 분류점에 대응하는 제I종 오류율과 제II종 오류율 그리고 오류율합을 구하여 크기를 비교하고자 정규혼합분포를 가정하여 다양한 정규혼합분포에 대하여 분류정확도의 조건함수와 전체부도율과의 관계를 탐색하였다. 즉 주어진 전체부도율에 의존하는 특정한 정규혼합분포의 경우에는 제I종 오류율이 최소일 때의 정확도 측도는 무엇이며 어떠한 정확도 측도가 최소의 제II종 오류율을 나타내는 지를 파악할 수 있으며 결과를 표 4와 5에 구현하였다.

표 4와 5로부터 유도한 결론은 다음과 같다. 우선 본 연구에서 고려한 정확도 측도들 중에서 제2범주에 속하는 MVD, J, 수정된$(0, 1)$기준, SSS, TR 통계량의 제I종 오류율이 최소임을 나타낸다. 제II종 오류율에 대하여는 전체부도율이 적은 값에서는 제4범주에 속하는 TA 통계량의 제II종 오류율이 최소이며, 전체부도율이 0.5에 가까운 값을 가질 때에는 제3범주에 속하는 대칭점 통계량의 제II종 오류율이 최소이다. 그리고 제I종과 제II종 오류율합이 최소인 경우는 전체부도율에 의존하는 어떠한 혼합분포에도 제2범주에 속하는 MVD, J, 수정된$(0, 1)$기준, SSS, TR 통계량의 오류율합이 최소임을 보여준다. 표본수가 많은 정상 차주의 분산은 부도 차주의 분산보다 일반적으로 크기 때문에 $\sigma_n^2 > \sigma_d^2$를 가정하였으나, $\sigma_n^2 < \sigma_d^2$ 그리고 γ가 0.5 이상인 경우에 제I종 오류율이 최소일 때에 해당되는 범주는 제3범주에 속하는 대칭점 통계량이 되며, 제II종 오류율이 최소일 때에 해당되는 범주는 제2범주에 속하는 MVD, J, 수정된$(0, 1)$, SSS, TR 기준 통계량으로 서로 교환되며 그외는 변함이 없음을 탐색하였다. 또한 $\sigma_n^2 = \sigma_d^2$ 인 경우에 제I종 오류율이 최소일 때는 제2범주에 속하는 측도들이며, 제II종 오류율이 최소일 때는 제3범주에 속하는 측도인 대칭점 측도이다.

그러므로 본 연구에서 얻은 결론을 이용하여 자료의 상황에 따라 어떤 분류정확도 측도를 사용할 것인가에 대하여는 다음과 같이 정리할 수 있다. 일반적으로 표본수가 많은 정상 차주의 분산은 부도 차주의 분산보다 크며, 이 경우에 제I종 오류율을 최소화하길 원하면 제2범주에 속하는 MVD, J, 수정된$(0, 1)$기준, SSS, TR 통계량을 사용하고, 제II종 오류율을 최소화하길 원하면 전체부도율이 적을 때에는 제4범주에 속하는 TA 통계량 그리고 전체부도율이 0.5에 가까운 값을 가질 때에는 제3범주에 속하는 대칭점 통계량을 사용한다. 제I종과 제II종 오류율합이 최소화하길 원하면 제2범주에 속하는 MVD, J, 수정된$(0, 1)$기준, SSS, TR 통계량을 사용하는 것을 제안한다. 다음으로 정상 차주의 분산이 부도 차주의 분산보다 작으며 전체부도율 가 0.5이상인 경우에, 제I종 오류율을 최소화하길 원하면 제3범주에 속하는 대칭점 통계량 그리고 제II종 오류율을 최소화하길 원하면 제2범주에 속하는 MVD, J, 수정된$(0, 1)$, SSS, TR 통계량을 사용하며 그외는 변함이 없다. 또한 정상 차주의 분산과 부도 차주의 분산이 동일한 경우에, 제I종 오류율을 최소화하길 원하면 제2범주에 속하는 측도들 그리고 제II종

오류율을 최소화하길 원하면 제3범주에 속하는 측도인 대칭점 측도를 사용한다.

혼합분포의 최적분류점은 정확도 측도들에 따라 다르기 때문에 분류점에 대응하는 오류율의 크기가 일정하지 않아 어떤 경우에 최적인지 판단하기 어렵다. 따라서 분석할 자료에 적합한 분포와 전체부도율을 계산해서 혼합분포를 추정한 다음 본 연구를 활용하면, 어떤 분류정확도 측도를 사용하여 제I종과 II종 오류율 또는 오류율합이 최소인지를 탐색할 수 있으며 자주 이용하는 정확도 측도의 장점과 단점을 파악할 수 있다.

본 연구에서는 스코어 확률변수의 누적분포함수를 구성하는 부도와 정상상태의 누적분포함수 각각을 가장 통계적으로 일반적인 정규분포로 가정하였다. 그리고 정규분포 이외의 분포를 가정한 연구는 충분히 가치있는 연구가 되기에 이를 향후 연구과제로 남겨두기로 한다.

참고 문헌

홍종선, 주재선, 최진수 (2010). 혼합분포에서의 최적분류점, <응용통계연구>, **23**, 13–28.

홍종선, 최진수 (2009). ROC와 CAP 곡선에서의 최적분류점, <응용통계연구>, **22**, 911–921.

Brasil, P. (2010). Diagnostic test accuracy evaluation for medical professionals, package *DiagnosisMed* in R.

Cantor, S. B. and Kattan, M. W. (2000). Determining the area under the ROC curve for a binary diagnostic test, *Medical Decision Making*, **20**, 468–470.

Cantor, S. B., Sun, C. C., Tortolero-Luna, G., Richards-Kortum, R. and Follen, M. (1999). A comparison of C/B ratios from studies using receiver operating characteristic curve analysis, *Journal of Clinical Epidemiology*, **52**, 885–892.

Connell, F. A. and Koepsell, T. D. (1985). Measures of gain in certainty from a diagnostic test, *American Journal of Epidemiology*, **121**, 744–753.

Engelmann, B., Hayden, E. and Tasche, D. (2003). Measuring the discriminative power of rating systems, series 2: Banking and Financial Supervision, **01**.

Fawcett, T. (2003). ROC graphs: notes and practical considerations for data mining researchers, *HP Laboratories*, 1501 Page Mill Road, Palo Alto, CA 94304.

Feinstein, A. R. (2002). *Principles of Medical Statistics*, Chapman & Hall/CRC, Boca Raton, FL.

Finley, J. P. (1884). Tornado predictions, *American Meteorological Journal*, **1**, 85–88.

Freeman, E. A. and Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa, *Ecological Modelling*, **217**, 48–58.

Greiner, M. M. and Gardner, I. A. (2000). Epidemiologic issues in the validation of veterinary diagnostic tests, *Preventive Veterinary Medicine*, **45**, 3–22.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*, Champman & Hall/CRC, Boca Raton, FL.

Lambert, J. and Lipkovich, I. (2008). A macro for getting more out of your ROC curve, *SAS Global Forum*, **231**.

Liu, C., White, M. and Newell1, G. (2009). Measuring the accuracy of species distribution models: a review, *18th World IMACS/MODSIM Congress*, http://mssanz.org.au/modsim09.

Moses, L. E., Shapiro, D. and Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations, *Statistics in Medicine*, **12**, 1293–1316.

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

Perkins, N. J. and Schisterman, E. F. (2006). The inconsistency of “optimal” cutpoints obtained using two criteria based on the receiver operating characteristic curve, *American Journal of Epidemiology*, **163**,

670–675.

Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments, *Machine Learning*, **42**, 203–231.

Sobehart, J. R. and Keenan, S. C. (2001). Measuring default accurately, Credit Risk Special Report, *Risk*, **14**, 31–33.

Tasche, D. (2006). Validation of internal rating systems and PD estimates, *arXiv.org, eprint arXiv: physics/0606071*.

Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M. and Moore, J. H. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction, *Genetic Epidemiology*, **31**, 306–315.

Ward, C. D. (1986). The differential positive rate, a derivative of receiver operating characteristic curves useful in comparing tests and determining decision levels, *Clinical Chemistry*, **32**, 1428–1429.

Youden, W. J. (1950). Index for rating diagnostic test, *Cancer*, **3**, 32–35.

Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*, Wiley, New York.

2011년 1월 접수; 2011년 3월 채택

Optimal Criterion of Classification Accuracy Measures for Normal Mixture

Hyun Sang Yoo[a], Chong Sun Hong[1,b]

[a]Research Institute of Applied Statistics, Sungkyunkwan University
[b]Department of Statistics, Sungkyunkwan University

Abstract

For a data with the assumption of the mixture distribution, it is important to find an appropriate threshold and evaluate its performance. The relationship is found of well-known nine classification accuracy measures such as MVD, Youden's index, the closest-to-(0, 1) criterion, the amended closest-to-(0, 1) criterion, SSS, symmetry point, accuracy area, TA, TR. Then some conditions of these measures are categorized into seven groups. Under the normal mixture assumption, we calculate thresholds based on these measures and obtain the corresponding type I and II errors. We could explore that which classification measure has minimum type I and II errors for estimated mixture distribution to understand the strength and weakness of these classification measures.

Keywords: Accuracy, classification, discrimination, error, sensitivity, specificity.

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, 53, Myeongnyun-dong 3-Ga, Jongro-Gu, Seoul 110-745, Korea. E-mail: cshong@skku.ac.kr