한국통계학회논문집
2012, 19권, 4호, 629–638

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.4.629

Bivariate ROC Curve and Optimal Classification Function

C. S. Hong[1,a], J. A. Jeong[b]

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

Abstract

We propose some methods to obtain optimal thresholds and classification functions by using various cut-off criterion based on the bivariate ROC curve that represents bivariate cumulative distribution functions. The false positive rate and false negative rate are calculated with these classification functions for bivariate normal distributions.

Keywords: Credit evaluation, default, classification, cutoff criteria, misclassification rate.

1. 서론

일반적인 ROC(Receiver Operation Characteristic) 곡선은 일변량 확률변수에 대한 누적분포함수로 구성되는데, Hong 등 (2012)은 이변량으로 확장하여 이변량 결합누적분포함수로 표현되는 ROC 곡선을 제안하였다. 신호탐지이론으로 출발한 ROC 곡선은 다양한 학분분야에서의 의사결정과 진단의 체계에서 폭넓게 연구되고 있으며, 본 연구에서는 신용평가 관점에서 논의하며 다음을 가정한다.

차주(borrower)는 이차원 스코어(score) 확률변수 X_1, X_2와 모수 공간 Θ에 의해서 특성을 나타낸다고 가정한다. 스코어 확률변수는 대출기관에서 차주의 신용가치를 예상하기 위해 차주에게 부여한 연속형 값을 갖는다. 대출기관은 차주의 신용가치에 관한 정보에 의한 스코어 변수를 통하여 차주의 미래상태 Θ를 예상한다. 차주의 모집단은 두 개의 부모집단 $\Theta = \{\theta_d, \theta_n\}$으로 구성되며 미래시점에 대출상환능력이 없는 부도(default; d)와 대출상환능력이 있는 정상(non-default; n)으로 구분된다. Θ가 θ_d일 때($\Theta = \theta_d$) 부도차주의 모집단에 속하고, Θ가 θ_n일 때($\Theta = \theta_n$) 정상차주의 모집단에 속한다. 이변량 스코어 확률변수 X_1, X_2의 결합누적분포함수 $F(x_1, x_2)$는 차주의 부도와 정상상태 하에서 스코어의 조건부 결합누적분포함수 $F_d(x_1, x_2)$와 $F_n(x_1, x_2)$의 선형 결합으로 다음과 같이 가정한다.

$$F(x_1, x_2) = \lambda F_d(x_1, x_2) + (1 - \lambda)F_n(x_1, x_2), \tag{1.1}$$

여기서 $F_d(x_1, x_2) = P(X_1 \le x_1, X_2 \le x_2|\theta_d)$, $F_n(x_1, x_2) = P(X_1 \le x_1, X_2 \le x_2|\theta_n)$ 그리고 λ는 전체부도율(total probability of default) $P(\Theta = \theta_d)$이다.

차주의 부도와 정상상태 하에서 이변량 확률변수 X_1, X_2의 모평균벡터를 각각 $(\mu_1\ \mu_2)'$와 $(\mu_3\ \mu_4)'$라고 하면, 기울기는 두 모평균벡터를 지나가는 $b = (\mu_4 - \mu_2)/(\mu_3 - \mu_1)$이며 (μ_1, μ_2)을 통과하는 직선 $X_2 = b(X_1 - \mu_1) + \mu_2$를 설정하여 X_2를 X_1의 일차함수 $h(x_1)$로 표현하면, 임의의 값 $y_n = F_n(x_1, x_2)$, $y_d = F_d(x_1, x_2)$에 대하여 유일하게 존재하는 $(x_1, h(x_1))$을 선정할 수 있으므로 (y_n, y_d)에 대응하는 $F_n(x_1, h(x_1))$, $F_d(x_1, h(x_1))$을 이용하여 이변량 ROC 곡선을 구현하였다.

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongro-gu, Seoul 110-745, Korea. E-mail: cshong@skku.edu

Table 1: μ_3 and μ_4 with slope

Slope	1/3	2/3	1	1.5	2
(μ_3, μ_4)	$(\sqrt{18/5}, \sqrt{2/5})$	$(\sqrt{36/13}, \sqrt{16/13})$	$\left(\sqrt{2}, \sqrt{2}\right)$	$(\sqrt{16/13}, \sqrt{36/13})$	$(\sqrt{4/5}, \sqrt{16/5})$

Hong 등 (2012)는 ROC 곡선에 대하여 많이 사용하는 최적분류기준으로 Kolmogorov-Smirnov 통계량과 동일한 MVD (Krzanowski와 Hand, 2009), Youden지수 (Youden, 1950), 수정된$(0, 1)$기준 (Perkins와 Schisterman, 2006), SSS (Connell과 Koepsell, 1985), True Rate (Lambert와 Lipkovich, 2008) 등을 이용하여 최적분류점(optimal cutoff, threshold)을 구하고 이 점을 통과하는 최적분류함수(optimal classification function)를 제안하였다. 특히 이변량 확률변수 X_1, X_2의 모분산벡터를 모두 동일한 1로 설정하면 최적분류점은 두 모평균벡터의 중간점이며 이 점을 지나면서 기울기가 $-1/b$인 직선을 최적분류함수로 제안하였다.

Hong 등 (2012)의 식 (2.3) 또는 본 연구의 식 (2.1)에서 설정한 차주의 부도와 정상상태의 이변량 정규분포함수에 대한 이변량 ROC 곡선을 바탕으로 최적분류기준에 의하여 최적분류점을 구한다. 최적분류점은 두 모평균 벡터의 중간점이 아닌 사실과, 구한 최적분류점을 통과하는 최적분류함수(optimal classification function)가 직선이 아닌 곡선 형태로 표현됨을 증명하고자 한다. 이러한 최적분류함수에 기인하여 실제 정상을 부도로 잘못 예측하는 비율(FPR; false positive rate 또는 false alram rate)과 실제 부도를 정상으로 예측하는 비율(FNR; false negative rate)를 구하고 분석한다.

본 연구의 구성은 다음과 같다. 2절에서는 차주의 부도와 정상상태의 확률분포함수를 ROC 곡선에 관한 연구에서 많이 사용하는 정규분포를 가정하고자 하며, 이변량 분포함수 중에서 상관관계를 가장 잘 표현하는 이변량 정규분포를 가정한다. 따라서 차주의 부도와 정상상태의 확률분포함수를 다양한 모평균벡터, 모분산벡터 그리고 모상관계수를 나타낼 수 있는 이변량 정규분포로 가정하고, Hong 등 (2012)이 제안한 이변량 ROC 곡선을 구현하고 이 곡선으로부터 많이 사용하는 최적분류기준을 사용하여 최적분류점을 구한다. 3절에서는 이변량 확률밀도함수로 정의된 차주의 부도와 정상상태를 판별할 수 있는 분류함수를 제안하고 이 분류함수와 2절에서 얻은 최적분류점과의 관계를 탐색한다. 4절에서는 본 연구에서 제안한 최적분류함수를 이용하여 다양한 이변량 정규분포인 경우에 FPR과 FNR을 구하고 모분산이 동일한 경우와 비교 분석한다. 마지막 5절에서는 다양한 이변량 분포의 성과를 표현하는 이변량 ROC 곡선으로부터 유도할 수 있는 최적분류점과 이를 바탕으로 구한 최적분류함수에 대하여 요약 정리하면서 결론을 유도한다.

2. 이변량 ROC곡선에서 최적분류점

식 (1.1)에서의 이변량 확률변수 X_1, X_2의 결합누적분포함수는 $F_d(x_1, x_2)$와 $F_n(x_1, x_2)$로 표현되는데, 본 연구에서는 $F_d(x_1, x_2)$와 $F_n(x_1, x_2)$을 따르는 확률변수를 각각 (X_{1d}, X_{2d})와 (X_{1n}, X_{2n})이라고 정의한다. 부도상태 하의 확률변수 (X_{1d}, X_{2d})는 상관계수 ρ를 갖는 표준정규분포를 따르며 즉, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ 그리고 정상상태의 (X_{1n}, X_{2n})는 다음과 같은 이변량 정규분포를 따른다고 가정한다.

$$\begin{pmatrix} X_{1d} \\ X_{2d} \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \qquad \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix} = N\left(\begin{pmatrix} \mu_3 \\ \mu_4 \end{pmatrix}, \begin{pmatrix} \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho\sigma_3\sigma_4 & \sigma_4^2 \end{pmatrix}\right). \tag{2.1}$$

Hong 등 (2012)이 제안한 이변량 ROC 곡선을 구하고 이 곡선으로부터 구한 최적분류점을 비교하여 살펴본 다음 다양한 분산과 상관계수인 경우로 확장한다. 부도상태 하의 분산이 1이고 정상상태의 분산은 1보다 작거나 큰 분산의 경우를 살펴보기 위하여 두 모평균벡터의 거리를 2로 고정한다. 따라서 다양한 기울기 $b = \mu_4/\mu_3$에 따라 (μ_3, μ_4)의 좌표는 Table 1과 같이 설정하였다.

Table 2: Threshold with $\sigma_N = (\sigma_3, \sigma_4)$

Slope	ρ	$(\sigma_3, \sigma_4) = (1, 1)$	$(\sigma_3, \sigma_4) = (0.75, 0.75)$	$(\sigma_3, \sigma_4) = (1.25, 1.25)$
1/3	−0.8	(0.95, 0.32)	(1.05, 0.35)	(0.85, 0.28)
	−0.6	(0.95, 0.32)	(1.05, 0.35)	(0.90, 0.30)
	−0.4	(0.95, 0.32)	(1.00, 0.33)	(0.95, 0.32)
	−0.2	(0.95, 0.32)	(1.00, 0.33)	(0.95, 0.32)
2/3	−0.8	(0.83, 0.56)	(0.95, 0.63)	(0.75, 0.50)
	−0.6	(0.83, 0.56)	(0.90, 0.60)	(0.80, 0.53)
	−0.4	(0.83, 0.56)	(0.90, 0.60)	(0.80, 0.53)
	−0.2	(0.83, 0.56)	(0.90, 0.60)	(0.85, 0.57)
	0	(0.83, 0.56)	(0.85, 0.57)	(0.85, 0.57)
1	−0.8	(0.70, 0.70)	(0.80, 0.80)	(0.65, 0.65)
	−0.6	(0.70, 0.70)	(0.80, 0.80)	(0.65, 0.65)
	−0.4	(0.70, 0.70)	(0.75, 0.75)	(0.70, 0.70)
	−0.2	(0.70, 0.70)	(0.75, 0.75)	(0.70, 0.70)
	0	(0.70, 0.70)	(0.75, 0.75)	(0.75, 0.75)
1.5	−0.8	(0.56, 0.83)	(0.60, 0.90)	(0.50, 0.75)
	−0.6	(0.56, 0.83)	(0.60, 0.90)	(0.50, 0.75)
	−0.4	(0.56, 0.83)	(0.60, 0.90)	(0.55, 0.83)
	−0.2	(0.56, 0.83)	(0.60, 0.90)	(0.55, 0.83)
	0	(0.56, 0.83)	(0.55, 0.83)	(0.55, 0.83)
	0.2	(0.56, 0.83)	(0.55, 0.83)	(0.60, 0.90)
2	−0.8	(0.45, 0.89)	(0.50, 1.00)	(0.40, 0.80)
	−0.6	(0.45, 0.89)	(0.50, 1.00)	(0.40, 0.80)
	−0.4	(0.45, 0.89)	(0.50, 1.00)	(0.45, 0.90)
	−0.2	(0.45, 0.89)	(0.45, 0.89)	(0.45, 0.90)
	0	(0.45, 0.89)	(0.45, 0.89)	(0.45, 0.90)
	0.2	(0.45, 0.89)	(0.45, 0.89)	(0.45, 0.90)
	0.4	(0.45, 0.89)	(0.45, 0.89)	(0.45, 0.90)

본 연구에서는 다양한 최적분류기준 중에서 가장 보편적이며 많이 사용하는 MVD (Krzanowski와 Hand, 2009), Youden지수 (Youden, 1950), 수정된$(0, 1)$기준 (Perkins와 Schisterman, 2006), SSS (Connell과 Koepsell, 1985), True Rate (Lambert와 Lipkovich, 2008; Hong과 Joo, 2010)의 기준을 바탕으로 구한 최적분류점(optimal cutoff, threshold)을 구하고자 한다.

정상상태 하의 분포에서 분산이 부도상태와 동일한 ($\sigma_3 = \sigma_4 = 1$) 연구에서 최적분류점은 두 모평균벡터의 중간지점이다. 부도와 정상상태의 분산이 다른 경우에는 최적분류점이 두 모평균벡터의 중점이 아니고 이론적으로 구하기 어려우므로, 다양한 분산 σ_3, σ_4와 상관계수 ρ인 경우의 이변량 ROC 곡선으로부터 1절에서 언급한 최적분류기준에 의하여 최적분류점을 구하고 Table 2에 나열하였다. Table 2에서의 기울기와 상관계수는 Table 3의 경우와 일치시키기 위하여 Table 3의 기울기과 상관계수에 대응하는 최적분류점만을 구하였다.

정상상태 하의 분산의 크기가 부도상태의 분산보다 크거나 작음에 따라 그리고 ρ값에 따라서 최적분류점의 위치가 변동한다. 정상상태 하의 표준편차가 1보다 작은 $\sigma_3 = \sigma_4 = 0.75$인 경우에는 최적분류점이 오른쪽으로 이동하여 (μ_3, μ_4)에 가깝게 위치하고, 반면 표준편차가 1보다 큰 $\sigma_3 = \sigma_4 = 1.25$의 경우에는 최적분류점이 왼쪽으로 이동하여 원점에 가깝게 위치한다. $\sigma_3 = \sigma_4 = 1$일 때는 상관계수 ρ값에 관계없이 두 평균벡터의 중점이 최적분류점이 된다. 그리고 ρ가 증가할수록 $\sigma_3 = \sigma_4 = 0.75$인 경우에는 최적분류점 아래 방향의 왼쪽으로 이동하고, $\sigma_3 = \sigma_4 = 1.25$인 경우에는 최적분류점이 위 방향의 오른쪽으로 이동한다.

3. 최적분류함수

정상상태 하의 분포에서의 분산이 부도상태와 동일한 $\sigma_3 = \sigma_4 = 1$경우에의 최적분류함수는 최적분류점을 지나면서 두 평균벡터와 직교하는 법선이다. 그러나 식 (2.1)과 같이 정상상태 하의 분포에서의 분산이 부도상태와 동일하지 않은 경우에 즉, 두 이변량 정규분포의 공분산행렬이 서로 다른 경우에 최적분류함수를 제안한다. 식 (2.1)에서 가정한 두 이변량 정규분포에서 $\sigma_3 = \sigma_4$라고 가정하고 로그가능도비(loglikelihood ratio)를 구하면 다음과 같다.

$$\begin{aligned}
Q(x_1, x_2) &= \log\left[\frac{L(\mu_3, \mu_4, \sigma, \rho)}{L(0, 0, 1, \rho)}\right] \\
&= \frac{1-\sigma^2}{2(1-\rho^2)\sigma^2}x_1^2 - \frac{\rho(1-\sigma^2)}{(1-\rho^2)\sigma^2}x_1x_2 + \frac{1-\sigma^2}{2(1-\rho^2)\sigma^2}x_2^2 \\
&\quad - \frac{(\mu_3-\rho\mu_4)}{(1-\rho^2)\sigma^2}x_1 - \frac{(\mu_4-\rho\mu_3)}{(1-\rho^2)\sigma^2}x_2 + \frac{\mu_3^2 - 2\rho\mu_3\mu_4 + \mu_4^2}{2(1-\rho^2)\sigma^2} + \ln\left(\sigma^2\right) \\
&= \frac{1-\sigma^2}{2(1-\rho^2)\sigma^2}x_2^2 - \left[\frac{\rho(1-\sigma^2)}{(1-\rho^2)\sigma^2}x_1 + \frac{(\mu_4-\rho\mu_3)}{(1-\rho^2)\sigma^2}\right]x_2 \\
&\quad + \left[\frac{1-\sigma^2}{2(1-\rho^2)\sigma^2}x_1^2 - \frac{(\mu_3-\rho\mu_4)}{(1-\rho^2)\sigma^2}x_1 + \frac{\mu_3^2 - 2\rho\mu_3\mu_4 + \mu_4^2}{2(1-\rho^2)\sigma^2} + \ln\left(\sigma^2\right)\right].
\end{aligned}$$

$Q(x_1, x_2) = 0$으로 설정하고 x_2에 관해 정리하면, 다음과 같은 판별함수를 얻을 수 있다.

$$\begin{aligned}
x_2 &= \frac{\rho(1-\sigma^2)x_1 + (\mu_4 - \rho\mu_3)}{1-\sigma^2} \\
&\quad \pm \sqrt{(\rho^2-1)x_1^2 + \frac{2\mu_3(1-\rho^2)}{1-\sigma^2}x_1 + \frac{1}{1-\sigma^2}\left[\frac{(\mu_4-\rho\mu_3)^2}{1-\sigma^2} - (\mu_3^2 - 2\rho\mu_3\mu_4 + \mu_4^2) - 2(1-\rho^2)\sigma^2\ln\sigma^2\right]}, \quad (3.1)
\end{aligned}$$

여기서 식 (3.1)의 제곱근 안에서 근의 판별식은 0보다 크며 두 근이 항상 존재한다. 존재하는 두 근을 x_p, x_q $(x_p < x_q)$라고 하자.

$$\begin{aligned}
x_p &= \frac{\mu_3}{1-\sigma^2} - \sqrt{\frac{\mu_3^2}{(1-\sigma^2)^2} - \frac{1}{(1-\sigma^2)(\rho^2-1)}\left[\frac{(\mu_4-\rho\mu_3)^2}{1-\sigma^2} - (\mu_3^2 - 2\rho\mu_3\mu_4 + \mu_4^2) - 2(1-\rho^2)\sigma^2\ln\sigma^2\right]}, \\
x_q &= \frac{\mu_3}{1-\sigma^2} + \sqrt{\frac{\mu_3^2}{(1-\sigma^2)^2} - \frac{1}{(1-\sigma^2)(\rho^2-1)}\left[\frac{(\mu_4-\rho\mu_3)^2}{1-\sigma^2} - (\mu_3^2 - 2\rho\mu_3\mu_4 + \mu_4^2) - 2(1-\rho^2)\sigma^2\ln\sigma^2\right]}.
\end{aligned}$$

식 (3.1)으로부터 다음 두 가지 최적분류함수를 제안한다. 임의의 $x_1 \in (x_p, x_q)$에 대하여

$$\begin{aligned}
g^-(x_1) &= \frac{\rho(1-\sigma^2)x_1 + (\mu_4 - \rho\mu_3)}{1-\sigma^2} \\
&\quad - \sqrt{(\rho^2-1)x_1^2 + \frac{2\mu_3(1-\rho^2)}{1-\sigma^2}x_1 + \frac{1}{1-\sigma^2}\left[\frac{(\mu_4-\rho\mu_3)^2}{1-\sigma^2} - \left(\mu_3^2 - 2\rho\mu_3\mu_4 + \mu_4^2\right) - 2(1-\rho^2)\sigma^2\ln\sigma^2\right]}, \quad (3.2) \\
g^+(x_1) &= \frac{\rho(1-\sigma^2)x_1 + (\mu_4 - \rho\mu_3)}{1-\sigma^2} \\
&\quad + \sqrt{(\rho^2-1)x_1^2 + \frac{2\mu_3(1-\rho^2)}{1-\sigma^2}x_1 + \frac{1}{1-\sigma^2}\left[\frac{(\mu_4-\rho\mu_3)^2}{1-\sigma^2} - \left(\mu_3^2 - 2\rho\mu_3\mu_4 + \mu_4^2\right) - 2(1-\rho^2)\sigma^2\ln\sigma^2\right]}. \quad (3.3)
\end{aligned}$$

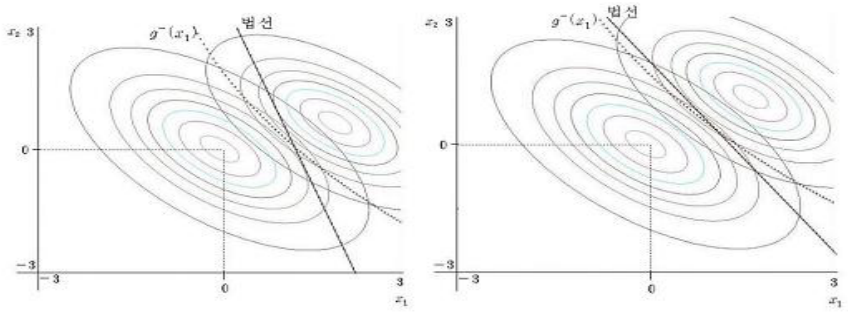

Figure 1: $\sigma_N = 0.75$, $\rho = -0.6$ *contour (left* $b = 1/3$*, right* $b = 2/3$*)*

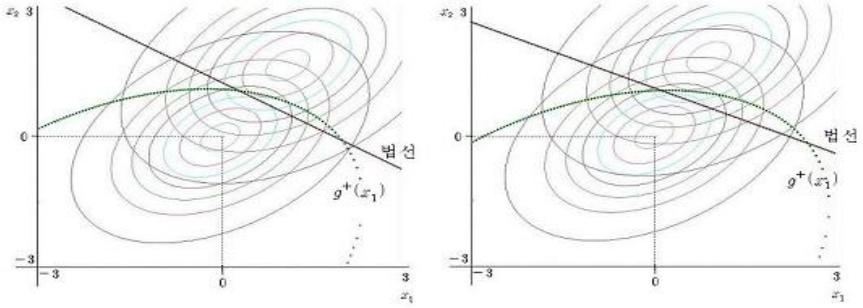

Figure 2: $\sigma_N = 1.25$, $\rho = 0.4$ *contour (left* $b = 1.5$*, right* $b = 2$*)*

4. FPR과 FNR

다양한 모분산과 모상관계수 그리고 기울기의 경우에 두 이변량 정규분포의 확률밀도함수를 등고선도(contour plot)로 Figure 1과 Figure 2에 표현하였으며, 식 (3.2) 또는 식 (3.3)의 최적분류함수와 Table 2에서 구한 최적분류점을 지나는 법선을 Figure 1과 Figure 2의 등고선도와 같이 나타내었다. 최적분류함수 식 (3.2)와 식 (3.3)은 Table 2에서 얻은 최적분류점을 통과하며 Figure 1과 Figure 2를 통해서도 탐색할 수 있다. 그러므로 본 연구에서 제안한 최적분류함수는 이변량 ROC 곡선으로 얻을 수 있으며 유용하게 활용할 수 있다.

최적분류함수 식 (3.2)와 식 (3.3)은 타원 형태이다. 정상상태 하의 표준편차가 1보다 작은 $\sigma_3 = \sigma_4 = 0.75$인 경우에 최적분류함수는 법선의 왼쪽에, 표준편차가 1보다 큰 $\sigma_3 = \sigma_4 = 1.25$의 경우에는 최적분류함수는 법선의 오른쪽에 나타난다. 따라서 $\sigma_3 = \sigma_4 = 0.75$에는 두 최적분류함수 중에서 식 (3.2)인 $g^-(x_1)$는 두 밀도함수가 중복되는 두 등고선 위를 지나가고 Table 2에서 구한 최적분류점을 통과한다. 그러나 $\sigma_3 = \sigma_4 = 1.25$에는 식 (3.3)인 $g^+(x_1)$가 최적분류점을 지난다. 그러므로 최적분류함수 $g^-(x_1)$과 $g^+(x_1)$ 중에서 최적분류점을 통과하는 법선에 가까운 함수 하나만 사용하여 오분류율을 구한다. 그러나 상관계수가 양수인 경우에는 최적분류함수 $g^-(x_1)$과 $g^+(x_1)$ 둘 다 두 밀도함수가 중복되는 등고선 위를 지나가며 두 최적분류함수의 경계점이 불명확하므로 FPR과 FNR를 구하기 위하여 적분식으로 표현하는 것이 쉽지 않다. 따라서 본 연구에서는 상관계수가 양수이며 큰 값을 갖는 경우는 제외하였다 (Table 2 참조). 3절에서 제시한 최적분류함수의 특징이 잘 드러나는 4가지를 Figure 1과 Figure 2에 제시하였다.

본 연구에서는 실제 부도를 정상으로 잘못 예측하는 비율(FNR)을 α로 그리고 실제 정상을 부도로 잘못 예측하는 비율(FPR)을 β로 정의하고, 식 (2.1)에서 가정한 다양한 이변량 정규분포의 경우에 α와 β를 구하기 위한 식을 다음과 같이 정리한다.

(I) $\sigma < 1$ 경우의 α와 β

$$\alpha = P(X_2 > g^-(X_1) \mid \theta_d) = P(X_2 > g^-(x_1) \mid X_1 \in (x_p, x_q), \theta_d) = 1 - \int_{x_p}^{x_q} \int_{-\infty}^{g^-(x_1)} f(x_2 \mid x_1) dx_2 f(x_1) dx_1,$$

여기서 $f(x_2 \mid x_1) = \Phi\left(x_2; \rho x_1, \sqrt{(1-\rho^2)}\right)$, $f(x_1) = \phi(x_1; 0, 1)$.

$$\beta = P(X_2 \le g^-(X_1) \mid \theta_n) = P(X_2 \le g^-(x_1) \mid X_1 \in (x_p, x_q), \theta_n) = \int_{x_p}^{x_q} \int_{-\infty}^{g^-(x_1)} f(x_2 \mid x_1) dx_2 f(x_1) dx_1,$$

여기서 $f(x_2 \mid x_1) = \Phi\left(x_2; \mu_4 + \rho(x_1 - \mu_3), \sqrt{\sigma_4^2(1-\rho^2)}\right)$, $f(x_1) = \phi(x_1; \mu_3, \sigma_3)$.

(II) $\sigma > 1$ 경우의 α와 β

$$\alpha = P(X_2 > g^+(X_1) \mid \theta_d) = P(X_2 > g^+(x_1) \mid X_1 \in (x_p, x_q), \theta_d) = 1 - \int_{x_p}^{x_q} \int_{-\infty}^{g^+(x_1)} f(x_2 \mid x_1) dx_2 f(x_1) dx_1,$$

여기서 $f(x_2 \mid x_1) = \Phi\left(x_2; \rho x_1, \sqrt{(1-\rho^2)}\right)$, $f(x_1) = \phi(x_1; 0, 1)$.

$$\beta = P(X_2 \le g^+(X_1) \mid \theta_n) = P(X_2 \le g^+(x_1) \mid X_1 \in (x_p, x_q), \theta_n) = \int_{x_p}^{x_q} \int_{-\infty}^{g^+(x_1)} f(x_2 \mid x_1) dx_2 f(x_1) dx_1,$$

여기서 $f(x_2 \mid x_1) = \Phi\left(x_2; \mu_4 + \rho(x_1 - \mu_3), \sqrt{\sigma_4^2(1-\rho^2)}\right)$, $f(x_1) = \phi(x_1; \mu_3, \sigma_3)$.

식 (2.1)에서 가정한 확률분포함수에서 다양한 기울기와 상관계수의 경우에 FNR인 α와 FPR인 β 그리고 두 종류의 오분류율합인 $\alpha + \beta$를 구하고 이를 Table 3에 정리하였다. Table 3을 바탕으로 기울기의 변화에 따른 오분류율과 상관계수의 변화에 대한 오분류율을 정상상태 하의 표준편차가 부도상태보다 작은 경우와 큰 경우로 구분하여 Figure 3부터 Figure 6에 나타냈다.

Figure 3과 Figure 4에서 오분류율 α, β, $\alpha + \beta$는 아래로 볼록한 모양에서 상관계수가 커질수록 완만한 직선형태가 된다. 따라서 상관계수가 작을 때는 기울기가 오분류율에 영향을 미치지만, 상관계수가 커질수록 기울기는 오분류율에 크게 영향을 미치지 않는다. 정상상태 하에서 표준편차가 1보다 작은 $\sigma_3 = \sigma_4 = 0.75$일 때 α가 β보다 더 큰 값을 갖고, 표준편차가 1보다 큰 $\sigma_3 = \sigma_4 = 1.25$인 경우에는 β가 α보다 큰 값을 가진다. 따라서 모분산의 크기에 따라서 α, β의 크기가 반대로 바뀌는 것을 알 수 있다. 또한 상관계수가 동일하다면, $\sigma_3 = \sigma_4 = 0.75$일 때의 오류율 보다 $\sigma_3 = \sigma_4 = 1.25$인 경우의 오류율이 더 크다. 그리고 $\sigma_3 = \sigma_4 = 0.75$일 때의 β는 상관계수가 커지면서 일정한 값으로 나타나고, $\sigma_3 = \sigma_4 = 1.25$일 때는 α가 상관계수가 커질수록 일정한 값을 갖는다.

Figure 5와 Figure 6에서는 동일한 기울기를 갖는 경우에 상관계수가 커지면 오분류율도 증가한다. 또한 Figure 3과 Figure 4에서 확인한 것과 같이 정상상태 하의 표준편차가 1보다 작은 $\sigma_3 = \sigma_4 = 0.75$일 때는 α가 β보다 더 큰 값이고, 표준편차가 1보다 큰 $\sigma_3 = \sigma_4 = 1.25$인 경우에는 β가 α보다 큰 값임을 Figure 5와 Figure 6을 통해서도 알 수 있다. 그리고 $\sigma_3 = \sigma_4 = 0.75$인 경우의 오류율합은 $\sigma_3 = \sigma_4 = 1.25$일 때의 오류율합 보다 작은 값으로 나타난다.

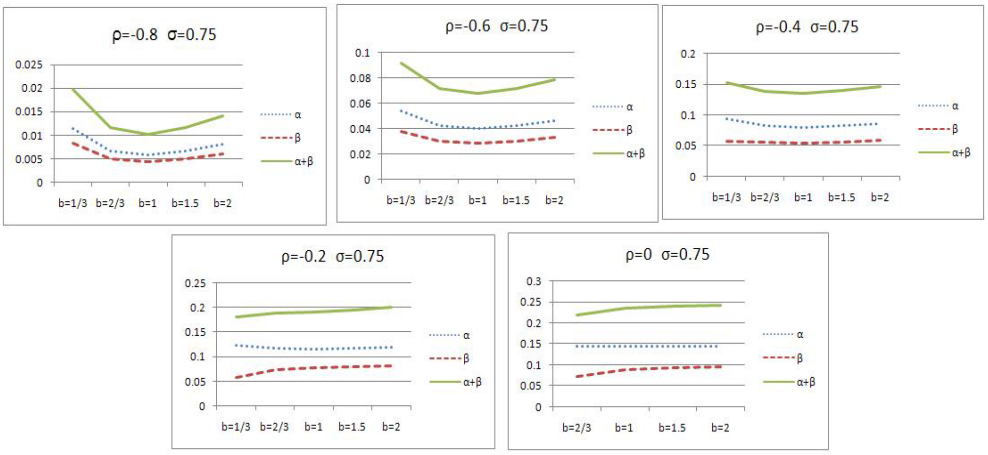

Figure 3: *FPR(β) and FNR(α) with slope (b)* ($\sigma_3 = \sigma_4 = 0.75$)

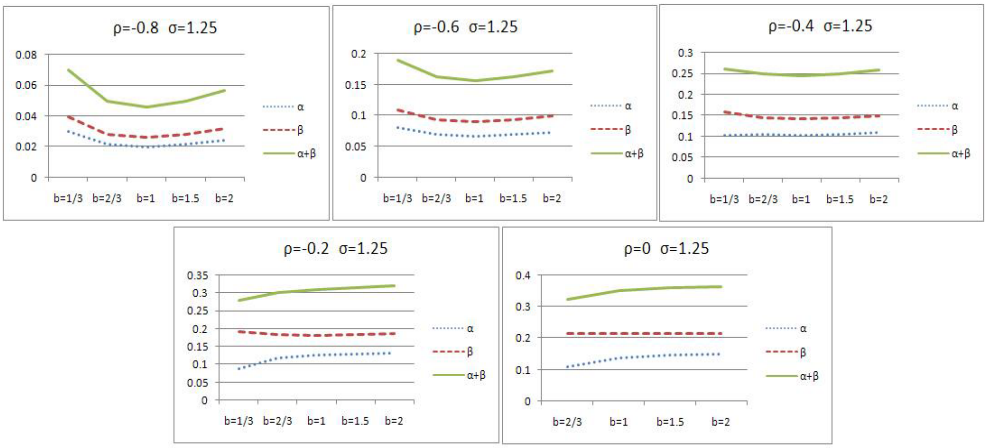

Figure 4: *FPR(β) and FNR(α) with slope (b)* ($\sigma_3 = \sigma_4 = 1.25$)

5. 결론

본 연구에서는 다양한 이변량 정규분포의 경우에 Hong 등 (2012)이 제안한 이변량 ROC 곡선을 구현하고, 이변량 ROC 곡선을 기반으로 최적분류함수를 구하였다.

부도와 정상상태의 확률분포가 정규분포를 따른다고 가정 하에 두 모평균벡터의 길이를 2로 고정시키고 모평균 사이의 기울기를 1/3부터 2까지 다양한 경우와 정상상태의 모분산을 부도상태의 모분산보다 작거나 큰 경우 그리고 다양한 상관계수의 경우를 고려하여, 실제 부도를 정상으로 잘못 예측하는 오분류율(FNR)과 실제 정상을 부도로 잘못 예측하는 오분류율(FPR)을 계산하였다. 이변량 ROC 곡선으로부터 구한 최적분류점을 통과하는 최적분류함수를 통해 계산하여 얻은 FPR과 FNR을 바탕으로 다음과 같은 결론을 유도하였다.

오분류율 FNR, FPR, FNR과 FPR의 합은 상관계수가 커질수록 아래로 볼록한 모양에서 완만한 직선형태가 된다. 그리고 상관계수가 작을 때는 기울기가 오분류율에 영향을 미치지만, 상관계수가 커질

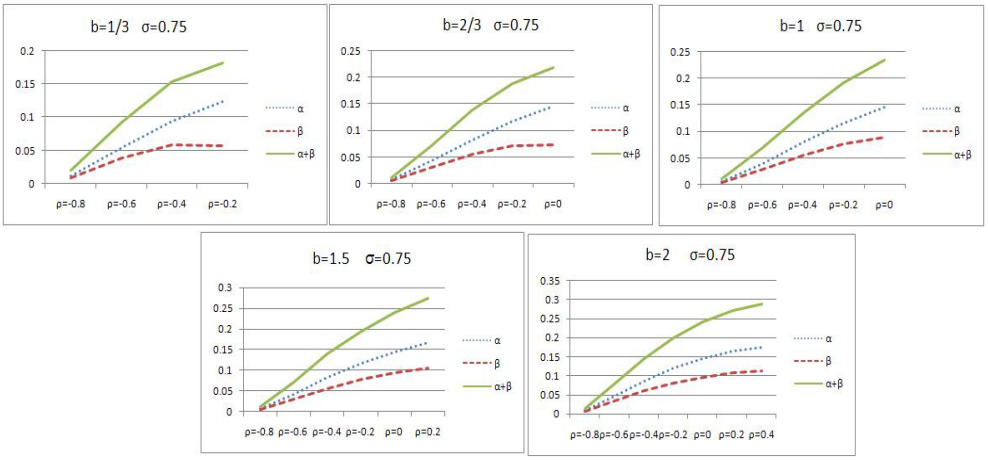

Figure 5: *FPR(β) and FNR(α) with Rho* ($\sigma_3 = \sigma_4 = 0.75$)

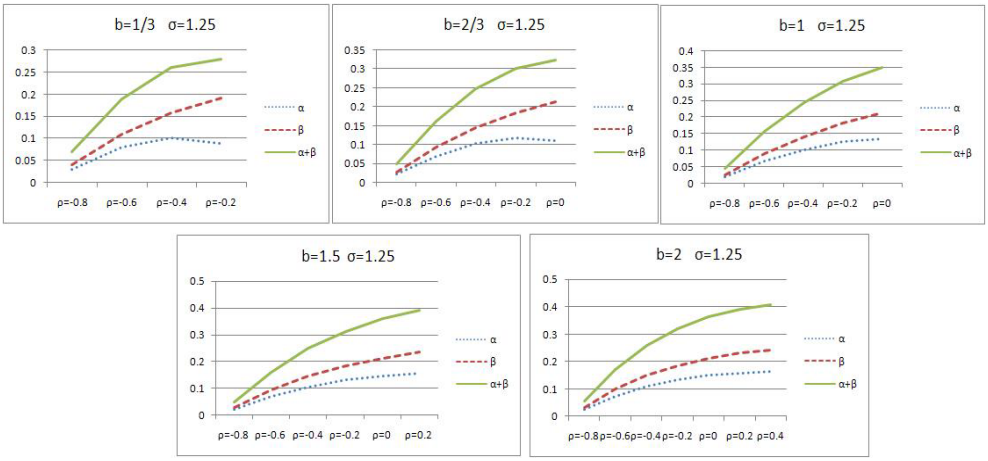

Figure 6: *FPR(β) and FNR(α) with Rho* ($\sigma_3 = \sigma_4 = 1.25$)

수록 기울기는 오분류율에 크게 영향을 미치지 않음을 파악하였다. 또한 정상상태 하에서 표준편차가 1보다 작은 경우에는 FNR이 FPR보다 더 큰 값을 갖고, 표준편차가 1보다 큰 경우에는 FPR이 FNR보다 큰 값을 가지기 때문에 모분산의 크기에 따라서 FNR, FPR의 크기가 반대로 바뀌는 것을 알 수 있었다. 상관계수가 동일하다면, 정상상태 하에서 표준편차가 1보다 작은 경우의 오류율 보다 표준편차가 1보다 큰 경우의 오류율이 더 크다. 그리고 정상상태 하에서 표준편차가 1보다 작은 경우의 FNR은 상관계수가 커지면서 일정한 값으로 나타나고, 표준편차가 1보다 큰 경우는 FPR은 상관계수가 커질수록 일정한 값을 갖는다. 기울기가 동일한 경우에 상관계수가 커지면 오분류율도 증가하며, 정상상태 하에서 표준편차가 1보다 작은 경우의 FNR과 FPR의 합은 표준편차가 1보다 큰 경우의 오류율합 보다 작은 값으로 나타남을 확인 할 수 있다.

일변량 분포함수를 표현하는 ROC 곡선을 이변량 분포함수로 확장하였으며, ROC 곡선의 분류기준으로부터 이변량 확률밀도함수들의 최적분류함수를 제안하였다. 본 연구에서 제안한 최적분류함수

Table 3: FNR(α) and FPR(β)

Slope	ρ	σ = 0.75			σ = 1.25		
		FNR	FPR	FNR + FPR	FNR	FPR	FNR + FPR
1/3	−0.8	0.0115	0.0830	0.0198	0.0302	0.0397	0.0699
	−0.6	0.0542	0.0378	0.0920	0.0804	0.1094	0.1897
	−0.4	0.0939	0.0586	0.1525	0.1020	0.1581	0.2601
	−0.2	0.1242	0.0571	0.1813	0.0886	0.1913	0.2799
2/3	−0.8	0.0067	0.0049	0.0116	0.0216	0.0281	0.0497
	−0.6	0.0423	0.0298	0.0722	0.0692	0.0928	0.2601
	−0.4	0.0822	0.0560	0.1382	0.1035	0.1445	0.2479
	−0.2	0.1167	0.0723	0.1890	0.1176	0.1833	0.3010
	0	0.1451	0.0734	0.2186	0.1096	0.2132	0.3227
1	−0.8	0.0059	0.0043	0.0102	0.0199	0.0259	0.0458
	−0.6	0.0400	0.0298	0.0682	0.0667	0.0893	0.1560
	−0.4	0.0796	0.0549	0.1345	0.1029	0.1414	0.2443
	−0.2	0.1150	0.0763	0.1913	0.1263	0.1815	0.3078
	0	0.1452	0.0891	0.2343	0.1357	0.2131	0.3489
1.5	−0.8	0.0067	0.0049	0.0116	0.0216	0.0281	0.0497
	−0.6	0.0423	0.0298	0.0722	0.0692	0.0928	0.1620
	−0.4	0.0822	0.0567	0.1388	0.1052	0.1445	0.2496
	−0.2	0.1167	0.0788	0.1955	0.1304	0.1833	0.3137
	0	0.1451	0.0953	0.2406	0.1468	0.2131	0.3599
	0.2	0.1683	0.1063	0.2746	0.1558	0.2361	0.3920
2	−0.8	0.0083	0.0060	0.0142	0.0245	0.0320	0.0565
	−0.6	0.0456	0.0327	0.0792	0.0734	0.9870	0.1721
	−0.4	0.0864	0.0595	0.1460	0.1086	0.1495	0.2581
	−0.2	0.1195	0.0808	0.2003	0.1328	0.1863	0.3191
	0	0.1452	0.0965	0.2418	0.1490	0.2131	0.3622
	0.2	0.1640	0.1072	0.2712	0.1592	0.2320	0.3912
	0.4	0.1751	0.1130	0.2882	0.1644	0.2429	0.4072

는 이변량 분포의 판별함수와 일치함을 보이고, 부도와 정상상태가 이분산일 때 오분류율을 구하였다. 이변량 ROC 곡선을 여러 분야에 적용하면, 다양한 활용을 기대할 수 있다. 그리고 정규분포 이외에 다른 분포를 가정하며 구한 최적분류점과 최적분류함수에 대해 추후 연구되어야 할 것이다. 또한 다양한 이변량 결합분포함수와 삼차원 이상의 다차원 결합분포함수의 ROC 곡선에 대한 연구를 향후 연구과제로 남겨 놓는다.

References

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

Hong, C. S. and Joo, J. S. (2010). Optimal thresholds from non-normal mixture, *The Korean Journal of Applied Statistics*, **23**, 943–953.

Hong, C. S., Kim, G. C. and Jeong, J. A. (2012). Bivariate ROC curve, *The Korean Journal of Applied Statistics*, **19**, 277–286.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*, Chapman & Hall/CRC, Monographs on Statistics & Applied Probability, 111, Florida.

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

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.

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

2012년 5월 14일 접수; 2012년 7월 2일 수정; 2012년 7월 13일 채택