The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.6.977
26(6), 977–985

Standard Criterion of VUS for ROC Surface

C. S. Hong[a,1] · E. S. Jung[a] · D. G. Jung[a]

[a]Department of Statistics, Sungkyunkwan University

(Received October 1, 2013; Revised October 28, 2013; Accepted October 28, 2013)

Abstract

Many situations are classified into more than two categories in real world. In this work, we consider ROC surface and VUS, which are graphical representation methods for classification models with three categories. The standard criteria of AUC for the probability of default based on Basel II is extended to the VUS for ROC surface; therefore, the standardized criteria of VUS for the classification model is proposed. The ranges of AUC, K-S and mean difference statistics corresponding to VUS values for each class of the standard criteria are obtained. The standard criteria of VUS for ROC surface can be established by exploring the relationships of these statistics.

Keywords: AUC, classification, default, FPR, risk, threshold, TPR, validation, VUS.

1. 서론

ROC(Receiver operating characteristic) 곡선은 성과(performance)를 기반으로 분류모형(classification model) 또는 분류자(classifiers)를 시각화하여 평가할 수 있는 유용한 방법이다. ROC 곡선은 분류자의 'Sensitivity(민감도)'와 '1-Specifity(1-특이도)' 사이에 교환(trade-off)을 나타내는 신호탐지 이론에서 오랫동안 사용되었으며, 의사결정과 의학진단의 체계 등에서 폭넓게 사용되었다. ROC 곡선의 특성에 관한 연구와 ROC 분석의 응용과 관련된 정보는 Provost와 Fawcett (2001), Sobehart와 Keenan (2001), Zho 등 (2007), Engelmann 등 (2003), Fawcett (2003), Hong과 Choi (2009), Hong 등 (2010) 이외의 많은 문헌에서 발견할 수 있다. ROC 곡선을 통해 판별력을 측정하는 객관적인 통계량으로는 ROC 곡선의 아래 면적을 계산한 AUC(Area Under ROC Curve 또는 AUROC)를 사용한다 (Bradley, 1997; Hanley와 McNeil, 1982).

각 국가의 중앙은행을 감독하는 국제결제은행(Bank of International Settlements: BIS)은 은행의 파산을 막기 위하여 2004년에 국제적인 위험(risk) 관리제도를 표준화하여 Basel II를 제안하였다. Joseph (2005)는 Basel II에 대한 부도확률(probability of default)을 바탕으로 평균차이(mean difference)에 대응하는 AUC의 판별력 판단기준을 제안하였다. Joseph (2005)가 제안한 판별력 판단기준은 Wilkie (2004)의 방법을 확장한 것으로 부도(불량)과 정상이 동일한 표준편차를 갖는 정규분포 가정하에서 평균차이에 대응하는 AUC 통계량을 바탕으로 Table 1.1을 제안하였다.

실제 현실에서는 두 가지 형태로 분류되어 있는 상황보다 더 많은 형태로 분류되어 있는 상황이 많이 존재하고, 이에 따라 다범주 분류모형의 판별력을 측정할 수 있는 방법이 요구된다. 신용평가모형에서도

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

Table 1.1. Standardized criteria for AUC and K-S

Validation Range	Meaning	Mean Diffrence	AUC	K-S Statistic
0–1	Random	0.00	0.5000	0
1–2	Doubtful	0.25	0.5700	0.0995
2–3	Poor	0.50	0.6380	0.1974
3–4	Marginal	0.75	0.7020	0.2923
4–5	Satisfactory	1.00	0.7600	0.3829
5–6	Good	1.25	0.8115	0.4680
6–7	Very Good	1.50	0.8555	0.5467
7–8	Strong	1.75	0.8920	0.6184
8–9	Very Strong	2.00	0.9215	0.6827
9–10	Excellent	2.25	0.9504	0.7394
10–11	Excellent	2.50	0.9710	0.7887
11–12	Excellent	2.75	0.9857	0.8309
12–13	Superior	3.00	0.9946	0.8664

기존에 정상, 부도의 두 범주에 의한 판별력 측정이 아닌 보다 많은 범주(정상, 주의, 부도)에 대하여 모형의 판별력을 측정할 필요성이 있으나, ROC 곡선의 경우에는 두 가지 형태에 대하여만 분류되는 구조적인 한계가 있다. 따라서 본 논문에서는 세 종류의 범주형태로 분류되는 상황에 대하여 이를 분류하는 모형을 시각적으로 표현하는 ROC 곡면(surface) 방법을 고려하고, ROC 곡면의 아래 부피를 측정한 VUS(Volume Under ROC Surface) 통계량에 대해 살펴보고자 한다.

ROC 곡면과 VUS 통계량의 경우 Scurfield (1996), Mossman (1999), Dreiseitl 등 (2000), Heckerling (2001), Fawcett (2003), Nakas와 Yiannoutsos (2004), Nakas 등 (2010), Patel과 Markey (2005), Wandishin과 Mullen (2009) 등에 의해 연구들이 많이 진행되었지만 VUS 통계량을 바탕으로 분류모형의 판별력을 판단하는 기준에 대하여는 명확하게 밝혀지지 않았다. 본 논문에서는 세 범주 분류모형의 판별력을 측정하는 VUS 통계량에 관한 판단기준을 연구한다. 이를 위해 Joseph (2005)의 연구 방법을 적용하여 VUS 값을 바탕으로 13단계의 판단기준을 제안하고, 각 판단기준의 VUS 통계량에 대응하는 AUC 통계량, K-S 통계량 그리고 세 분포의 평균차이에 대한 범위를 구하여 이들의 관계에 대해 살펴봄으로써 세 범주 분류모형의 판별력 판단기준을 설정한다.

본 논문의 구성은 다음과 같다. 2절에서는 ROC 곡선과 AUC 통계량을 간략히 소개하고 이를 3차원으로 확장한 ROC 곡면과 VUS 통계량을 설명한다. 그리고 AUC와 VUS 통계량과의 관계에 대해 설명한다. 3절에서는 Joseph (2005)의 연구 방법을 적용하여 13단계의 Validation Range를 설정하고 각 Validation Range에 따른 VUS 값을 바탕으로 판단기준을 제안한다. 또한, 각 Validation Range에서의 VUS 값에 대응하는 AUC 통계량, K-S 통계량, 그리고 평균차이와의 관계를 파악하고 이 결과와 더불어 VUS 통계량에 의한 판단기준을 설정한다. 마지막으로 4절에서는 VUS를 바탕으로 제안한 판단기준의 결과에 대해 정리하고 향후 연구과제에 대해 토론한다.

2. ROC 곡면과 VUS

모형에서 대상을 두가지 범주로 나뉘어져 있는 경우에 부도와 정상을 나타내는 스코어 확률변수 X_1과 X_2는 $F_1(\cdot)$과 $F_2(\cdot)$ 분포를 각각 따른다고 정의한다. ROC 곡선은 각 절단점(cut-off value, threshold)의 스코어 s에서 얻는 비율들로 구성되어 있으며, 실제 부도를 부도로 정확히 예측하는 비율 TPR(true positive rate) $F_1(s)$과 실제 정상을 부도로 잘못 예측하는 비율 FPR(false positive rate) $F_2(s)$을 각각 Y축과 X축 좌표에 대응시킨 그래프로 표현된다 (Tasche, 2006).

Table 2.1. Confution matrix

		Event		
		e_1	e_2	e_3
Decision	d_1	$F_1(c_1)$	$F_2(c_1)$	$F_3(c_1)$
	d_2	$F_1(c_2) - F_1(c_1)$	$F_2(c_2) - F_2(c_1)$	$F_3(c_2) - F_3(c_1)$
	d_3	$1 - F_1(c_2)$	$1 - F_2(c_2)$	$1 - F_3(c_2)$

AUC 통계량은 ROC 곡선의 하단 부분의 넓이를 의미하며, ROC 함수를 적분한 값이다. AUC는 0.5와 1사이의 값을 갖고 1에 가까울수록 분류모형에 대한 판별력이 높으며 다음과 같이 정의한다 (Lim, 2005; Joseph, 2005).

$$\text{AUC} = P(X_1 < X_2) = \int_0^1 \text{ROC}(u)du,$$

여기서 $\text{ROC}(u) = F_1(F_2^{-1}(u))$, $u \in (0, 1)$.

Hosmer와 Lemeshow (2000)는 AUC가 0.5이면 판별력이 전혀 없고 0.7∼0.8 사이의 값이면 모형이 채택할만한 수준, 0.8∼0.9 정도의 값이면 매우 판별력이 좋은 모형이라고 하였으며, Joseph (2005)는 부도와 정상이 정규분포를 따른다는 가정하에서 AUC를 바탕으로 하는 판단기준을 Table 1.1과 같이 제안하였다.

분류모형에서 대상을 세가지 범주로 판별하는 문제를 고려하자. 실제상태의 모수공간은 $\Omega = \{e_1, e_2, e_3\}$으로, 분류모형의 결과상태는 $\{d_1, d_2, d_3\}$으로 정의한다. 실제상태를 조건으로 결과를 예측하는 조건부확률 $P(d_i|e_i)$ $(i = 1, 2, 3,\ j = 1, 2, 3)$을 구할 수 있고 이를 통해 각각의 정분류와 오분류로 이루어진 3×3 혼동행렬(confusion matrix)을 구한다. 확률변수 X_1, X_2, X_3가 각각 $F_1(\cdot)$, $F_2(\cdot)$, $F_3(\cdot)$ $(F_1(x) \geq F_2(x) \geq F_3(x))$ 분포를 따른다고 가정하였을 때 X축을 따라 움직이는 임의의 절단점 c_1과 c_2에 해당되는 각각의 $F_1(\cdot)$, $F_2(\cdot)$, $F_3(\cdot)$를 계산하여 각 범주에 대한 정분류율과 오분류율을 구할 수 있고, 그 결과는 Table 2.1과 같이 3×3 혼동행렬로 정리된다.

확률변수 X와 절단점 c_1, c_2 $(c_1 \leq c_2)$와의 관계에 의해서 예측한 분류결과 d_1, d_2, d_3 ($X \leq c_1$이면 d_1, $c_1 \leq X \leq c_2$이면 d_2, $c_2 \leq X$이면 d_3)를 정의할 수 있다. 또한 실제 대상의 상태를 정확히 예측한 비율을 p_1, p_2, p_3라 한다면, $F_1(c_1) = p_1$, $F_2(c_2) - F_2(c_1) = p_2$, $1 - F_3(c_2) = p_3$로 정의할 수 있으며 임의의 절단점 c_1, c_2에 대해 정의된 p_1, p_2, p_3들을 3차원 좌표에 대응시켜 ROC 곡면을 $(F_1(c_1), F_2(c_2) - F_2(c_1), 1 - F_3(c_2))$로 정의하고, Figure 2.1과 같이 표현한다. ROC 곡면은 각 범주의 실제 대상의 상태를 정확히 예측하는 비율로 표현되며 이에 근거하여 동일한 X와 Y축의 값에 대응하는 Z축이 높을수록 판별력이 좋은 모형이다. 즉 ROC 곡면이 점 $(1, 1, 1)$에 가까워짐에 따라 또는 ROC 곡면 아래의 부피가 증가할수록 모형의 판별력이 좋다고 평가한다.

ROC 곡면 아래의 부피를 VUS(volume under ROC surface) 통계량이라 하고, 이는 ROC 곡면의 아래 부분을 중적분하여 계산한 값이다. 실제 대상의 상태를 제대로 판별하지 못한 정보 가치가 없는 모형의 VUS 값은 1/6이며 (Figure 2.1의 오른쪽 그림), 완벽하게 판별한 모형의 VUS 값은 1이다. 즉 VUS 값이 1에 가까울수록 분류모형의 판별력이 높다고 판단할 수 있다. VUS 통계량은 다음과 같이 정리한다 (Nakas와 Yiannoutsos, 2004).

$$\begin{aligned}\text{VUS} &= P(X_1 \leq X_2 \leq X_3) \\ &= \int_0^1 \int_0^{F_1\left(F_3^{-1}(1-p3)\right)} \text{ROC}_s(p_1, p_3)dp_1dp_3 = \int_0^1 F_1\left(F_2^{-1}(u)\right)\left[1 - F_3\left(F_2^{-1}(u)\right)\right]du,\end{aligned}$$

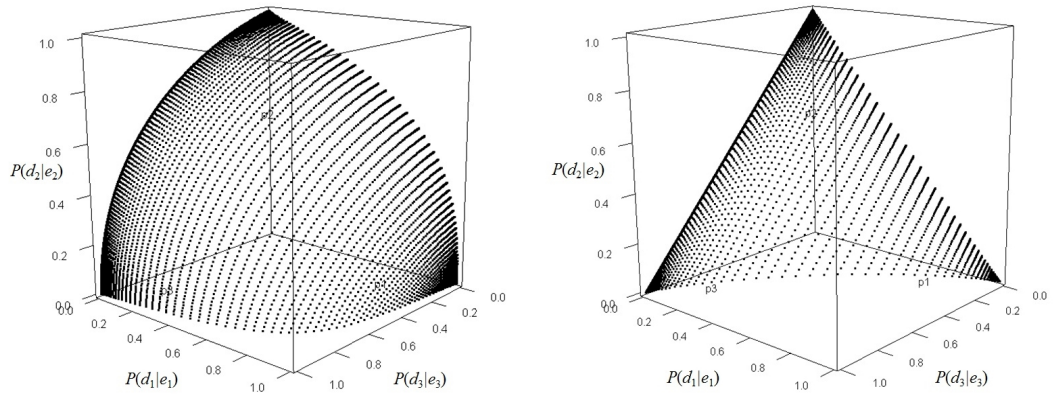

Figure 2.1. ROC surfaces of a predicted model and a random model

여기서 $\mathrm{ROC}_s(p_1, p_3) = F_2(F_3^{-1}(1-p_3)) - F_2(F_1^{-1}(p_1))$, $0 \le p_1, p_3 \le 1$.

절단점 c_1, c_2의 다양한 조건에 따라 하나의 3×3 혼동 행렬에서 세 가지 경우의 2×2 혼동 행렬 얻을 수 있으며, 이를 통해 ROC 곡선과 AUC 통계량을 구할 수 있다. AUC 통계량은 $P(X_i < X_j)(i = 1, 2, j = 2, 3, i \neq j)$으로 표현 가능하므로 세 종류의 AUC 통계량은 여섯 종류의 VUS 통계량에 의해 다음과 같은 관계를 갖는다 (Scurfield, 1996).

$$\begin{aligned}
\mathrm{AUC}_{12} &= \mathrm{VUS}_{123} + \mathrm{VUS}_{132} + \mathrm{VUS}_{312}, \\
\mathrm{AUC}_{23} &= \mathrm{VUS}_{123} + \mathrm{VUS}_{213} + \mathrm{VUS}_{231}, \\
\mathrm{AUC}_{13} &= \mathrm{VUS}_{123} + \mathrm{VUS}_{132} + \mathrm{VUS}_{213}.
\end{aligned}$$

여기서 $\mathrm{AUC}_{ij} = P(X_i \le X_j)$, $\mathrm{AUC}_{ijk} = P(X_i \le X_j \le X_k)$.

3. VUS 판단기준

VUS 통계량은 ROC 곡면의 주관적인 판단을 객관화하는 통계량으로 사용될 수 있으나 VUS 통계량을 통해 모형의 판별력을 판단할 수 있는 기준이 명확하지 않으므로 본 연구에서는 2차원의 AUC 통계량의 판단기준을 제안한 Joseph (2005)의 연구를 확장하여 3차원의 VUS 통계량의 판단기준에 대하여 연구한다.

스코어 확률변수의 분포함수를 정규분포를 가정한 Wilkie (2004)의 연구 (Thomas 등 (2004)의 4장 참조)를 확장하여 Joseph (2005)는 AUC 통계량의 판단기준을 제안하였다. 본 연구는 VUS 통계량의 판단기준을 설정하기 위하여 확률변수 X_2와 X_3의 분포함수 $F_2(\cdot)$와 $F_3(\cdot)$를 Joseph (2005)가 사용한 연구방법을 가정하고, 확률변수 X_1의 분포함수 $F_1(\cdot)$의 모평균은 $F_3(\cdot)$의 모평균의 부호를 음수로 설정한다. 즉 $F_2(\cdot) \equiv \Phi(x; 0, 1)$을 기준으로 $F_3(\cdot) \equiv \Phi(x; \mu, 1)$ 그리고 $F_1(\cdot) \equiv \Phi(x; -\mu, 1)$으로 설정한다. $F_3(\cdot)$의 모평균 μ의 범위를 Joseph (2005)와 유사하게 0.0부터 3.0까지 0.05 간격으로 설정하고 따라서 $F_1(\cdot)$의 모평균은 -3.0부터 0.0까지의 범위로 설정하여 $F_1(x) \ge F_2(x) \ge F_3(x)$의 성질을 유지한다. 그리고 $F_1(\cdot)$과 $F_3(\cdot)$의 모평균의 절대값이 동일한 경우의 VUS 값을 바탕으로 판단기준을 13단계로 나누고 각 단계에서 AUC_{12}와 AUC_{23} 값을 바탕으로 $\mathrm{AUC}_{12} + \mathrm{AUC}_{23}$의 값을 구하였다. 그리고 13단계

Table 3.1. Standardized criteria for VUS and other statistics

Validation Range	Meaning	VUS	Range of $AUC_{12} + AUC_{23}$	Range of $\mu + \mu$	Range of $KS_{12} + KS_{23}$
0–1	Random	0.1667	1.0000	(0.00, 0.05)	(0.0000, 0.0199)
1–2	Doubtful	0.2453	(1.0000, 1.1403)	(0.05, 0.05)	(0.0199, 0.1989)
2–3	Poor	0.3372	(1.1403, 1.2763)	(0.05, 1.10)	(0.1990, 0.4204)
3–4	Marginal	0.4365	(1.2763, 1.4041)	(1.00, 2.00)	(0.3948, 0.6827)
4–5	Satisfactory	0.5362	(1.4041, 1.5205)	(1.55, 3.15)	(0.5978, 0.9262)
5–6	Good	0.6301	(1.5205, 1.6232)	(2.05, 3.50)	(0.7766, 1.0638)
6–7	Very Good	0.7139	(1.6232, 1.7112)	(2.50, 3.85)	(0.9361, 1.1956)
7–8	Strong	0.7850	(1.7112, 1.7841)	(3.05, 4.15)	(1.1027, 1.3113)
8–9	Very Strong	0.8430	(1.7841, 1.8427)	(3.55, 4.50)	(1.2444, 1.4131)
9–10	Excellent	0.8885	(1.8427, 1.8888)	(4.00, 4.85)	(1.3654, 1.5114)
10–11	Excellent	0.9229	(1.8888, 1.9229)	(4.55, 5.15)	(1.4857, 1.5885)
11–12	Excellent	0.9482	(1.9229, 1.9482)	(5.00, 5.55)	(1.5774, 1.6641)
12–13	Superior	0.9661	(1.9482, 1.9661)	(5.55, 5.95)	(1.6662, 1.7262)

로 구분한 각 단계에서 $F_1(\cdot)$과 $F_3(\cdot)$의 모평균이 어느 정도 변화하는지를 살펴보기 위하여 모평균의 변화 범위와 Table 1.1에서와 같이 K-S 통계량의 변화 범위를 구하여 Table 3.1에 정리하였다.

Table 3.1을 통해 먼저 VUS 값이 0.1667(= 1/6)인 경우, 설정한 세 모형이 모두 동일하여 Random하다고 판단되는 경우 ($AUC_{12} = 0.5$, $AUC_{12} = 0.5$)의 VUS 값이다. Validation Range 1–2의 경우에서는 가정한 분포 중 평균차가 0.25인 경우에 VUS 값이 0.2453이므로 VUS 값이 0.1667부터 0.2453까지인 경우를 Validation Range 1–2라고 설정하며 'Doubtful'이라고 판단한다. 이에 해당되는 AUC_{12}와 AUC_{23}의 합은 (1.0000, 1.1403) 사이의 값을 나타낸다. Validation Range 2–3의 경우는 VUS 값이 조금 커진 0.2453부터 0.3372까지이며 'Marginal'이라고 판단한다. 이에 대응하는 AUC_{12}와 AUC_{23}의 합은 $(1.1403, 1.2763)$ 사이의 값을 나타낸다.

Validation Range의 단계가 상승할수록 VUS 값이 커지며 AUC_{12}와 AUC_{23}의 합도 증가한다. 그리고 각각의 Validation Range 단계에서 AUC_{12} 값이 커지면 AUC_{23} 값이 작아지며 반대로 AUC_{12} 값이 작아지면 AUC_{23} 값이 커진다. 13단계의 Validation Range 각각에서의 AUC_{12}와 AUC_{23} 값을 Figure 3.1에 표현하였다. Figure 3.1에서와 같이 Validation Range 단계가 상승할수록 AUC_{12}와 AUC_{23}의 합은 커지게 되고, AUC_{12}와 AUC_{23}의 관계는 기울기가 -1에 가까운 직선으로 표현할 수 있다.

각각의 Validation Range 단계에 대응하는 $F_3(\cdot)$의 모평균 μ와 $F_1(\cdot)$에서의 모평균 $-\mu$의 범위 즉, 평균차이의 범위를 살펴보기 위하여 Figure 3.1과 유사하게 두 분포함수의 모평균차이인 μ의 합에 대한 범위를 구하여 Table 3.1에서 살펴보았고, μ들의 관계를 Figure 3.2에 표현하였다. 이를 통해 Validation Range 단계가 올라갈수록 모평균 μ의 합도 커지는 것을 알 수 있고, μ들은 점점 우측 상단으로 움직이는 것을 확인할 수 있다. 각 Validation Range 단계에서의 μ들의 분포는 기울기가 -1과 유사한 비선형 형태로 표현되는 것을 볼 수 있으며 이 때문에 μ의 합의 범위는 3–4단계부터 11–12단계까지 겹쳐서 나타나게 된다. 따라서 Table 3.1의 모평균의 합의 범위보다 Figure 3.2를 바탕으로 모평균들의 값을 탐색하는 것을 추천한다.

Joseph (2005)가 제안한 Table 1.1에서 각각의 Validation Range 단계에서의 K-S 값을 제시하였는데 본 연구에서도 각 Validation Range 단계의 VUS 값에 대응하는 $F_1(\cdot)$과 $F_2(\cdot)$ 그리고 $F_2(\cdot)$와 $F_3(\cdot)$에 대한 K-S 값 KS_{12}와 KS_{23}를 구하여 Table 3.1에 $KS_{12} + KS_{23}$의 범위를 제시하였으며, Figure 3.3에

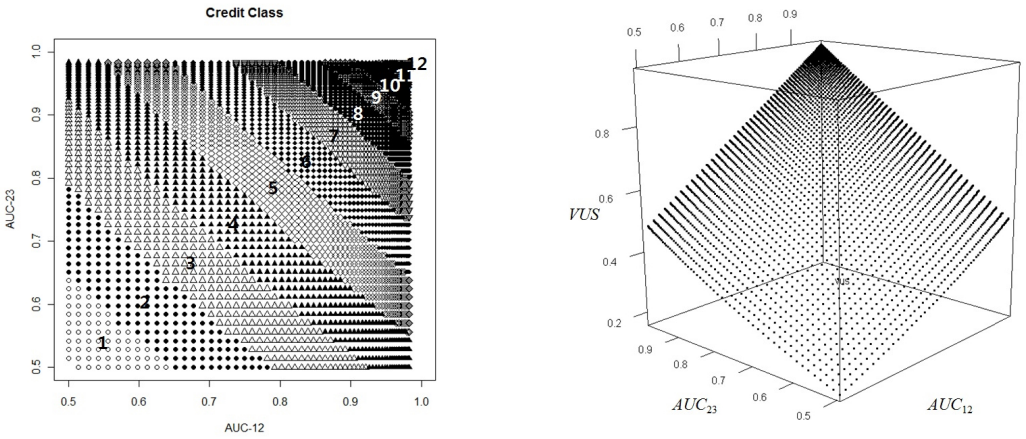

Figure 3.1. AUC_{12} and AUC_{23} for each validation range

서 KS_{12}와 KS_{23}의 관계에 대해 표현하였다. 이를 살펴보면 Validation Range 단계가 올라갈수록 K-S 통계량의 합의 크기도 커지는 것을 알 수 있다. Figure 3.3의 경우 Figure 3.2와 유사한 형태를 띄고 있으며, 이로 인해 3–4단계부터 11–12단계에 해당하는 K-S 통계량의 합의 범위의 경우 상위 단계와 겹쳐서 나타나는 것을 확인할 수 있다.

예를 들어 VUS 값이 0.50이라면 Validation Range가 3–4 단계이고, AUC_{12}와 AUC_{23}가 각각 0.7과 0.8이고 그 합이 1.5이면, Validation Range가 4–5 단계로 판단할 수 있다. 그리고 $F_3(\cdot)$의 모평균 μ와 $F_1(\cdot)$에서의 모평균 $-\mu$가 각각 1.75와 -0.60이라면 $\mu + \mu = 2.35$로 Validation Range가 5–6 단계이고, KS_{12}와 KS_{23}가 각각 0.35와 0.75이면 $KS_{12} + KS_{23} = 1.10$으로 Validation Range가 6–7 단계인 'Very Good'의 의미를 갖는 기준이라고 판단할 수 있다.

4. 결론

ROC 곡선과 AUC 통계량은 분류모형의 정확도를 평가할 수 있는 유용한 방법이다. 하지만 분류하는 범주의 수가 두 종류인 경우에만 판별이 가능하다. 본 논문에서는 ROC 곡면과 VUS 통계량을 이용하여 범주의 수가 세 종류인 경우에 모형의 판별력을 판단할 수 있는 기준에 대해 연구하였다.

VUS 통계량의 판단기준을 설정하기 위하여 Joseph (2005)의 연구를 활용하여 분포함수 $F_2(\cdot)$를 기준으로 분포함수 $F_1(\cdot)$과 $F_3(\cdot)$를 좌우 대칭되고 각 모평균의 절대값을 동일한 범위로 설정하였다. 다양한 경우의 조합에서 VUS 값을 기준으로 판단기준을 13단계로 나누었고 각 단계에서 기준으로 설정한 VUS 값에 대응하는 AUC_{12}와 AUC_{23}, 모평균차이 그리고 K-S 통계량을 살펴보았다. VUS 값에 대응하는 AUC, μ와 K-S 통계량들은 $F_1(\cdot)$과 $F_2(\cdot)$에 대응하는 값이 커지면 $F_2(\cdot)$와 $F_3(\cdot)$에 대응하는 값이 작아지며 반대로 $F_1(\cdot)$과 $F_2(\cdot)$에 대응하는 값이 작아지면 $F_2(\cdot)$와 $F_3(\cdot)$에 대응하는 값이 커진다. 그러므로 VUS 값에 대응하는 AUC, μ와 K-S 통계량들의 합을 구하고 탐색하였다. 그들의 합의 범위를 제시함으로써 VUS 통계량을 바탕으로 설정한 판단기준과 함께 설명할 수 있다.

Validation Range의 단계가 상승할수록 VUS 값과 AUC_{12}와 AUC_{23}의 합, 평균차이, K-S 통계량이 모두 증가한다. 또한 각각의 Validation Range 단계에서 AUC_{12}와 AUC_{23}는 기울기가 -1에 가까운 음

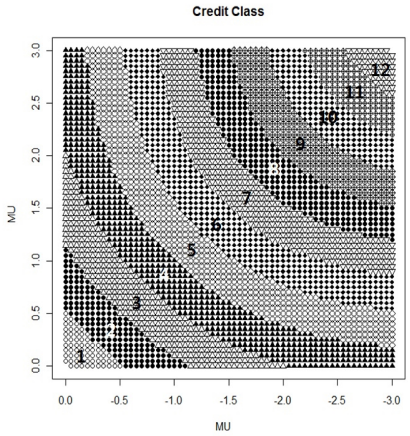

Figure 3.2. μ and μ for each validation range

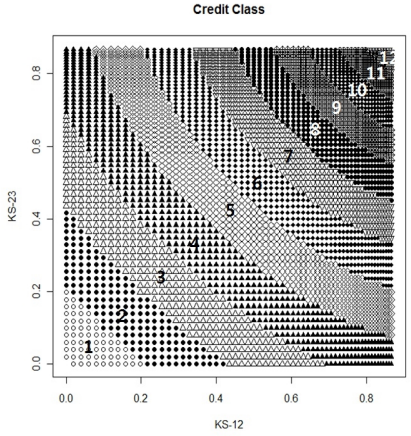

Figure 3.3 KS_{12} and KS_{23} for each validation range

의 선형관계를 갖고 있으며, 평균차이, K-S 통계량은 비선형관계로 나타났다. 본 연구의 3절에서 얻은 Table 3.1과 Figure 3.1부터 Figure 3.3까지의 결과를 바탕으로 VUS 값에 따라 13단계의 Validation Range를 제시하여 판단기준을 설정하였으며, VUS 값에 대응하는 AUC_{12}와 AUC_{23}, $F_3(\cdot)$의 모평균 μ와 $F_1(\cdot)$에서의 모평균 $-\mu$, 그리고 KS_{12}와 KS_{23} 값들과 각각의 합에 대하여도 판단기준을 제안하고 예를 들어 설명하였다.

범주의 수가 두 종류와 세 종류인 분류모형은 ROC 곡선과 ROC 곡면을 이용하여 탐색할 수 있으며 각각에 대한 AUC와 VUS로 판별력을 측정할 수 있지만, 범주의 수가 세 종류 이상인 경우에는 그래픽적인 방법으로 표현하는데 한계가 있다. 그러나 Scurfield (1996)가 언급하고 본 논문의 3절에서 설명한 AUC와 VUS의 관계를 확장하면, 범주의 수준이 네 종류 이상인 경우의 분류방법에도 적용이 가능하다. 따라서 본 연구를 범주의 수가 네 종류 이상의 분류하는 연구에서도 모형의 판별력을 판단할 수 있는 기준을 설정할 수 있고 이는 향후 연구 과제로 남겨둔다.

References

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30**, 1145–1159.

Dreiseitl, S., Ohno-Machado, L. and Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis, *Medical Decision Making*, **20**, 323–331.

Engelmann, B., Hayden, E. and Tasche, D. (2003). Measuring the discriminative power of rating systems, *Risk*, 82–86.

Fawcett, T. (2003). *ROC graphs: notes and practical considerations for data mining researchers*, HP Labs Tech Report HPL-2003-4.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143**, 29–36.

Heckerling, P. S. (2001). Parametric three-way receiver operating characteristic surface analysis using mathematica, *Medical Decision Making*, **21**, 409–417.

Hong, C. S. and Choi, J. S. (2009). Optimal threshold from ROC and CAP curves, *The Korean Journal of Applied Statistics*, **22**, 911–921.

Hong, C. S., Joo, J. S. and Choi, J. S. (2010). Optimal thresholds from mixture distributions, *The Korean Journal of Applied Statistics*, **23**, 13–28.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, John Wiley & Sons, New York.

Joseph, M. P. (2005). A PD validation framework for Basel II internal ratings-based systems, *Credit Scoring and Credit Control IV*.

Lim, C. K. (2005). Introduction of goodness-of-fit test methods for credit evaluation system, *Financial Supervisory Service, Risk Review*, 33–54.

Mossman, D. (1999). Three-way ROCs, *Medical Decision Making*, **19**, 78–89.

Nakas, C. T., Alonzo, T. A. and Yiannoutsos, C. T. (2010). Accuracy and cut off point selection in three class classification problems using a generalization of the Youden index, *Statistics in Medicine*, **29**, 2946–2955.

Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements, *Statistics in Medicine*, **23**, 3437–3449.

Patel, A. C. and Markey, M. K. (2005). Comparison of three-class classification performance metrics: a case study in breast cancer CAD, *International Society for Optical Engineering*, **5749**, 581–589.

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

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability, *Journal of Mathematical Psychology*, **40**, 253–269.

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, *The Analytics of Risk Model Validation*, **28**, 169–196.

Thomas, L. C., Edelman, D. B. and Crook, J. N. (2004). *Readings in Credit Scoring: Foundations, Developments, and Aims*, Oxford finance, Oxford University Press, New York.

Wandishin, M. S. and Mullen, S. J. (2009). Multiclass ROC analysis, *Weather and Forecasting*, **24**, 530–547.

Wilkie, A. D. (2004). *Measures for comparing scoring systems, in readings in credit scoring-recent decelopments, Advances, and Aims*, Oxford Finance.

Zou, K. H. (2002). Receiver operation characteristic literature research, On-line ibliography available from: http://www.spl.havard.edu/pages/ppl/zou/roc.html.

ROC 곡면에서 VUS의 판단기준

홍종선[a,1] · 정의석[a] · 정동근[a]

[a]성균관대학교 통계학과

(2013년 10월 1일 접수, 2013년 10월 28일 수정, 2013년 10월 28일 채택)

요 약

현실세계에는 두 가지 범주 이상으로 분류되는 경우가 많이 존재한다. 본 논문은 분류범주가 세 종류인 분류모형을 시각적으로 표현하는 방법인 ROC 곡면과 이 곡면 아래의 체적을 나타내는 VUS 통계량을 고려한다. 바젤 II에 기반한 부도확률에 관한 AUC 통계량의 판단기준을 ROC 곡면에서의 VUS에 대하여 확장하여, VUS에 의한 판별력 판단기준 13단계를 제안한다. 제안한 판단기준 각 단계에서의 VUS값에 대응하는 AUC, K-S 통계량 그리고 세 분포의 평균차이에 대한 범위를 탐색하고, 이들의 관계를 살펴봄으로써 VUS 통계량의 판별력 판단기준을 설정한다.

주요용어: 부도, 분류, VUS, 절단점, AUC, 위험, 오분류율, 정분류율, 판별.

[1]교신저자: (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 교수.
E-mail: cshong@skku.edu