Journal of the Korean Data & Information Science Society
2014, **25**(3), 473–483

http://dx.doi.org/10.7465/jkdi.2014.25.3.473
한국데이터정보과학회지

ROC 다면체 아래 체적의 판단기준

홍종선[1] · 정동근[2]

[12]성균관대학교 통계학과

접수 2014년 3월 25일, 수정 2014년 4월 17일, 게재확정 2014년 4월 28일

요 약

ROC 곡선과 ROC 곡면을 확장한 4차원 이상의 공간에서의 ROC 다면체는 시각적인 표현이 어렵기 때문에 활용하기 어려우나, ROC 다면체 아래 공간을 측정하는 HUM 통계량에 대하여는 AUC와 VUS 통계량을 기반으로 정의가 가능하고 값을 구할 수 있으므로 본 연구는 네 가지 범주의 분류모형의 판별력을 측정하는 확률을 정의하고 연구한다. 그리고 Basel II를 기반한 부도확률에 대한 AUC의 판별력 판단기준을 제안한 연구를 확장하여, 네 범주 분류모형의 판별력을 측정하는 HUM 통계량에 관한 판단기준을 13단계로 구분하여 제안하고 활용하는 방법을 설명한다. 다양한 분포함수에 대하여 얻은 HUM 값을 바탕으로 제안한 판단기준을 탐색하기 위하여 삼원구획그림을 활용하여 판단기준을 설명한다.

주요용어: 부도, 신용평가, 위험, 절단점, 판단기준, 판별력.

1. 서론

국제결제은행 (BIS; Bank for International Settlement)에서는 금융기관이 보유하고 있는 익스포져 (exposure)를 신용, 시장, 운영 리스크로 구분하고 은행의 파산을 막기 위하여 2004년에 국제적인 위험 (risk) 관리 제도를 표준화하여 모형화하고 특정화하도록 하는 신BIS협약 (Basel Ⅱ)을 발표하였다. Joseph (2005)은 Basel II에 의한 부도확률 (probability of default)에 대해 ROC (receiver operating characteristic) 곡선을 통해 판별력을 측정하는 통계량으로 ROC 곡선의 아래 면적을 계산한 AUC (area under the ROC curve)에 대한 판별력 판단기준을 제안하였다. Joseph (2005)가 제안한 판별력 판단기준은 Wilkie (2004)의 방법을 확장한 것으로, 신용평가모형에서 부도 (불량)와 정상 모형의 분포함수가 동일한 표준편차를 갖는 정규분포 가정 하에서 모평균 차이에 대응하는 AUC의 판단기준을 제안하였다.

신용평가모형에서 많이 사용하는 두 가지 범주인 정상과 부도에 의한 판별력 측정은 보다 많은 다항 범주 (예를 들어 삼항 범주인 경우는 정상, 위험, 부도)에 대하여 모형의 판별력을 측정할 필요성이 있다. 세 종류의 범주형태로 분류되는 상황에 대하여 Scurfield (1996), Mossman (1999), Dreiseitl 등 (2000), Heckerling (2001), Fawcett (2003), Nakas와 Yiannoutsos (2004), Nakas 등 (2010), Patel과 Markey (2005), Zou 등 (2007), Wandishin과 Mullen (2009)은 이차원의 ROC 곡선을 삼차원으로 확장한 ROC 곡면 (surface)과 ROC 곡선에서의 AUC 통계량에 대응하는 ROC 곡면에서의 VUS (volume under the ROC surface) 통계량에 대해 정의하고 연구하였다. 그리고 Joseph (2005)의 연구 방법을

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

확장하여 Hong 등 (2013)은 삼차원 ROC 곡면에서의 VUS에 대하여 13단계의 판별력 판단기준을 제안하였다. 13단계의 판단기준에서 VUS 통계량의 범위를 제시하고 각 단계서 VUS 통계량에 대응하는 AUC 통계량, K-S (Kolmogorov - Smirnov) 통계량 그리고 세 분포의 모평균 차이들에 대한 값의 범위를 탐색하고 이 통계량들의 크기 관계를 살펴봄으로써 VUS 통계량의 판별력 판단기준을 설정하였다.

이차원에서 두 가지 범주의 판별에 대한 ROC 곡선과 AUC 통계량 그리고 삼차원에서 세 가지 범주의 판별에 대한 ROC 곡면과 VUS 통계량을 확장하여, Li와 Fine (2008)은 네 범주 이상의 판별에 대하여는 ROC 다면체 (manifold)라고 하고 ROC 다면체 아래의 공간에 대한 통계량을 HUM (hypervolume under the ROC manifold) 통계량으로 정의하고 추론을 제안하였다. ROC 곡선은 2차원 평면에, ROC 곡면은 3차원 공간에 나타낼 수 있지만 네 범주 이상은 시각적인 표현이 불가능할 뿐만 아니라, AUC와 VUS와 같이 면적과 체적으로 표현하는데 어려움이 존재한다. 본 연구에서는 두가지 범주에 관한 AUC와 세 가지 범주에 관한 VUS의 정의를 확장하여, 네 가지 범주 분류모형의 판별력을 측정하는 통계량을 제안한다. 일반적인 다항범주에 대하여 정의한 HUM 통계량을 네 범주로 제한하여 HUM^4 통계량으로 표기하고 HUM^4 통계량의 정의와 성질에 대하여 구체적이고 상세하게 설명하고 논의한다. HUM^4과 VUS와의 관계를 탐색하면서 Joseph (2005)와 Hong 등 (2013)의 연구방법을 바탕으로 HUM^4 통계량에 관한 판단기준을 제안하는 것을 연구목적으로 한다.

본 논문의 구성은 다음과 같다. 2절에서는 네 가지 범주 분류모형의 판별력을 측정하는 확률인 HUM^4 통계량을 정의하고, 이산형과 연속형 확률변수인 경우에 HUM^4 통계량을 구하는 방법을 각각 설명한다. 그리고 3차원의 VUS 통계량과 4차원의 HUM^4 통계량과의 관계에 대해 살펴본다. 3절에서는 Joseph (2005)와 Hong 등 (2013)의 연구방법을 확장한 HUM^4 통계량에 대한 판단기준을 제안한다. 이를 위하여 독립적인 네 종류의 분포함수를 다양하게 설정하고 이에 대응하는 HUM^4 값을 구하고 이를 바탕으로 13단계의 판단기준을 제안한다. 각 단계에서의 판단범위 (validation range)를 설정하고 각 판단범위에서의 HUM^4 값에 대응하는 모평균 차이와의 관계를 삼원구획그림 (ternary plot)을 이용하여 탐색하고 설명한다. 그리고 설정한 네 종류의 분포함수를 보다 다양하게 설정하고 이런 경우에 대하여 HUM^4 통계량의 활용에 대하여 살펴본다. 마지막으로 4절에서는 본 연구의 결과에 대해 정리하여 결론을 유도하고 향후 연구과제에 대해 토론한다.

2. HUM 통계량

독립적인 네 종류의 확률변수 X_1, X_2, X_3, X_4를 고려하자. 각각의 확률변수의 누적분포함수를 $F_i(\cdot)$로 표기하고, 임의의 x에 대하여 $F_1(x) \geq F_2(x) \geq F_3(x) \geq F_4(x)$를 가정한다.

우선 두 확률변수 X_1과 X_2에 대하여 ROC 곡선은 임의의 절단점 (cut-off, threshold) c에 대하여 $(F_2(c),\ F_1(c))$의 좌표를 각각 X축과 Y축 좌표에 대응시켜 단위 1의 정사각형의 그래프로 나타내거나, $u_1 \in (0,\ 1)$에 대하여 $(u_1,\ F_1(F_2^{-1}(u_1)))$로 표현한 곡선이다. ROC 곡선의 아래 면적인 AUC는 확률 $AUC = P(X_1 \leq X_2)$로 나타낸다. 확률변수가 이산형인 경우와 연속형인 경우에 AUC는 각각 다음과 같이 구한다.

$$AUC \quad = \quad P(X_1 < X_2) + \frac{1}{2}P(X_1 = X_2) = \int_0^1 ROC(u_1)du_1,$$

여기서 $ROC(u_1) = F_1(F_2^{-1}(u_1))$ 이다. AUC 값의 범위는 (1/2!, 1)로 Hosmer (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 두 분포가 정규분포를 따른다는 가정 하에서 AUC를 바탕으로 하는 판단기준을 제안하였다.

세 종류의 확률변수 X_1, X_2, X_3에 대하여 Scurfield (1996)는 두 개의 절단점 $c_1,\ c_2$ $(c_1 \le c_2)$인 경우에 대해 ROC 곡면을 정의하였다. ROC 곡면은 ROC 곡선과는 달리 세가지 정분류를 각각 X축과 Y축 그리고 Z축 좌표에 대응시켜 단위 1의 정육면체안의 3차원 곡면으로 표현하는데 임의의 절단점 $c_1,\ c_2$에 대하여 $(F_1(c_1),\ F_2(c_2) - F_2(c_1),\ 1 - F_3(c_2))$로 구현한 곡면이다. 그리고 VUS를 3차원으로 표현되는 ROC 곡면의 아래 면적을 나타내며 $VUS = P(X_1 \le X_2 \le X_3)$로 정의한다 (Nakas와 Yiannoutsos, 2004). 확률변수가 이산형인 경우에는 다음과 같이 표현한다.

$$\begin{aligned} VUS &= P(X_1 < X_2 < X_3) + \frac{1}{2}P(X_1 = X_2 < X_3) \\ &\quad + \frac{1}{2}P(X_1 < X_2 = X_3) + \frac{1}{6}P(X_1 = X_2 = X_3). \end{aligned}$$

그리고 연속형 확률변수인 경우에는 다음과 같이 두 종류의 적분으로 표현되어 값을 구할 수 있다.

$$\begin{aligned} VUS &= \int_0^1 \int_0^{F_1(F_3^{-1}(1-u_3))} ROC_s(u_1, u_3)\ du_1 du_3 \\ &= \int_0^1 F_1(F_2^{-1}(u))[1 - F_3(F_2^{-1}(u))]du, \end{aligned}$$

여기서 $ROC_s(u_1, u_3) = F_2(F_3^{-1}(1-u_3)) - F_2(F_1^{-1}(u_1))$, $F_2(x) = u$이다. VUS의 범위는 $(1/3!,\ 1)$이다. Hong 등 (2013)은 확률변수 X_1, X_2, X_3이 표준정규분포를 따른다는 가정 하에서 Joseph (2005)의 판단기준을 확장하여 VUS 통계량의 판단기준을 제시하였으며 13단계의 분류기준을 제안하였다.

본 연구에서는 네 종류의 확률변수 X_1, X_2, X_3 그리고 X_4를 고려하자. AUC와 VUS의 확장으로 확률 $P(X_1 \le X_2 \le X_3 \le X_4)$에 대하여 연구한다. 우선 3차원 공간의 ROC 곡면을 4차원으로 확장하기 위하여 임의의 절단점 $c_1,\ c_2,\ c_3$ $(c_1 \le c_2 \le c_3)$에 대한 4차원 좌표 $(F_1(c_1), F_2(c_2) - F_2(c_1), F_3(c_3) - F_3(c_2), 1 - F_4(c_3))$로 정의할 수 있으나, 이를 기하학적으로 표현할 수 없다. 그러나 이차원의 ROC 곡선 아래의 면적을 나타내는 확률 $AUC = P(X_1 \le X_2)$과 3차원의 ROC 곡면 아래의 부피를 나타내는 확률 $VUS = P(X_1 \le X_2 \le X_3)$를 구하는 방법을 확장하여, 4차원에서의 확률 $P(X_1 \le X_2 \le X_3 \le X_4)$를 구하고자 한다. Li와 Fine (2008)은 일반적인 다항범주에 대하여 HUM 통계량을 정의하였는데 본 연구에서는 네 범주로 제한한 HUM 통계량인 HUM^4에 관하여 설명하고 추론한다. 즉 네 가지 분류모형의 판별력을 측정하는 확률 $P(X_1 \le X_2 \le X_3 \le X_4)$을 HUM^4로 표기하고 정의 2.1에서 구체적으로 정의한다.

정의 2.1 독립적인 확률변수 X_1, X_2, X_3, X_4를 고려하자. 각각의 누적분포함수 $F_i(\cdot)$로 표기하고 임의의 x에 대하여 $F_1(x) \ge F_2(x) \ge F_3(x) \ge F_4(x)$를 가정할 때, 네 가지 범주 분류모형의 판별력을 측정하는 확률 $P(X_1 \le X_2 \le X_3 \le X_4)$을 HUM^4로 표기하고 다음과 같이 정의한다.

$$HUM^4 = P(X_1 \le X_2 \le X_3 \le X_4).$$

네 확률변수가 이산형인 경우와 연속형인 경우에 HUM^4 를 구하는 방법을 정리 2.1과 정리 2.2에서 각각 표현한다.

정리 2.1 이산형 확률변수인 경우에 HUM^4은 다음과 같이 표현한다.

$$\begin{aligned} HUM^4 = \ & P(X_1 < X_2 < X_3 < X_4) + \frac{1}{2}P(X_1 = X_2 < X_3 < X_4) + \frac{1}{2}P(X_1 < X_2 = X_3 < X_4) \\ & + \frac{1}{2}P(X_1 < X_2 < X_3 = X_4) + \frac{1}{4}P(X_1 = X_2 < X_3 = X_4) + \frac{1}{6}P(X_1 = X_2 = X_3 < X_4) \\ & + \frac{1}{6}P(X_1 < X_2 = X_3 = X_4) + \frac{1}{24}P(X_1 = X_2 = X_3 = X_4). \end{aligned}$$

증명: 각 변수들의 관계는 $\{X_1 < X_2 < X_3 < X_4\}$인 경우는 한가지이며, $\{X_1 = X_2 < X_3 < X_4$, $X_1 < X_2 = X_3 < X_4\}$ 그리고 $\{X_1 < X_2 < X_3 = X_4\}$인 경우는 두 종류가 있으므로 각각의 확률에 1/2를 곱해준다. 그리고 $\{X_1 = X_2 < X_3 = X_4\}$인 경우는 네 종류가 있으므로 이 경우의 확률에 1/4을 곱하고, $\{X_1 < X_2 = X_3 = X_4\}$와 $\{X_1 = X_2 = X_3 < X_4\}$인 경우는 3!의 종류가 있으므로 각각의 확률에 1/6을 곱해준다. 마지막 $\{X_1 = X_2 = X_3 = X_4\}$인 경우는 4!의 종류가 있으므로 이 경우의 확률에 1/24를 곱하여 각 변수들의 경우로 기대되는 결과를 계산한 $\{X_1 \le X_2 \le X_3 \le X_4\}$의 확률이 HUM^4이다. □

정리 2.2 독립적인 확률변수가 모두 연속형인 경우에 HUM^4는 다음과 같이 표현된다.

$$HUM^4 = \int_0^1 \int_0^{F_2(F_3^{-1}(u_3))} F_1(F_2^{-1}(u_2))[1 - F_4(F_3^{-1}(u_3))]du_2du_3 \ .$$

증명:

$$\begin{aligned} HUM^4 &= \iiint \int_{x_1<x_2<x_3<x_4} dF_1(x_1)dF_2(x_2)dF_3(x_3)dF_4(x_4) \\ &= \iint_{x_2<x_3} [\int_{x_3}^{\infty} dF_4(x_4) \int_{-\infty}^{x_2} dF_1(x_1)]dF_2(x_2)dF_3(x_3) \\ &= \iint_{x_2<x_3} F_1(x_2)[1 - F_4(x_4)]dF_2(x_2)dF_3(x_3) \\ &= \int_0^1 \int_0^{F_2(F_3^{-1}(u_3))} F_1(F_2^{-1}(u_2))[1 - F_4(F_3^{-1}(u_3))]du_2du_3, \end{aligned}$$

여기서 $F_2(x_2) = u_2$, $F_3(x_3) = u_3$이며, $x_2 < x_3$이므로 $F_2^{-1}(u_2) < F_3^{-1}(u_3)$가 되며 $0 < u_2 < F_2(F_3^{-1}(u_3))$로 표현되어 적분 범위에 반영된다. 이를 이용하여 4중 적분으로 HUM^4을 구한다. □

확률변수가 연속형일 때 정리 2.2를 통해 HUM^4을 계산하고 HUM^4 값의 범위는 (1/4!, 1)이다. AUC 통계량과 VUS 통계량의 관계 (Scurfield, 1996; Hong 등 2013)를 확장하여 VUS와 HUM^4의 관계를 정리 2.3에서 설명한다.

정리 2.3 4차원의 HUM^4와 3차원의 VUS의 관계는 다음과 같다.

$$\begin{aligned} VUS_{123} &= {HUM^4}_{1234} + {HUM^4}_{1243} + {HUM^4}_{1423} + {HUM^4}_{4123} \\ VUS_{124} &= {HUM^4}_{1243} + {HUM^4}_{1234} + {HUM^4}_{1324} + {HUM^4}_{3124} \\ VUS_{134} &= {HUM^4}_{1342} + {HUM^4}_{1324} + {HUM^4}_{1234} + {HUM^4}_{2134} \\ VUS_{234} &= {HUM^4}_{2341} + {HUM^4}_{2314} + {HUM^4}_{2134} + {HUM^4}_{1234}, \end{aligned}$$

여기서 $VUS_{ijk} = P(X_i \le X_j \le X_k)$ ${HUM^4}_{ijkl} = P(X_i \le X_j \le X_k \le X_l)$, $i, j, k, l = 1, 2, 3, 4$로 표기한다.

증명: 첫번째 경우인 $VUS_{123} = {HUM^4}_{1234} + {HUM^4}_{1243} + {HUM^4}_{1423} + {HUM^4}_{4123}$에 대하여 증명하면 다음과 같으며 나머지 관계의 증명은 생략한다.

$$
\begin{aligned}
& HUM_{1234} + HUM_{1243} + HUM_{1423} + HUM_{4123} \\
= \; & P(X_1 < X_2 < X_3 < X_4) + \frac{1}{2}P(X_1 = X_2 < X_3 < X_4) + \frac{1}{2}P(X_1 < X_2 = X_3 < X_4) + \frac{1}{2}P(X_1 < X_2 < X_3 = X_4) \\
& + \frac{1}{4}P(X_1 = X_2 < X_3 = X_4) + \frac{1}{6}P(X_1 = X_2 = X_3 < X_4) + \frac{1}{6}P(X_1 < X_2 = X_3 = X_4) + \frac{1}{24}P(X_1 = X_2 = X_3 = X_4) \\
& + P(X_1 < X_2 < X_4 < X_3) + \frac{1}{2}P(X_1 = X_2 < X_4 < X_3) + \frac{1}{2}P(X_1 < X_2 = X_4 < X_3) + \frac{1}{2}P(X_1 < X_2 < X_4 = X_3) \\
& + \frac{1}{4}P(X_1 = X_2 < X_4 = X_3) + \frac{1}{6}P(X_1 = X_2 = X_4 < X_3) + \frac{1}{6}P(X_1 < X_2 = X_4 = X_3) + \frac{1}{24}P(X_1 = X_2 = X_4 = X_3) \\
& + P(X_1 < X_4 < X_2 < X_3) + \frac{1}{2}P(X_1 = X_4 < X_2 < X_3) + \frac{1}{2}P(X_1 < X_4 = X_2 < X_3) + \frac{1}{2}P(X_1 < X_4 < X_2 = X_3) \\
& + \frac{1}{4}P(X_1 = X_4 < X_2 = X_3) + \frac{1}{6}P(X_1 = X_4 = X_2 < X_3) + \frac{1}{6}P(X_1 < X_4 = X_2 = X_3) + \frac{1}{24}P(X_1 = X_4 = X_2 = X_3) \\
& + P(X_4 < X_1 < X_2 < X_3) + \frac{1}{2}P(X_4 = X_1 < X_2 < X_3) + \frac{1}{2}P(X_4 < X_1 = X_2 < X_3) + \frac{1}{2}P(X_4 < X_1 < X_2 = X_3) \\
& + \frac{1}{4}P(X_4 = X_1 < X_2 = X_3) + \frac{1}{6}P(X_4 = X_1 = X_2 < X_3) + \frac{1}{6}P(X_4 < X_1 = X_2 = X_3) + \frac{1}{24}P(X_4 = X_1 = X_2 = X_3) \\
= \; & P(X_1 < X_2 < X_3 < X_4) + P(X_1 < X_2 < X_4 < X_3) + P(X_1 < X_4 < X_2 < X_3) + P(X_4 < X_1 < X_2 < X_3) \\
& + P(X_1 < X_2 < X_3 = X_4) + P(X_1 < X_2 = X_4 < X_3) + P(X_1 = X_4 < X_2 < X_3) + \frac{1}{2}P(X_1 = X_2 < X_3 < X_4) \\
& + \frac{1}{2}P(X_1 = X_2 < X_3 = X_4) + \frac{1}{2}P(X_1 = X_2 < X_4 < X_3) + \frac{1}{2}P(X_1 = X_2 = X_4 < X_3) + \frac{1}{2}P(X_4 < X_1 = X_2 < X_3) \\
& + \frac{1}{2}P(X_1 < X_2 = X_3 < X_4) + \frac{1}{2}P(X_1 < X_2 = X_3 = X_4) + \frac{1}{2}P(X_1 < X_4 < X_2 = X_3) + \frac{1}{2}P(X_1 = X_4 < X_2 = X_3) \\
& + \frac{1}{2}P(X_4 < X_1 < X_2 = X_3) + \frac{1}{6}P(X_1 = X_2 = X_3 < X_4) + \frac{1}{6}P(X_4 < X_1 = X_2 = X_3) + \frac{1}{6}P(X_1 = X_2 = X_3 = X_4) \\
= \; & P(X_1 < X_2 < X_3) + \frac{1}{2}P(X_1 = X_2 < X_3) + \frac{1}{2}P(X_1 < X_2 = X_3) + \frac{1}{6}P(X_1 = X_2 = X_3) = VUS_{123}.
\end{aligned}
$$

□

AUC 통계량과 VUS 통계량의 관계 그리고 정리 2.3에서 살펴본 VUS 통계량과 HUM^4 통계량의 관계를 이용하여 AUC와 HUM^4의 관계도 쉽게 파악할 수 있다.

3. HUM 판단기준

HUM^4의 통계량값이 1에 가까울수록 판별력이 좋은 모형이며 1/4!=0.0417일 때 Random 모형이라고 한다. 본 연구에서는 HUM^4 값의 범위에서 단계별 기준을 설정하고 모형의 판별력에 대한 의미를 부여하는 기준을 각 단계별로 제안하고자 한다.

2차원의 AUC 통계량의 판단기준을 제안한 Joseph (2005)의 연구와 3차원의 VUS 통계량의 판단기준을 제안한 Hong 등 (2013)의 연구를 확장하여, 4차원에서의 HUM^4 통계량의 판단기준을 설정한다. 우선 X_1, X_2, X_3, X_4의 분포함수를 Joseph (2005)와 Hong 등 (2013)이 사용한 분포함수와 유사하게 각각의 분포함수의 모평균을 μ_i 그리고 모분산을 1로 가정한 정규분포인 $F_i(x) \equiv \Phi(x; \mu_i, 1)$로 설정한다. $i+1$번째 정규분포와 i번째 분포의 모평균차이를 $\delta_i = \mu_{i+1} - \mu_i$로 표기하고, 각각의 δ_i를 0부터 9까지 0.25 간격으로 설정하여 자료를 생성하였으며 분포함수의 가정인 $F_1(x) \geq F_2(x) \geq F_3(x) \geq F_4(x)$의 성질을 유지한다. 그리고 모평균 $\mu_1, \mu_2, \mu_3, \mu_4$들의 전체평균이 0이 되도록 설정하여 $(\mu_1, \mu_2, \mu_3, \mu_4)$가 (0, 0, 0, 0)부터 (-13.5, -4.5, 4.5, 13.5)까지 나타내었다. 판단범위 (validation range)의 기준을 설정하기 위하여 동일한 모평균차이 $\delta_1 = \delta_2 = \delta_3 \equiv \delta_0$를 0부터 3까지 0.25간격으로 세분하였으며 따라서 모평균이 (0, 0, 0, 0)부터 (-4.5, -1.5, 1.5, 4.5)까지의 13단계로 설정하였다. 이러한 판단범위의 기준을 Table 3.1에 정리하였다.

Table 3.1 Standard criteria of HUM

Validation range	Meaning	μ_1	μ_2	μ_3	μ_4	δ_0	HUM^4
0 - 1	Random	0	0	0	0	0	0.0417
1 - 2	Doubtful	-0.375	-0.125	0.125	0.375	0.25	0.0890
2 - 3	Poor	-0.75	-0.25	0.25	0.75	0.5	0.1621
3 - 4	Marginal	-1.125	-0.375	0.375	1.125	0.75	0.2583
4 - 5	Satisfactory	-1.5	-0.5	0.5	1.5	1	0.3693
5 - 6	Good	-1.875	-0.625	0.625	1.875	1.25	0.4840
6 - 7	Very Good	-2.25	-0.75	0.75	2.25	1.5	0.5930
7 - 8	Strong	-2.625	-0.875	0.875	2.625	1.75	0.6897
8 - 9	Very Strong	-3	-1	1	3	2	0.7709
9 - 10	Excellent	-3.375	-1.125	1.125	3.375	2.25	0.8359
10 - 11	Excellent	-3.75	-1.25	1.25	3.75	2.5	0.8859
11 - 12	Excellent	-4.125	-1.375	1.375	4.125	2.75	0.9229
12 - 13	Superior	-4.5	-1.5	1.5	4.5	3	0.9494

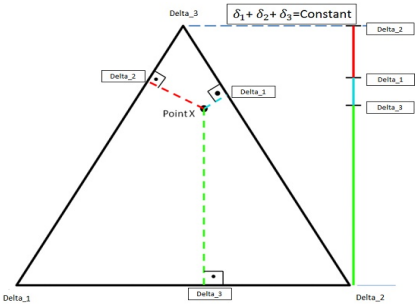

Figure 3.1 Ternary plot ($\mu_4 - \mu_1 = 3$)

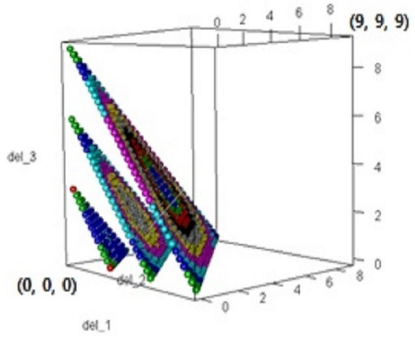

Figure 3.2 3D plot by $\delta_1 + \delta_2 + \delta_3 = 3, 6, 9$

Table 3.1을 통해 HUM^4 통계량 값이 0.0417 (=1/24)부터 0.0890까지인 경우에는 설정한 네 모형이 모두 동일하여 모형의 판별을 할 수 없으므로 'Random'하다고 해석하고 판단범위 (validation range)는 0-1단계로 한다. HUM^4 통계량 값이 0.9229부터 0.9494일 때는 네 모형을 완벽하게 판별할 수 있는 'Superior'라고 의미를 부여하고 판단범위를 12-13단계로 제안한다. 따라서 HUM^4 값이 주어지면, Table 3.1을 바탕으로 판별 수준을 결정하는 판단범위를 선정하면서 네 모형에 관한 판별 수준의 의미를 설명할 수 있는 기준을 제안한다.

동일한 모평균차이 δ_0가 선형적으로 증가할 때, 판단범위의 단계가 상승하고 HUM^4값도 증가하지만 비선형으로 증가한다. 세 종류의 모평균차이 δ_i가 각각 증가할수록 HUM^4값의 관계를 Figure 3.2에 구현하였다. 3차원 육면체로 나타나면서 HUM^4값이 비선형이므로 본 연구에서는 세 종류의 모평균차이 δ_i와 HUM^4의 관계를 살펴보기 위하여 삼원구획그림 (ternary plot)을 이용한다 (Zachos, 2012). 3차원 자료를 2차원으로 표현할 수 있는 삼원구획그림은 색 또는 점 형태를 달리하여 네 개의 변수를 2차원 평면에 표현할 수도 있지만 세 개의 기준변수의 합 (여기에서는 $\delta_1 + \delta_2 + \delta_3$)이 일정하다는 제약이 있다.

Figure 3.1의 삼원구획그림에서 $\delta_1 + \delta_2 + \delta_3$의 값이 일정한 경우에 대해 δ_i의 변화에 대해 나타낼 수 있다. δ_i가 등간격일 때에는 정삼각형인 삼원구획그림의 중심에 위치하며 삼원구획그림안에 위치한 점 (point x)부터 삼각형의 각각의 변까지의 직각을 이루는 수선의 길이는 $\delta_1 + \delta_2 + \delta_3$의 크기와 같으며, 이는 $\mu_4 - \mu_1$과 같다. 이 관계는 삼각형의 넓이를 이용하여 쉽게 나타낼 수 있다. 세 기준

변수의 합 $\sum_{i=1}^{3}\delta_i$는 $\mu_4-\mu_1$이므로 네 분포함수들의 평균들 중에서 가장 큰 평균과 가장 작은 평균의 거리이다. 따라서 고정된 합 $\sum_{i=1}^{3}\delta_i$의 의미는 네 종류의 분포함수들의 산포도 (dispersion)를 측정한다고 파악할 수 있으므로 세 기준변수의 합이 일정한 경우 몇 가지를 살펴본다. Figure 3.2에서 $\mu_4-\mu_1=\delta_1+\delta_2+\delta_3$이 3, 6, 9일 경우에 해당하는 HUM^4을 삼원구획그림으로 구현하여 Figure 3.3부터 3.5에 나타내었다. 원점에 가까운 삼각형이 $\mu_4-\mu_1=3$이고 가장 큰 삼각형이 $\mu_4-\mu_1=9$인 경우이다.

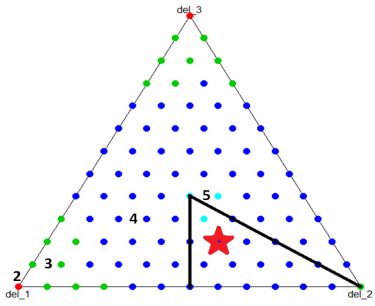

Figure 3.3 Ternary plot ($\mu_4-\mu_1=3$)

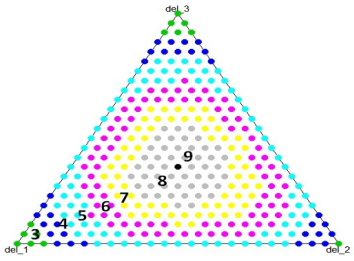

Figure 3.4 Ternary plot ($\mu_4-\mu_1=6$)

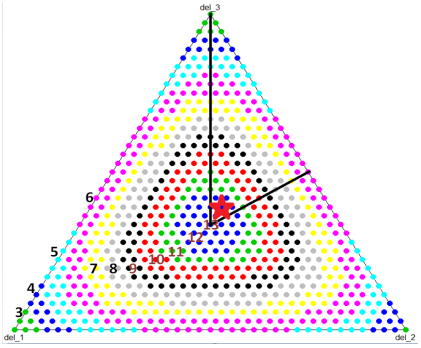

Figure 3.5 Ternary plot ($\mu_4-\mu_1=9$)

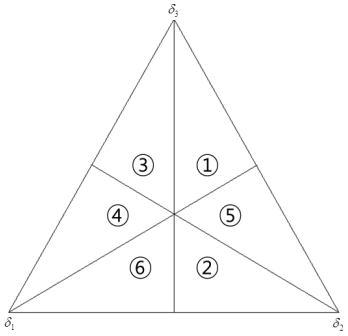

Figure 3.6 Relation of δ_i in ternary plot

삼원구획그림에서의 모평균차이 δ_i 사이의 관계를 나타내기 위하여 Figure 3.6을 고려한다. 각 꼭지점에서 마주보는 변에 직각인 수선들이 만나는 점은 모평균차이들이 모두 동일한 $\delta_1=\delta_2=\delta_3$인 지점이 되며, Figure 3.6과 같이 세 종류의 수선들로 삼원구획그림을 여섯 가지의 직각삼각형으로 나눌 수 있다. 각 삼각형에서의 δ_i들의 크기 관계를 표현할 수 있는데, (1)번 삼각형에 존재하는 점들은 $\delta_1<\delta_2<\delta_3$의 크기일 때이며, (2)번 삼각형은 $\delta_3<\delta_1<\delta_2$, (3)번 삼각형에 존재하는 점들은 $\delta_2<\delta_1<\delta_3$, (4)번 삼각형은 $\delta_2<\delta_3<\delta_1$, (5)번 삼각형은 $\delta_1<\delta_3<\delta_2$ 그리고 (6)번 삼각형에 존재하는 점들은 $\delta_3<\delta_2<\delta_1$인 관계를 갖고 있다. 이러한 관계는 모평균차이 δ_i와 관련된 모평균 μ_i와도 연관성이 있다.

Figure 3.3은 $\mu_4-\mu_1=3$인 경우이며, $\delta_0=1$일 때 가운데 점에 값이 위치하고 그때의 판단범위는 4-5이며 HUM^4 값은 0.3693이다. 예를 들어 $\mu_4-\mu_1=3$인 경우에 Figure 3.3의 삼원구획그림을 이용할 수 있는데 특히, $\delta_1=1$, $\delta_2=1.5$, $\delta_3=0.5$일 때는 $\delta_3<\delta_1<\delta_2$의 관계이므로 Figure 3.6의 (2)번 삼각형 내에 존재하며 Figure 3.3의 삼원구획그림에서 '$\star$' 표시된 점으로 나타나므로 판단범위는

3-4단계임을 파악할 수 있다. 그리고 Figure 3.5는 $\mu_4 - \mu_1 = 9$일 때의 삼원구획그림이며 판단범위가 2-3부터 12-13단계까지 나타난다. $\delta_1 = 2.5$, $\delta_2 = 3$, $\delta_3 = 3.5$인 경우에는 $\delta_1 < \delta_2 < \delta_3$의 관계이므로 Figure 3.6의 (1)번 삼각형 내에 존재하며 Figure 3.5의 삼원구획그림에서 '⋆' 표시된 점으로 나타나므로 판단범위는 11-12단계로 설명한다.

HUM^4값을 알고 있으면, Table 3.1을 통하여 판단범위를 선택할 수 있으며 네 모형의 판별 수준의 의미를 설명할 수 있다. 그러나 HUM^4값을 알고 있지 않는 경우라도 모평균들의 최대 범위 $\mu_4 - \mu_1$ 즉, 모평균차이들의 합 그리고 모평균차이들의 크기 순서 등으로 네 모형의 판별 수준을 결정하는 판단범위를 파악할 수 있고 판별 수준을 파악할 수 있다.

Table 3.2 Standard criteria of HUM

Validation range	δ_0	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$
0 - 1	0	0.0417	0.0417	0.0417
1 - 2	0.25	0.0890	0.0622	0.0547
2 - 3	0.5	**0.1621**	0.0890	0.0704
3 - 4	0.75	**0.2583**	0.1223	0.0890
4 - 5	1	**0.3693**	**0.1621**	0.1105
5 - 6	1.25	0.4840	0.2077	0.1349
6 - 7	1.5	0.5930	**0.2583**	**0.1621**
7 - 8	1.75	0.6897	0.3126	0.1919
8 - 9	2	0.7709	**0.3693**	0.2241
9 - 10	2.25	0.8359	0.4268	**0.2583**
10 - 11	2.5	0.8859	0.4840	0.2942
11 - 12	2.75	0.9229	0.5397	0.3313
12 - 13	3	0.9494	0.5930	**0.3693**

Table 3.1에서의 판단기준을 위해 분포함수의 표준편차 σ를 1로 설정하였다. Table 3.2에서 σ=2와 3으로 설정하였을 때 얻은 HUM^4 통계량값을 제시하였다. Table 3.2를 통하여 $\sigma = 1$에서의 HUM^4값은 σ의 증가 속도만큼 δ_0가 증가했을 때의 HUM^4값과 동일하다. 즉 판단범위 3-4단계에서 $\sigma = 1$일 때의 HUM^4값은 0.2583이고 이때의 δ_0는 0.75이다. $\sigma = 2$일 때 동일한 HUM^4값을 갖는 경우는 δ_0가 두배로 증가한 $\delta_0 = 1.5$이며 판단범위 6-7단계이고, $\sigma = 3$에서 동일한 HUM^4값인 경우에는 δ_0가 세 배로 증가한 $\delta_0 = 2.25$이며 판단범위 9-10단계임을 파악할 수 있다. 그리고 σ=1일 때 판단범위의 특정한 단계 (예를 들어 3-4단계)를 기준으로 전후의 한 단계 차이는 σ=2일 때는 판단범위 6-7단계의 전후의 두 단계 차이인 4-5와 8-9단계로 나타나며, σ=3일 때는 판단범위 3-4단계의 전후의 세 단계 차이인 6-7과 12-13단계로 나타남을 파악한다.

Table 3.3은 표준편차 $\sigma = 2$인 경우에서의 HUM^4 판단기준이다. Table 3.3에서 $\sigma = 2$인 경우에 각 판단범위에서의 HUM^4값은 Table 3.1과 동일하나, Table 3.3에서의 δ_0의 변화 속도는 Table 3.1에서의 δ_0의 변화 속도의 두배이다. 즉, Table 3.1에서 판단범위의 단계가 상승할 때마다 δ_0가 0.25만큼 증가하나, Table 3.3에서는 δ_0가 0.5만큼 증가한다. 따라서 Table 3.3에서 판단범위의 0-1단계로부터 단계가 상승할수록 μ_1의 값은 0에서 0.75만큼씩 감소하고 μ_4값은 0.75만큼씩 증가한다. 그리고 판단범위의 단계가 상승할수록 μ_2값은 0에서 0.25만큼씩 감소하고 μ_3값은 0.25만큼씩 증가한다. 즉 Table 3.3에서 μ_i들의 변화 속도는 Table 3.1의 속도보다 σ의 증가속도와 동일하게 두배로 증가함을 탐색할 수 있다. HUM^4의 판단기준은 모평균 μ의 변화뿐만 아니라 표준편차 σ의 변화도 함께 반영하여 고려할 수 있다.

분포함수들의 분산이 일정한 경우는 Table 3.1의 정보를 판단기준으로 활용할 수 있고, 분포함수들의 분산이 일정하지 않은 경우는 Table 3.2과 Table 3.3의 결과를 바탕으로 분포들의 표준편차의 크기에 따라 조정하여 판단기준을 설정한다.

Table 3.3 Standard criteria of HUM ($\sigma = 2$)

Validation range	Meaning	μ_1	μ_2	μ_3	μ_4	δ_0	HUM^4
0 - 1	Random	0	0	0	0	0	0.0417
1 - 2	Doubtful	-0.75	-0.25	0.25	0.75	0.5	0.0890
2 - 3	Poor	-1.5	-0.5	0.5	1.5	1	0.1621
3 - 4	Marginal	-2.25	-0.75	0.75	2.25	1.5	0.2583
4 - 5	Satisfactory	-3	-1	1	3	2	0.3693
5 - 6	Good	-3.75	-1.25	1.25	3.75	2.5	0.4840
6 - 7	Very Good	-4.5	-1.5	1.5	4.5	3	0.5930
7 - 8	Strong	-5.25	-1.75	1.75	5.25	3.5	0.6897
8 - 9	Very Strong	-6	-2	2	6	4	0.7709
9 - 10	Excellent	-6.75	-2.25	2.25	6.75	4.5	0.8359
10 - 11	Excellent	-7.5	-2.5	2.5	7.5	5	0.8859
11 - 12	Excellent	-8.25	-2.75	2.75	8.25	5.5	0.9229
12 - 13	Superior	-9	-3	3	9	6	0.9494

4. 결론

시각적으로 표현할 수 있는 이차원 ROC 곡선과 삼차원 ROC 곡면의 활용은 매우 유용한 통계적 방법이다. 네 범주 이상인 ROC 다면체는 시각적인 표현이 어렵기 때문에 활용하는데 한계가 있다. 그러나 ROC 곡선과 ROC 곡면의 아래 면적과 부피인 AUC와 VUS 통계량을 확장하여 ROC 다면체 아래 공간을 측정하는 HUM 통계량에 대하여는 정의가 가능하고 값을 구할 수 있으므로 본 연구는 ROC 다면체에 대응하는 HUM 통계량을 연구하였다.

네 가지 범주의 분류모형에 대하여 연구하기 위하여 ROC 통계량의 정의를 바탕으로 네 모형의 판별력을 측정하는 확률을 HUM^4로 표기하기로 하고 HUM^4에 대하여 연구하였다. HUM^4 통계량에 대하여 정의를 내리고 HUM^4를 계산하는 방법에 대하여 살펴보았으며, HUM^4과 VUS와의 관계에 대하여 탐색하였다.

Joseph (2005)와 Hong 등 (2013)의 연구방법을 확장하여 네 범주 분류모형의 판별력을 측정하는 HUM^4 통계량에 관한 판단기준을 제안하였다. 다양한 분포함수에 대한 HUM^4 통계량을 바탕으로 판단기준을 13단계로 구분하고 이에 대하여 탐색하면서 설명하였다. 우선 표준편차가 모두 1인 경우 모평균 차이들의 변화에 따라 HUM^4 통계량과의 관계를 살펴보기 위하여 삼원구획그림을 활용하여 설명하였다. 그리고 표준편차를 변경하였을 경우에는 모평균 차이들의 변화에 따라 HUM^4 통계량과의 관계가 표준편차의 변화량만큼의 선형관계가 존재함을 발견하였다.

본 연구는 ROC manifold 중에서 네 가지 범주의 분류모형에 대하여만 연구하였는데, 이를 확장하여 다섯 이상의 범주를 가진 분류모형에서의 판별력을 판단할 수 있는 기준을 설정할 수 있다. 5차원 이상의 HUM 통계량을 정의하기는 쉬우나 이를 구하는 방법은 쉽게 얻을 수 없으므로 이를 향후 연구과제로 남겨둔다. 그리고 본 연구에서 다룬 네 종류의 분포함수의 표준편차를 동일하게 설정했지만 현실 세계에서 활용할 수 있도록 각 분포별 표준편차를 다양하게 변경하여 HUM 통계량의 판단기준 설정과 이러한 판단기준을 네 가지 범주 분류모형의 판별의 실제 사례에 적용하는 현실 문제를 향후 연구과제로 남겨두기로 한다.

References

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.

Fawcett, T. (2003). *ROC graphs: Notes and practical considerations for data mining researchers*, HP Labs Tech Report HPL-2003-4, HP Laboratories, `http://www.hpl.hp.com/techreports/2003/`.

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

Hong, C. S., Jung, E. S. and Jung, D. G. (2013). Standardized criterion of VUS for ROC surface. *The Korean Journal of Applied Statistics*, **26**, 977-985.

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

Li, J. and Fine, J. P. (2008). ROC analysis with multiple classes and multiple tests: Methodology and its application in microarray studies. *Biostatistics*, **9**, 566-576.

Joseph, M. P. (2005). A PD validation framework for Basel II internal ratings-based systems. *Quantitative Analyst Basel II Project,* Commonwealth Bank of Australia.

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

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

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.

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.

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

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,* edited by L.C. Thomas, D.B. Edelman, and J.N. Crook, Oxford University Press, Oxford.

Zachos, C. K. (2012). Ternary plots for neutrino mixing visualization. ANL-HEP-PR-12-31, arXiv:1205.4772 v3 [hep-ph].

Zou, K. H., O'Malley, A. J. and Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, **115**, 654-657.

Journal of the Korean Data & Information Science Society
2014, **25**(3), 473–483

http://dx.doi.org/10.7465/jkdi.2014.25.3.473
한국데이터정보과학회지

Standard criterion of hypervolume under the ROC manifold

C. S. Hong[1] · D. G. Jung[2]

[12]Department of Statistics, Sungkyunkwan University

Received 25 March 2014, revised 17 April 2014, accepted 28 April 2014

Abstract

Even though the ROC manifold for more than three dimensional space which is an extension of the ROC curve and surface has difficulty to represent graphically, the hypervolume under the ROC manifold (HUM) statistic can be defined and obtained based on AUC and VUS measures for the ROC curve and the ROC surface. Hence the definition and characteristics of the HUM for four dimensional space are studied in this work. By extension of the standard criterion of AUC for probabilities of default based on Basel II, the 13 classes of standard criterion of HUM are proposed in order to discriminate four classification models and some application methods are discussed. In order to explore the standard criterion of HUM whose values are obtained from various distributions, ternary plot is used and explained.

Keywords: Credit evaluation, default, discriminant, risk, threshold point, validation range.

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