The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.2.187
32(2), 187–198

Partial AUC and optimal thresholds

Chong Sun Hong[a1] · Hyun Su Cho[a]

[a]Department of Statistics, Sungkyunkwan University

(Received February 11, 2019; Revised February 25, 2019; Accepted February 27, 2019)

Abstract

Extensive literature exists on how to estimate optimal thresholds based on various accuracy measures using receiver operating characteristic (ROC) and cumulative accuracy profile (CAP) curves. This paper now proposes an alternative measure to represented the specific partial area under the ROC and CAP curves. The relationship between ROC and CAP functions is examined using differential equations of the new defined partial area under curves. In addition, the relationship with the optimal thresholds under conditions of various accuracy measures for the ROC and CAP functions is also derived. We assume there are two kinds of distribution functions composing the mixed distribution as various normal distributions before finding the optimal thresholds. Corresponding type 1 and 2 errors are also explored and discussed under various conditions for accuracy measures.

Keywords: accuracy, classification, confusion matrix, default, optimal threshold

1. 서론

두 종류 함수의 혼합분포가 주어진 경우 '최적분류점(optimal threshold)'은 이에 대한 판별력(power of discrimination)을 극대화시켜주는 분류점(절단점; threshold, cut-off point)이다. 최적분류점을 추정하는 연구는 신호탐지 이론으로 시작하여 신용평가, 제약, 의학, 고객분류 등 다양한 분야에서 응용되고 있다 (Swets, 1988; Irwin와 Irwin, 2012; Dodd와 Pepe, 2003). 본 연구는 신용평가의 관점에서 두 종류로 구분할 수 있는 부도(default/bad/positive)와 정상(non-default/good/negative) 차주에 대한 혼합분포를 분류하는 연구에 대하여 설명하고자 한다.

두 종류로 구성된 부도와 정상에 대한 모수공간을 $\theta = \{\theta_d, \theta_n\}$로 가정하고, 임의의 실수 x에 대하여 각각의 분포에 대한 조건부 누적분포를 $F_d(x) \equiv P(X \le x|\theta_d)$, $F_n(x) \equiv P(X \le x|\theta_n)$로 정의하면 이에 대한 혼합분포는 다음과 같이 설정할 수 있다.

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

여기서 γ는 전체 부도율(total probability of default)로 n_d와 n_n가 각각 부도와 정상으로 판별된 표본수인 경우에 전체 부도율은 전체 차주에서 부도 차주의 비율인 $\gamma = n_d/(n_d + n_n)$로 추정한다. 또한 본 연구에서는 모든 x에 대해서 $F_d(x) \ge F_n(x)$를 가정한다.

[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-Ro, Jongno-Gu, Seoul 03063, Korea. E-mail: cshong@skku.edu

두 분포함수를 판별하는 임의의 분류점을 설정하면 주어진 데이터에 관하여 혼동행렬(confusion matrix)을 표현할 수 있으며 이를 기반으로 true positive rate (TPR, sensitivity)과 false positive rate (FPR, $1-$specificity)을 다음과 같이 정의한다 (Pepe, 2003; Tasche, 2006).

$$\text{TPR} = F_d(x), \qquad \text{FPR} = F_n(x).$$

Receiver operating characteristic (ROC) 곡선은 분류모형(classification model) 임의의 분류점 변화에 따른 TPR과 FPR간의 교환(trade-off) 관계를 이차원 평면에 나타낸 성과(performance) 기반의 시각적 도구이며 수평과 수직축을 $F_d(x)$, $F_n(x)$로 나타낸다 (Zweig과 Campbell, 1993; Fawcett와 Provost, 1997, 2001; Pepe, 2000; Fawcett, 2006). ROC 곡선은 정사각형 모양의 공간에서 나타나기 때문에 자료의 비대칭성을 시각적으로 확인하기 어려운데, 이러한 점을 보완할 수 있는 방법이 바로 cumulative accuracy profile (CAP) 곡선이며 (Berry와 Linoff, 1999; Sobehart 등, 2000), CAP 곡선의 수직축은 ROC 곡선과 동일하지만 수평축을 Alarm Rate라고 정의하는 $F(x)$로 표현한다.

$$(F_n(x), F_d(x)) = (p, \text{ROC}(p)), \qquad (F(x), F_d(x)) = (q, \text{CAP}(q)),$$

여기서 $\text{ROC}(p) = F_d(F_n^{-1}(p))$ 그리고 $\text{CAP}(q) = F_d(F^{-1}(q))$이다.

분류모형에서 제1종 오류와 제2종 오류 각각 또는 두 오류의 합을 최소화하는 최적분류점을 구하기 위한 다양한 정확도 측도(accuracy measures)들이 존재한다. 그 중에서 대표적인 측도들로는 maximum vertical distance (MVD) (Krzanowski와 Hand, 2009), Youden index (J) (Youden, 1950), the closest-to-$(0,1)$ criterion와 the amended closest-to-$(0,1)$ criterion (Perkins와 Schisterman, 2006), sum of sensitivity and specificity (SSS) (Connel과 Koepsell, 1985), symmetry point (Moses 등, 1993; Pepe, 2003), accuracy area (Brasil, 2010), total accuracy (TA) (Lambert와 Lipkovich, 2008), true rate (TR) (Velez 등, 2007; Hong 등, 2010) 등이 있는데 모든 종류의 정확도 측도들은 ROC와 CAP 함수 또는 TPR과 FPR 즉, $F_d(x)$와 $F_n(x)$로 정의된다.

ROC와 CAP 곡선 아래의 면적을 area under the curve (AUC)라고 정의하는데, 우선 ROC 곡선 아래의 면적을 나타내는 area under the ROC curve (AUROC)를 이를 확장한 CAP 곡선 아래의 면적인 area under the CAP curve (AUCAP)를 다음과 같이 각각 정의한다 (Bradley, 1997; Joseph, 2005; Krzanowski와 Hand, 2009).

$$\text{AUROC} = \int_0^1 \text{ROC}(p)dp, \qquad \text{AUCAP} = \int_0^1 \text{CAP}(q)dq.$$

AUROC는 0.5에서 1 사이의 값을 가지며, 1에 접근하는 값을 가질수록 우수한 분류모형으로 판단할 수 있으며 0.5 값을 갖는 모형은 random한 분류모형이라고 한다. AUCAP는 0.5보다 큰 값을 가지지만 1보다 작은 $1-\gamma/2$를 가지며 이에 접근하는 값 또는 정확율(the accurate ratio)을 가질수록 우수한 모형으로 판단 한다 (Engelmann 등, 2003; Vuk와 Curk, 2006). 곡선 아래의 전체 면적인 AUROC와 AUCAP 중에서 특정한 부분의 면적을 '부분 AUC(partial AUC)'로 정의하는데, ROC와 CAP 함수 중에서 p_0와 p_1 사이의 함수 아래 면적을 다음과 같이 각각 정의한다 (Jiang 등, 1996). 임의의 p_0와 p_1 $(0 < p_0 < p_1 < 1)$에 대하여,

$$\text{AUROC}(p_0, p_1) = \int_{p_0}^{p_1} \text{ROC}(p)dp, \qquad \text{AUCAP}(p_0, p_1) = \int_{p_0}^{p_1} \text{CAP}(q)dq. \tag{1.2}$$

부분 AUC는 전체 구간이 아닌 관심이 있고 외부 요인으로 인해 제약된 특정 구간에서 분류모형의 성능을 비교할 때 사용되며, 전체 구간에서의 분류모형 성능과 관심 있는 구간에서의 성능에 차이가 존재

할 수 있기 때문에 제약, 의학 분야 등 다양한 분야에서 응용된다 (Centor, 1991; Swets등, 2000; Zou, 2002; Fawcett, 2003).

ROC와 CAP 곡선을 바탕으로 최적분류점을 추정하는 많은 연구 문헌 중에서 Hong과 Choi (2009)와 Hong 등 (2010)은 ROC와 CAP 함수들과 다양한 정확도 측도에 해당하는 일차원 접선의 접점을 활용하여 최적분류점을 추정하는 방법을 제안하였다. Yoo와 Hong (2011)은 많은 종류의 정확도 측도들 중에서 대표적인 아홉 가지 측도들을 설명하고 각 측도들의 관계를 정리하고, 정확도 측도가 포함하는 분류정확도의 조건함수를 유도하여 성격에 따라 네 종류의 범주로 구분하였다. 본 연구에서는 이러한 선행연구들을 확장하고 발전시켜 부분 AUC와 정확도 측도들의 최적분류점과의 관계를 유도한 후 미분방정식을 활용하여 최적분류점을 추정하는 방법을 제안한다.

ROC와 CAP 곡선 AUC의 특정 부분 면적을 나타내는 부분 AUC를 p_0와 p_1 사이로 정의한 식 (1.2)와 다르게 본 연구의 2절에서는 0부터 시작하는 부분에 대한 대안적인 부분 AUC를 제안한다. 그리고 대안적인 부분 AUC와 ROC와 CAP 함수와의 관계를 일차 미분방정식으로 유도하고, 대안적인 부분 AUC의 이차 미분방정식을 이용하여 ROC와 CAP 곡선에서의 최적분류점들과의 관계를 유도하여 다양한 최적분류점을 추정하기 위한 여러 조건으로 다시 정의한다. 3절에서는 부도와 정상의 분포함수를 정규분포로 가정하여 다양한 정규분포의 경우에서의 최적분류점을 유도한다. 4절에서는 여러 종류의 정규분포에 따른 최적분류점 그리고 제1종과 제2종 오류의 크기를 살펴본다. 마지막으로 5절에서는 본 연구를 종합하여 결론을 유도한다.

2. Partial AUC와 최적분류점

본 연구에서는 식 (1.2)에서 정의된 부분 AUC에서 p_0부터 p_1 사이 대신에 0부터 u까지의 부분 AUC인 $\mathrm{AUROC}(u)$와 $\mathrm{AUCAP}(u)$를 다음과 같이 대안적으로 정의한다.

정의 2.1 ROC와 CAP 함수에 대한 각각의 부분 AUC인 $\mathrm{AUROC}(u)$와 $\mathrm{AUCAP}(u)$를 다음과 같이 정의한다. 임의의 $u(0 < u < 1)$에 대하여.

$$\mathrm{AUROC}(u) = \int_0^u \mathrm{ROC}(p)dp,$$
$$\mathrm{AUCAP}(u) = \int_0^u \mathrm{CAP}(q)dq.$$

$\mathrm{AUROC}(u)$와 $\mathrm{AUCAP}(u)$를 정의 2.1과 같이 정의하면, ROC와 CAP 함수들과 다음과 같은 함수식을 유도할 수 있다.

정리 2.1 $\mathrm{AUROC}(u)$와 $\mathrm{AUCAP}(u)$의 일차 미분을 통해 다음과 같은 함수식을 각각 갖는다.

$$\frac{d}{du}\mathrm{AUROC}(u) = F_d\left(F_n^{-1}(u)\right),$$
$$\frac{d}{du}\mathrm{AUCAP}(u) = F_d\left(F^{-1}(u)\right).$$

$\mathrm{AUROC}(u)$와 $\mathrm{AUCAP}(u)$의 일차 미분 계수가 각각 $\mathrm{ROC}(p) = F_d(F_n^{-1}(p))$와 $\mathrm{CAP}(q) = F_d(F^{-1}(q))$이므로 쉽게 증명된다. 또한 $\mathrm{AUROC}(u)$와 $\mathrm{AUCAP}(u)$의 이차 미분 방정식을 통해 각각 다음과 같은 정리를 유도할 수 있다.

정의 2.2 $\mathrm{AUROC}(u)$와 $\mathrm{AUCAP}(u)$의 0차 미분을 통해 다음과 같은 ROC와 CAP 함수 각각에 대한 접선의 기

Table 2.1. The second derivatives of AUROC(u) and AUCAP(u)

AUROC(u)의 이차미분 계수	AUCAP(u)의 이차미분 계수	정확도 측도
1	1	TR, J, SSS, MVD, 수정된 $(0,1)$기준
$(1-\gamma)/\gamma$	$1/\gamma$	TA

AUROC = area under the receiver operating characteristic curve; AUCAP = area under the cumulative accuracy profile curve. TR = true rate; J = Youden index; SSS = sum of sensitivity and specificity; MVD = maximum vertical distance; TA = total accuracy.

울기인 확률밀도함수의 비율로 유도할 수 있다.

$$\frac{d^2}{du^2}\text{AUROC}(u) = \frac{f_d(F_n^{-1}(u))}{f_n(F_n^{-1}(u))},$$
$$\frac{d^2}{du^2}\text{AUCAP}(u) = \frac{f_d(F^{-1}(u))}{f(F^{-1}(u))}.$$

증명:

$$\begin{aligned}\frac{d^2}{du^2}\text{AUROC}(u) &= \frac{d}{du}F_d\left(F_n^{-1}(u)\right) = \frac{dF_n^{-1}(u)}{du}\frac{d}{dF_n^{-1}(u)}F_d\left(F_n^{-1}(u)\right)\\ &= \left(\frac{dF_n(x)}{dx}\right)^{-1}\left(\frac{dF_d(x)}{dx}\right)\\ &= \frac{f_d(x)}{f_n(x)} = \frac{f_d\left(F_n^{-1}(u)\right)}{f_n\left(F_n^{-1}(u)\right)},\end{aligned}$$

여기서 $F_n^{-1}(u) = x$ 그리고 $dF_n(x) = du$이다.

$$\begin{aligned}\frac{d^2}{du^2}\text{AUCAP}(u) &= \frac{d}{du}F_d\left(F^{-1}(u)\right) = \frac{dF^{-1}(u)}{du}\frac{d}{dF_n^{-1}(u)}F_d\left(F^{-1}(u)\right)\\ &= \left(\frac{dF(x)}{dx}\right)^{-1}\left(\frac{dF_d(x)}{dx}\right)\\ &= \frac{f_d(x)}{f(x)} = \frac{f_d\left(F^{-1}(u)\right)}{f\left(F^{-1}(u)\right)},\end{aligned}$$

여기서 $F^{-1}(u) = x$ 그리고 $dF(x) = du$이다. □

Yoo와 Hong (2011)은 미분 가능한 연속형 확률밀도함수들의 조건에 따라 모두 일곱 종류의 범주를 생성하였고 다양한 정확도 측도들을 총 네개의 범주로 분류하였는데 본 연구의 정리 2.2를 바탕으로 AUROC(u)와 AUCAP(u)의 이차 미분값이 특정한 상수일 때 정확도 측도에 대한 조건을 Table 2.1과 같이 정리하였다. Table 2.1을 통해 해당 정확도 측도들에 따른 최적분류점을 구할 수 있으며 이는 분류모형의 전체 구간에서 두 집단을 분류하는 최적의 값이라고 할 수 있다. 따라서 AUROC(u)와 AUCAP(u)의 이차 미분값이 즉, ROC(u)와 CAP(u)의 일차 미분값이 1인 경우에 해당하는 x값은 TR, J, SSS, MVD, 수정된 $(0,1)$기준에 의한 최적분류점(the amended closest-to-$(0,1)$ criterion) x_0이며, AUROC(u)와 AUCAP(u)의 이차 미분값이 각각 $(1-\gamma)/\gamma = n_n/n_d$와 $1/\gamma = (n_d+n_n)/n_d$인 경우 해당하는 x값은 TA에 의한 최적분류점으로 구할 수 있다, 여기서 n_d와 n_n은 부도와 정상분포의 표본수이다.

3. 정규분포에서의 최적분류점

부도와 정상의 두 조건부 확률밀도함수 $f_d(x)$와 $f_n(x)$를 모평균과 모분산이 각각 다른 정규분포 $f_d(x) = \phi(x|\mu_d, \sigma_d)$와 $f_n(x) = \phi(x|\mu_n, \sigma_n)$로 설정하여, AUROC$(u)$의 2차 미분 방정식의 해를 상수 c, 예를 들어 다음 식과 같이 $(1-\gamma)/\gamma$로 설정하여 최적분류점을 구한다. 자세한 유도과정은 부록을 참조한다.

$$\frac{d^2}{du^2}\text{AUROC}(u) = \frac{f_d(x)}{f_n(x)} = \frac{\sigma_n}{\sigma_d}\exp\left(\frac{-(x-\mu_d)^2}{2\sigma_d^2} + \frac{(x-\mu_n)^2}{2\sigma_n^2}\right) = \frac{1-\gamma}{\gamma}.$$

1) $\sigma_d = \sigma_n = \sigma$인 경우

$$x_0 = \frac{\sigma^2}{\mu_d - \mu_n} ln\frac{1-\gamma}{\gamma} + \frac{\mu_n + \mu_d}{2}.$$

2-1) $\sigma_d \le \sigma_n$인 경우

$\frac{1-\gamma}{\gamma} \le \exp\left(\frac{\mu_n^2\sigma_d^2 - \mu_d^2\sigma_n^2}{2\sigma_d^2\sigma_n^2} - \frac{\mu_n\sigma_d^2 - \mu_d\sigma_n^2}{2\sigma_d^2\sigma_n^2(\sigma_d^2 - \sigma_n^2)}\right)\frac{\sigma_n}{\sigma_d}$의 조건에서

$$x_0 = \sqrt{\frac{(\mu_n\sigma_d^2 - \mu_d\sigma_n^2)^2}{(\sigma_d^2 - \sigma_n^2)^2} - \frac{\mu_n^2\sigma_d^2 - \mu_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2} + \frac{2\sigma_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2}\ln\left(\frac{\sigma_d}{\sigma_n}\frac{1-\gamma}{\gamma}\right)} + \left(\frac{\mu_n\sigma_d^2 - \mu_d\sigma_n^2}{\sigma_d^2 - \sigma_n^2}\right).$$

2-2) $\sigma_d > \sigma_n$인 경우

$\frac{1-\gamma}{\gamma} > \exp\left(\frac{\mu_n^2\sigma_d^2 - \mu_d^2\sigma_n^2}{2\sigma_d^2\sigma_n^2} - \frac{\mu_n\sigma_d^2 - \mu_d\sigma_n^2}{2\sigma_d^2\sigma_n^2(\sigma_d^2 - \sigma_n^2)}\right)\frac{\sigma_n}{\sigma_d}$의 조건에서

$$x_0 = -\sqrt{\frac{(\mu_n\sigma_d^2 - \mu_d\sigma_n^2)^2}{(\sigma_d^2 - \sigma_n^2)^2} - \frac{\mu_n^2\sigma_d^2 - \mu_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2} + \frac{2\sigma_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2}\ln\left(\frac{\sigma_d}{\sigma_n}\frac{1-\gamma}{\gamma}\right)} + \left(\frac{\mu_n\sigma_d^2 - \mu_d\sigma_n^2}{\sigma_d^2 - \sigma_n^2}\right).$$

AUCAP(u)의 경우, 앞에서 가정한 $f_d(x) = \phi(x|\mu_d, \sigma_d)$와 혼합분포 $f(x) = \gamma f_d(x) + (1-\gamma)f_n(x)$에 대하여 2차 미분 방정식의 해를 다음과 같이 $1/2\gamma$로 설정하여 최적분류점을 구한다. AUCAP(u)에서의 최적분류점은 AUROC(u)에서의 과정에서 $f_n(x)$를 $f(x)$ 그리고 $(1-\gamma)/\gamma$를 $1/2\gamma$로 변환한 후 전개한 결과와 유사하므로 생략한다.

$$\frac{d^2}{du^2}\text{AUCAP}(u) = \frac{f_d(x)}{f(x)} = \frac{\sigma}{\sigma_d}\exp\left(\frac{-(x-\mu_d)^2}{2\sigma_d^2} + \frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{2\gamma}.$$

1) $\sigma_d = \sigma$인 경우

$$x_0 = \frac{\sigma^2}{\mu_d - \mu_n}\ln\frac{1}{2\gamma} + \frac{\mu + \mu_d}{2}.$$

2-1) $\sigma_d \le \sigma$인 경우에

$\frac{1}{2\gamma} \le \exp\left(\frac{\mu^2\sigma_d^2 - \mu_d^2\sigma^2}{2\sigma_d^2\sigma^2} - \frac{\mu\sigma_d^2 - \mu_d\sigma^2}{2\sigma_d^2\sigma^2(\sigma_d^2 - \sigma^2)}\right)\frac{\sigma}{\sigma_d}$의 조건에서

$$x_0 = \sqrt{\frac{(\mu\sigma_d^2 - \mu_d\sigma^2)^2}{(\sigma_d^2 - \sigma^2)^2} - \frac{\mu^2\sigma_d^2 - \mu_d^2\sigma^2}{\sigma_d^2 - \sigma^2} + \frac{2\sigma_d^2\sigma^2}{\sigma_d^2 - \sigma^2}\ln\left(\frac{\sigma_d}{\sigma}\frac{1}{2\gamma}\right)} + \left(\frac{\mu\sigma_d^2 - \mu_d\sigma^2}{\sigma_d^2 - \sigma^2}\right).$$

Table 4.1. Optimal thresholds

γ	μ_n						
	2.0	2.5	3.0	3.5	4.0	4.5	5.0
0.3	0.576	0.911	1.218	1.508	1.788	2.062	2.331
0.4	0.797	1.088	1.365	1.634	1.899	2.16	2.419
0.5(TR)	1.000	1.250	1.500	1.750	2.000	2.250	2.500
0.6	1.203	1.412	1.635	1.866	2.101	2.34	2.581
0.7	1.424	1.589	1.782	1.992	2.212	2.438	2.669

TR = true rate.

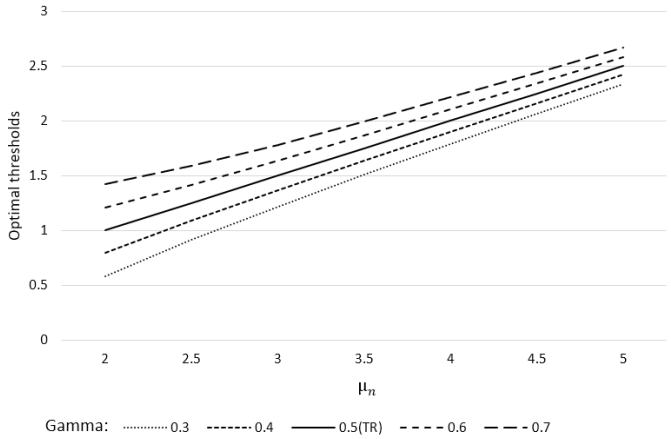


Figure 4.1. Optimal thresholds.

2-2) $\sigma_d > \sigma$인 경우에

$\frac{1}{2\gamma} > \exp\left(\frac{\mu^2\sigma_d^2 - \mu_d^2\sigma^2}{2\sigma_d^2\sigma^2} - \frac{\mu\sigma_d^2 - \mu_d\sigma^2}{2\sigma_d^2\sigma^2(\sigma_d^2 - \sigma^2)}\right)\frac{\sigma}{\sigma_d}$의 조건에서

$$x_0 = -\sqrt{\frac{(\mu\sigma_d^2 - \mu_d\sigma^2)^2}{(\sigma_d^2 - \sigma^2)^2} - \frac{\mu^2\sigma_d^2 - \mu_d^2\sigma^2}{\sigma_d^2 - \sigma^2} + \frac{2\sigma_d^2\sigma^2}{\sigma_d^2 - \sigma^2}\ln\left(\frac{\sigma_d}{\sigma}\frac{1}{2\gamma}\right)} + \left(\frac{\mu\sigma_d^2 - \mu_d\sigma^2}{\sigma_d^2 - \sigma^2}\right).$$

4. 모의실험

4.1. 모평균의 변화에 따른 최적분류점의 변화

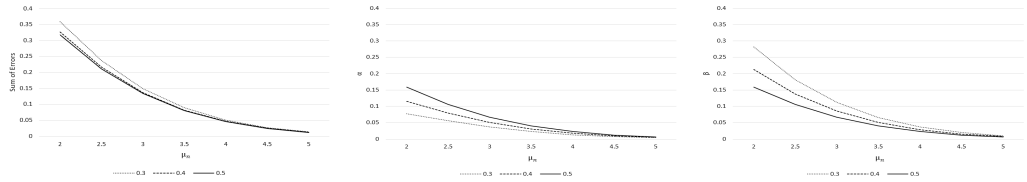
본 연구에서 부도의 분포를 표준정규분포 $f_d(x) = \phi(x|\mu_d = 0, \sigma_d^2 = 1)$로 가정하고 정상의 정규분포 함수 $f_n(x) = \phi(x|\mu_n, \sigma_n^2 = 1)$에서 평균 μ_n을 2.0부터 5.0까지 0.5 간격으로 변화시키고 표준편차를 1.0으로 가정한 경우 정확도 측도 TA의 모수 값에 따른 최적분류점을 구하여 Table 4.1에 그 위치를 정리하였고 μ_n의 증가함에 따라 최적분류점 변화를 살펴보기 위하여 Figure 4.1에 표현하였다.

Table 4.1과 Figure 4.1을 통하여 μ_n의 값이 작을 때 $\gamma > 0.5$인 경우 최적분류점의 값이 $\gamma = 0.5$인 경우보다 크며 $\gamma < 0.5$인 경우 작은 것을 확인하였다. 그리고 μ_n이 증가함에 따라 $\gamma = 0.5$인 TR에 의한 최적분류점의 값으로 수렴하는 경향이 있다. 또한 가정한 두 분포의 분산이 동일한 경우 TR에 의한 최적분류점은 $\mu_n/2$임을 확인하였다.

Table 4.2. Values of $\alpha, \beta, \alpha+\beta$

γ		μ_n						
		2.0	2.5	3.0	3.5	4.0	4.5	5.0
0.3 (= 0.7)	α	0.077	0.056	0.037	0.023	0.013	0.007	0.004
	β	0.282	0.181	0.112	0.066	0.037	0.020	0.010
	$\alpha+\beta$	0.359	0.237	0.149	0.089	0.050	0.027	0.014
0.4 (= 0.6)	α	0.115	0.079	0.051	0.031	0.018	0.01	0.005
	β	0.213	0.138	0.086	0.051	0.029	0.015	0.008
	$\alpha+\beta$	0.327	0.217	0.137	0.082	0.047	0.025	0.013
0.5 (TR)	α	0.159	0.106	0.067	0.04	0.023	0.012	0.006
	β	0.159	0.106	0.067	0.04	0.023	0.012	0.006
	$\alpha+\beta$	0.317	0.211	0.134	0.08	0.046	0.024	0.012

TR = true rate.



(a) Values of $\alpha+\beta$ (b) Values of α (c) Values of β

Figure 4.2. Values of errors.

Table 4.3. Optimal thresholds

γ	σ_n				
	1.0	1.5	2.0	2.5	3.0
0.3	0.576	0.442	0.502	0.654	0.812
0.5(TR)	1.000	1.087	1.238	1.377	1.492
0.7	1.424	1.605	1.759	1.879	1.973

TR = true rate.

4.2. 모평균의 변화에 따른 오류들의 변화

4절에서 가정한 분포의 변화에 따른 제1종 오류(α)와 제2종 오류(β) 그리고 오류합($\alpha+\beta$)을 구하여 구하여 Table 4.2에 정리하였고 μ_n의 증가함에 따라 각각의 오류 변화를 살펴보기 위하여 Figure 4.2(a)–(c)에 시각적으로 구현하였다.

Table 4.2과 Figure 4.2(a)–(c)를 바탕으로 분포의 μ_n이 증가할수록 모든 γ에 대하여 α, β, $\alpha+\beta$가 기하급수적으로 감소하며 $\gamma = 0.5$인 TR에 의한 값으로 수렴하였다. 또한 $\gamma = 0.5$인 경우 $\alpha = \beta$이며 $\gamma \neq 0.5$인 경우 $\alpha < \beta$임을 확인하였다. 따라서 α와 β 모두 감소추세이기 때문에 오류합의 감소추세가 기하급수적임을 확인할 수 있었다. 여기서 어떠한 γ에서의 오류는 $1-\gamma$에서의 값과 같다.

4.3. 모분산의 변화에 따른 최적분류점의 변화

4절에서 가정한 정상의 정규분포의 모평균 μ_n을 2.0으로 가정하고, 표준편차 σ_n을 1.0부터 3.0까지 0.5 간격으로 변화시키는 경우 정확도 측도 TA의 모수 값에 따른 최적분류점을 구하여 Table 4.3에 정리하였고 σ_n이 증가함에 따라 최적분류점 변화를 살펴보기 위하여 Figure 4.3에 표현하였다.

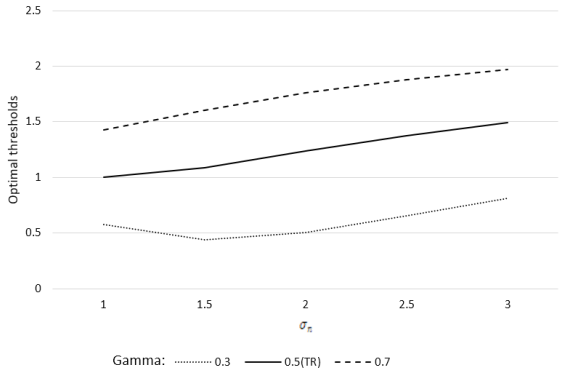
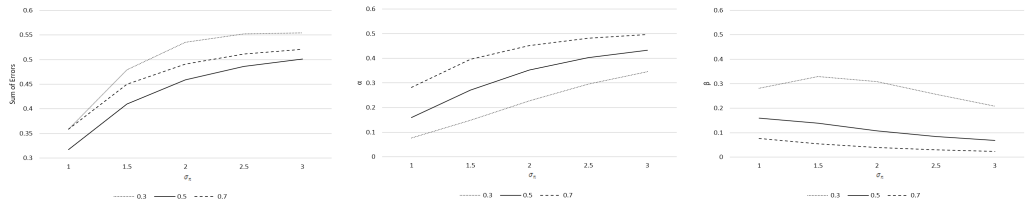


Figure 4.3. Optimal thresholds.

Table 4.4. Values of $\alpha, \beta, \alpha+\beta$

γ		σ_n				
		1.0	1.5	2.0	2.5	3.0
0.3	α	0.077	0.149	0.227	0.295	0.346
	β	0.282	0.329	0.308	0.257	0.208
	$\alpha+\beta$	0.359	0.479	0.535	0.552	0.554
0.5 (TR)	α	0.159	0.271	0.352	0.402	0.433
	β	0.159	0.139	0.108	0.084	0.068
	$\alpha+\beta$	0.317	0.41	0.459	0.486	0.501
0.7	α	0.282	0.396	0.452	0.481	0.496
	β	0.077	0.054	0.039	0.03	0.024
	$\alpha+\beta$	0.359	0.45	0.491	0.511	0.521

TR = true rate.



(a) Values of $\alpha+\beta$ (b) Values of α (c) Values of β

Figure 4.4. Values of errors.

Table 4.3과 Figure 4.3을 통하여 정상 분포의 σ_n이 증가할수록 모든 γ에 대하여 최적분류점이 증가하며, $\gamma > 0.5$인 경우 μ_n으로 수렴하는 것을 확인하였다. $\gamma < 0.5$인 경우에는 최적분류점이 소폭 감소한 후 다시 증가하는데 이는 $\gamma < 0.5$의 경우에 3절에서의 제곱근안의 $-\ln\sigma_n$이 초반에 기하급수적으로 증가한 이후 완만하게 증가하기 때문이다. 또한 γ 값이 클 수록 최적분류점의 값이 더 크며 μ_n의 증가에 따른 변화와 달리 $\gamma = 0.5$인 TR에 의한 값에 수렴하지 않음을 확인하였다.

4.4. 모분산의 변화에 따른 오류들의 변화

본 연구의 4절에서 가정한 분포의 변화에 따른 제1종 오류(α)와 제2종 오류(β) 그리고 오류합($\alpha+\beta$)을

구하여 Table 4.4에 정리하고 Figure 4.4(a)–(c)에 표현하면서 변화를 살펴보았다.

Table 4.4와 Figure 4.4(a)–(c)를 바탕으로 σ_n이 증가할수록 모든 γ에 대하여 α와 $\alpha+\beta$값이 증가함을 탐색하였다. 반면 β값은 감소하며 $\gamma > 0.5$인 경우 $\gamma = 0.5$인 TR에 의한 값으로 수렴한다. $\gamma < 0.5$인 경우 해당 값이 일정부분 증가한 후 다시 감소하는데 이런 현상은 4.3절에서 설명하였듯이 $\gamma < 0.5$인 경우 σ_n이 증가함에 따라 최적분류점이 감소한 후 다시 증가하기 때문이다. 또한 이러한 α, β, $\alpha+\beta$의 추세는 μ_n의 증가에 따른 변화 만큼의 기하급수적인 모습은 아님을 확인하였다.

5. 결론

신용평가, 제약, 의학, 고객분류 등 다양한 분야에서 두 종류 함수의 혼합분포가 주어진 경우 분류모형을 구현하는 ROC와 CAP 곡선을 바탕으로 최적분류점을 추정하는 많은 연구 문헌 중에서 Hong과 Choi (2009), Yoo와 Hong (2011), 그리고 Hong 등 (2010)은 ROC와 CAP 함수들과 다양한 정확도 측도에 해당하는 일차원 접선의 접점을 활용하여 최적분류점을 추정하는 방법을 제안하였다. 본 연구에서는 ROC와 CAP 곡선의 AUC의 특정한 부분의 면적을 나타내는 부분 AUC를 제안하고, 대안적인 부분 AUC와 ROC와 CAP 곡선와 관계를 일차 미분방정식으로 유도하고, 이차 미분방정식을 이용하여 ROC와 CAP 곡선에서의 최적분류점들과의 관계를 유도하여 다양한 최적분류점을 추정하기 위한 여러 조건으로 다시 정리하였다. 혼합분포를 구성하는 두 종류의 분포함수를 다양한 정규분포로 가정하여 최적분류점을 유도하고 이에 대응하는 제1종과 제2종 오류를 구하였다.

부도와 정상 분포를 가정한 후 평균과 표준편차의 변화에 따른 최적분류점과 이에 대응하는 오류합의 변화를 다양한 정확도 측도들 중에서 대표적인 TR과 TA 변화에 따라 살펴보았다. 설정한 두 분포의 평균 차이가 증가함에 따라 최적분류점의 값 역시 증가하며, 전체 부도율이 0.5인 즉, TR인 경우로 수렴함을 확인하였다. 또한 이 경우에 오류합이 기하급수적으로 감소하며 마찬가지로 전체 부도율이 0.5인 TR인 경우로 수렴함을 보였다. 혼합분포들의 표준편차 차이가 증가함에 따라 최적분류점이 증가하며, 전체 부도율이 0.5이하인 경우에는 최적분류점이 일정부분 감소한 후 다시 증가함을 확인하였다. 그러나 이런 모든 경우에도 오류합이 증가함을 살펴볼 수 있었다.

분류하는 혼합모형에서 정확도 측도에 따른 최적분류점을 추정하는 문제에서는 본 연구에서 제안한 부분곡선아래면적의 일차와 이차 미분 방정식을 이용하여 최적분류점의 추정을 효율적으로 해결할 수 있음을 확인하였다. 따라서 서로 다른 정확도 측도들을 비교하고 이에 대한 최적분류점의 변화하는 모습을 탐색해야 하는 경우에 ROC와 CAP 곡선의 함수와 접선함수의 접점을 통해 추정하는 기존 방법들보다 본 연구에서 제안한 방법을 활용하는 것이 많은 장점이 있음을 확인하였다.

본 연구는 실제로 부도와 정상 기업을 판별해야 하는 신용평가자료 또는 양성과 음성 반응을 분류해야 하는 임상실험자료 등에 대하여 두 집단의 최적분류점을 구하는 경우에 적용할 수 있으며, 본 연구에서 제안한 부분 AUC를 바탕으로 다양한 정확도 측도들에 대한 최적분류점을 구하고 이에 대한 특성을 쉽게 파악할 수 있는 장점이 있다.

부록

1) $\sigma_d = \sigma_n = \sigma$인 경우

$$\begin{aligned}\frac{f_d(x)}{f_n(x)} &= \exp\left(\frac{2(\mu_d-\mu_n)x}{2\sigma^2} + \frac{(\mu_n^2-\mu_d^2)}{2\sigma^2}\right)\\ &= \exp\left(\frac{(\mu_d-\mu_n)x}{\sigma^2} + \frac{(\mu_n^2-\mu_d^2)}{2\sigma^2}\right) = \frac{1-\gamma}{\gamma}.\end{aligned}$$

따라서

$$x_0 = \frac{\sigma^2}{\mu_d - \mu_n} \ln \frac{1-\gamma}{\gamma} + \frac{\mu_n + \mu_d}{2}.$$

2) $\sigma_d \neq \sigma_n$인 경우

$$\frac{f_d(x)}{f_n(x)} = \frac{\sigma_n}{\sigma_d} \exp\left(\frac{-(x-\mu_d)^2}{2\sigma_d^2} + \frac{(x-\mu_n)^2}{2\sigma_n^2} \right) = \frac{1-\gamma}{\gamma}.$$

$$\left(\sigma_d^2 - \sigma_n^2\right)\left(x - \left(\frac{\mu_n\sigma_d^2 - \mu_d\sigma_n^2}{\sigma_d^2 - \sigma_n^2}\right)\right)^2 - \frac{\left(\mu_n\sigma_d^2 - \mu_d\sigma_n^2\right)^2}{\sigma_d^2 - \sigma_n^2} + \left(\mu_n^2\sigma_d^2 - \mu_d^2\sigma_n^2\right) = 2\sigma_d^2\sigma_n^2 \ln\left(\frac{\sigma_d}{\sigma_n}\frac{1-\gamma}{\gamma}\right).$$

$\mu_d \leq \mu_n$의 가정에서 최적분류점은 다음과 같다.

2-1) $\sigma_d \leq \sigma_n$인 경우

$$x_0 = \sqrt{\frac{(\mu_n\sigma_d^2 - \mu_d\sigma_n^2)^2}{(\sigma_d^2 - \sigma_n^2)^2} - \frac{\mu_n^2\sigma_d^2 - \mu_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2} + \frac{2\sigma_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2} \ln\left(\frac{\sigma_d}{\sigma_n}\frac{1-\gamma}{\gamma}\right)} + \left(\frac{\mu_n\sigma_d^2 - \mu_d\sigma_n^2}{\sigma_d^2 - \sigma_n^2}\right).$$

2-2) $\sigma_d > \sigma_n$인 경우

$$x_0 = -\sqrt{\frac{(\mu_n\sigma_d^2 - \mu_d\sigma_n^2)^2}{(\sigma_d^2 - \sigma_n^2)^2} - \frac{\mu_n^2\sigma_d^2 - \mu_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2} + \frac{2\sigma_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2} \ln\left(\frac{\sigma_d}{\sigma_n}\frac{1-\gamma}{\gamma}\right)} + \left(\frac{\mu_n\sigma_d^2 - \mu_d\sigma_n^2}{\sigma_d^2 - \sigma_n^2}\right).$$

위의 식에서 제곱근은 양수 조건이 성립되어야 하므로

- $\sigma_d \leq \sigma_n$인 경우에는

$$\frac{1-\gamma}{\gamma} \leq \exp\left(\frac{\mu_n^2\sigma_d^2 - \mu_d^2\sigma_n^2}{2\sigma_d^2\sigma_n^2} - \frac{\mu_n\sigma_d^2 - \mu_d\sigma_n^2}{2\sigma_d^2\sigma_n^2(\sigma_d^2 - \sigma_n^2)} \right) \frac{\sigma_n}{\sigma_d}$$

그리고

- $\sigma_d > \sigma_n$인 경우에는

$$\frac{1-\gamma}{\gamma} > \exp\left(\frac{\mu_n^2\sigma_d^2 - \mu_d^2\sigma_n^2}{2\sigma_d^2\sigma_n^2} - \frac{\mu_n\sigma_d^2 - \mu_d\sigma_n^2}{2\sigma_d^2\sigma_n^2(\sigma_d^2 - \sigma_n^2)} \right) \frac{\sigma_n}{\sigma_d}$$

의 조건이 성립하여야 한다.

References

Berry, M. J. A. and Linoff, G. (1999). *Data Mining Techniques: For Marketing, Sales, and Customer Support* (3rd ed), John Wiley & Sons, New York.

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.

Brasil, P. (2010). DiagnosisMed: Diagnostic test accuracy evaluation for medical professionals, Package Diagnosis Med in R, from: http://www.CRAN.R-project.org/src/contrib/Archive/DiagnosisMed

Centor, R. M. (1991). Signal detectability: the use of ROC analysis, *Med Decision Making*, **11**, 102–106.

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

Dodd, L. E. and Pepe, M. S. (2003). Partial AUC Estimation and Regression, *Biometrics*, **59**, 614–623.

Engelmann, B., Hayden, E. and Tasche, D. (2003). Measuring the discriminative power of rating systems, *Discussion paper, Series 2: Banking and Financial Supervision*, No. 01/2003.

Fawcett, T. (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, *HP Laboratories, Palo Alto*, HPL-2003-4.

Fawcett, T. (2006). An introduction to ROC analysis, *Pattern Recognition Letters*, **27**, 861–874.

Fawcett, T. and Provost, F. (1997). Analysis and visualization of classier performance comparison under imprecise class and cost distributions, *Knowledge Discovery and Data Mining*, **97**, 43–48.

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

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 threshold from mixture distributions, *The Korean Journal of Applied Statistics*, **23**, 13–28.

Irwin, J. R. and Irwin, C. T. (2012). Appraising Credit Ratings: Does the CAP Fit Better than the ROC?, *International Monetary Fund Working paper*, WP. 12/122.

Jiang, Y., Metz, C., and Nishikawa, R. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests, *Radiology*, **201**, 745–750.

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

Krzanowski, W. J. and Hand, D, J. (2009). *ROC Curves for Continuous Data*, CRC Press, New York.

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

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. (2000). Receiver operating characteristic methodology, *Journal of the American Statistical Association*, **95**, 308–311.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Test for Classification and Prediction* (17th ed), Oxford 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.

Sobehart, J. R., Keenan, S. C., and Stein, R. M. (2000). Benchmarking quantitative default risk models: a validation methodology, *Moody's Investors Service*.

Swets, J. A., Dawes, R. M., and Monahan, J. (2000). Better Decisions through Science, *Scientific American*, 82–87.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems, *American Association for the Advancement of Science*, **240**, 1285–1293.

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.

Vuk, M. and Curk, T. (2006). ROC curve, lift chart and calibration plot, *Metodološki Zvezki*, **3**, 89–108.

Yoo, H. S. and Hong, C. S. (2011). Optimal criterion of classification accuracy measures for normal mixture, *The Korean Journal of Applied Statistics*, **18**, 343–355.

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

Zou, K. H. (2002). Receiver operating characteristic literature research, from: http://www.spl.harvard.edu/pages/ppl/zou/roc.html

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, **39**, 561–577.

부분 AUC와 최적분류점들

홍종선[a,1] · 조현수[a]

[a]성균관대학교 통계학과

(2019년 2월 11일 접수, 2019년 2월 25일 수정, 2019년 2월 27일 채택)

요 약

ROC와 CAP 곡선을 이용하여 다양한 정확도 측도를 바탕으로 최적분류점을 추정하는 많은 연구가 있다. 본 연구에서는 ROC와 CAP 곡선의 특정한 부분 면적을 나타내는 대안적인 통계량을 제안한다. 새롭게 정의된 부분 면적을 나타내는 통계량의 미분방정식을 이용하여 ROC와 CAP 함수와의 관계를 살펴보고, 다음으로는 ROC와 CAP 곡선에 대한 다양한 정확도 측도들의 조건에서의 최적분류점과의 관계를 유도한다. 혼합분포를 구성하는 두 종류의 분포함수를 다양한 정규분포로 가정하여 최적분류점을 설정하고, 다양한 정확도 측도들의 조건에서의 최적분류점에 대응하는 제1종과 제2종 오류의 크기를 탐색하고 토론한다.

주요용어: 정확도, 부도, 분류, 오류, 최적분류점, 혼동행렬

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: cshong@skku.edu