응용통계연구 (2010)
23(1), 13–28

혼합분포에서 최적분류점

홍종선[1] · 주재선[2] · 최진수[3]

[1]성균관대학교 경제학부, [2]한국여성정책연구원 통계패널센터, [3]성균관대학교 응용통계연구소

(2009년 9월 접수, 2009년 11월 채택)

요 약

혼합분포를 가정한 신용평가연구에서 부도차주를 정상으로 예측하거나 정상차주를 부도로 예측하는 오류를 최소화하는 분류점을 추정하는 방법을 토론한다. 확률변수 스코어와 정상과 부도상태의 모수공간으로 정의된 확률밀도함수들에 대하여 강력검정과 일반화가능도비검정을 이용하여 최적분류점의 추정방법을 제안하고, ROC와 CAP 곡선에서 분류정확도를 측정하는 정확도(accuarcy)와 진실율(true rate)을 이용하여 이 측도를 최대로 하는 최적분류점을 확률밀도함수의 관계식으로 추정하는 방법을 제안한다. 다양한 정규분포에서 가설검정, 정확도 그리고 진실율을 이용하는 세가지 방법의 최적분류점을 구하고 각 최적분류점에 대응하는 제 I 종과 제 II 종 오류합의 크기를 비교하여 효율성을 토론한다.

주요용어: 가능도비검정, 강력검정, 부도, 분류점, 스코어, 정확도, 진실율, 오류, 판별력, CAP, ROC.

1. 서론

두 종류의 분포함수가 혼합된 분포로부터 추출한 확률표본의 판별력을 극대화하는 분류점(threshold, cut-off)을 추정하는 연구는 공학에서 신호탐지이론으로 시작하여 의사결정론, 마케팅 판매에 관한 데이터 마이닝 그리고 의학진단체계에서 폭 넓게 사용되었다. 최근에는 신용평가연구에서도 많이 이용되는데 본 연구에서는 대표적인 신용평가연구에 관하여 설명하고자 한다.

신용평가모형에서 차주(borrower)의 신용가치를 기준으로 대출상환능력이 없는 부도상태(default)와 능력이 있는 정상상태(non-default)로 판별하는 문제를 고려하자. 확률변수 S는 스코어 변수로 대출기관에서 차주의 신용가치를 예측하기 위해 차주에게 부여한 연속형 실수값이다. 차주가 대출상환능력이 없는 부도상태와 능력이 있는 정상상태로 가정하여 모수공간을 $\Theta = \{\theta_d, \theta_n\}$으로 정의한다. 차주의 모집단은 미래시점에 대출상환능력이 없는 부도와 대출상환능력이 있는 정상으로 구분된 두 개의 부모집단으로 구성되어 있다고 가정한다. 차주의 상태가 θ_d일 때 부도차주의 모집단에 속하고, 차주의 상태가 θ_n값일 때 정상차주의 모집단에 속한다. 따라서 $f_d(s)$와 $f_n(s)$을 각각 차주의 부도와 정상상태에서 스코어의 조건부 확률밀도함수 $f_d(s) = f(s\,|\,\theta_d)$와 $f_n(s) = f(s\,|\,\theta_n)$로 정의하며, 스코어 확률밀도함수 $f(s)$는 다음과 같이 표현한다.

$$f(s) = \gamma f_d(s) + (1-\gamma) f_n(s), \quad s \in (-\infty,\ \infty), \tag{1.1}$$

여기서 γ는 부도율총합(total probability of default)이다. 두 종류의 차주 상태가 주어진 스코어 변수 S의 조건부 누적분포함수를 각각 $F_d(\cdot)$와 $F_n(\cdot)$이라 하면, 조건부 누적분포함수 $F_d(s) = P(S \leq$

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

$s\,|\,\theta_d)$은 스코어 s에서 부도차주를 부도차주로 정확히 예측한 비율로써, 'hit rate', 'recall', 'sensitivity' 또는 'true positive rate'라 하며, $F_n(s) = P(S \le s\,|\,\theta_n)$는 스코어 s에서 정상차주를 부도차주로 잘못 예측한 비율로써, 'false alarm rate', '1 − specificity' 또는 'false positive rate'라고 한다. 한편 비조건부 누적함수 $F(s) = P(S \le s)$는 전체 모집단에서 부도로 예측되는 비율이며 'alarm rate'라고 한다.

차주의 신용가치에 관한 정보인 스코어 변수 S를 바탕으로 차주의 미래 상황을 예측하기 위하여 부도상태인 P개 차주의 확률표본은 $f_d(s)$로부터 추출하고, N개의 정상차주의 확률표본은 $f_n(s)$로부터 추출되어 총표본 크기는 $P+N$으로 가정한다. 확률표본을 특정한 분류점(threshold, cut-off score)을 기준으로 분류한 결과는 실제 상황과 예측 결과를 통해 2×2 분할표 혹은 혼동행렬(confusion matrix)로 나타난다 (Fawcett, 2003). 혼동행렬은 1998년 Kohavi와 Provost가 고안한 것으로 실제부도를 부도로 예측한 것을 TP로, 실제정상을 정상으로 예측한 것으로 정분류한 경우를 TN로 표기한다. 반면에 FP는 실제 정상임에도 부도로 예측한 것이고 FN는 실제 부도를 정상으로 예측한 것으로 오분류한 경우이다. 혼동행렬에서 실제 부도와 정상의 표본수인 P와 N은 각각 $P = \mathrm{TP} + \mathrm{FN}$, $N = \mathrm{TN} + FP$이다. 이에 예측 부도와 예측 정상의 개수는 각각 TP + FP와 TN + FN으로 할당된다. 특정한 분류점 x_c에서 $F_d(s)$, $F_n(s)$ 그리고 $F(s)$의 추정은 각각 $\hat{F}_d(x_c) = \mathrm{TP}/P$, $\hat{F}_n(x_c) = \mathrm{FP}/N$ 그리고 $\hat{F}(x_c) = (\mathrm{TP}+\mathrm{FP})/(P+N)$. 또한 $\hat{F}(x_c) = \hat{\gamma}\hat{F}_d(x_c) + (1-\hat{\gamma})\hat{F}_n(x_c)$이고 부도율총합의 추정량은 $\hat{\gamma} = P/(P+N)$이다.

오분류한 경우인 FN은 제 I 종 오류 그리고 FP는 제II종 오류 빈도수로 간주할 수 있고 (Stein 2005), 특정한 분류점 x_c에서 제 I 종 오류(α)와 제II종 오류(β)는 다음과 같다.

$$\alpha = P(S > x_c \mid \theta_d) = 1 - F_d(x_c), \qquad \beta = P(S \le x_c \mid \theta_n) = F_n(x_c).$$

신용평가모형에서 제 I 종 오류는 대출 혹은 금융거래를 허용하였을 때 거래자가 이를 상환하지 못함으로써 발생하는 손실이며, 제II종 오류는 정상 대출 혹은 금융거래임에도 이를 거부함으로써 발생하는 잠재적 손실이다. 이에 신용평가모형은 제 I 종과 제II종 오류합이 최소가 되는 지점을 찾는 것이 중요하며, 이를 최소로 하는 분류점 또는 절단점은 손실을 최소화하는 최적분류점(optimal threshold)이 된다.

분류의 정확도를 측정하는 통계량으로 정확도(accuracy)와 진실율(true rate)을 이용하여 ROC(Receiver Operation Characteristic)와 CAP(Cumulative Accuracy Profile) 곡선에서 최적분류점을 발견하는 방법이 존재한다. ROC와 CAP 곡선에 관한 연구는 Hanley와 McNeil (1982), Swets (1988), Berry과 Linoff (1999), Sobehart 등 (2000), Provost와 Fawcett (1997, 2001), Sobehart와 Keenan (2001), Zou (2002), Engelmann 등 (2003), Fawcett (2003), Drummond와 Holte (2006) 등에서 찾아볼 수 있다. 정확도(AC)와 진실율(TR)의 추정량은 다음과 같이 정의한다:

$$\widehat{\mathrm{AC}} = \frac{\mathrm{TP}+\mathrm{TN}}{P+N}, \qquad \widehat{\mathrm{TR}} = \frac{1}{2}\left(\frac{\mathrm{TP}}{P} + \frac{\mathrm{TN}}{N}\right).$$

정확도는 가중평균(weighted mean)으로, 진실율은 산술평균(arithmetic mean)으로 다음과 같다.

$$\begin{aligned}\widehat{\mathrm{AC}} &= \left(\frac{P}{P+N}\right)\frac{\mathrm{TP}}{P} + \left(\frac{N}{P+N}\right)\left(1 - \frac{\mathrm{FP}}{N}\right) \\ &= \hat{\gamma}\hat{F}_d(s) + (1-\hat{\gamma})\left(1 - \hat{F}_n(s)\right), \qquad (1.2)\end{aligned}$$

$$\widehat{\mathrm{TR}} = \frac{1}{2}\left[\hat{F}_d(s) + \left\{1 - \hat{F}_n(s)\right\}\right]. \qquad (1.3)$$

식 (1.2)의 정확도에서 $\hat{\gamma} = 1/2$일 때 식 (1.3)의 진실율이다.

Vuk과 Curk (2006)은 정확도를 이용하고, 홍종선과 최진수 (2009)는 진실율을 이용하여 동일한 성과를 나타내는 선형식(iso-performance line)을 유도하여 이 식과 ROC와 CAP 곡선과의 접점이 정확도 또는 진실율을 최대로 하는 최적분류점을 발견했다. 홍종선과 최진수 (2009)는 총표본 크기 $P+N$ 중에서 부도차주수인 P와 정상 차주수인 N의 비율이 큰 차이가 나는 경우(일반적으로 P는 N보다 훨씬 작으며 따라서 부도율총합 γ는 0.5 미만이다)와 제 I 종 오류 비용이 제 II 종 오류 비용보다 큰 액수의 비용함수를 고려하는 경우에는 진실율을 최대로 하는 최적분류점이 정확도를 최대로 하는 최적분류점보다 효율적인 것을 보였다. 그리고 조건부 누적분포함수 $F_d(\cdot)$와 $F_n(\cdot)$를 비교하는 검정통계량으로 Kolmogorov - Smirnov 통계량과 진실율이 선형적이라는 사실을 유도하여 진실율이 정확도보다 좋은 분류정확도를 측정하는 통계량이라고 주장했다.

최적분류점을 추정하기 위하여 Vuk과 Curk (2006)과 홍종선과 최진수(2009)는 식 (1.1)에서 가정된 확률밀도함수의 누적분포함수로 표현되는 ROC와 CAP 곡선을 바탕으로 연구하였지만, 본 연구에서는 ROC와 CAP 곡선을 구성하는 확률밀도함수들의 관계식으로부터 정확도와 진실율을 최대로 하는 최적분류점을 추정하는 이론을 제안한다. 그리고 모수공간에서의 원소에 대한 가설을 확률밀도함수를 비교검정하는 가설로 변환하여 통계적 가설검정모형을 만든 후, 강력검정(most powerful test)과 일반화가능도비검정(generalized likelihood ratio test)을 이용하여 최적분류점을 추정하는 방법을 제안한다. 또한 제안한 세가지 방법인 정확도와 진실율 그리고 가설검정방법을 이용하여 추정한 최적분류점으로부터 각각 제 I 종과 제 II 종의 오류합을 구하고 이를 비교하여 오류합이 가장 작은 방법에 대하여 토론하고자 한다.

본 연구의 구성은 다음과 같다. 2절에서는 부도와 정상상태의 조건부 밀도함수 $f_d(\cdot)$와 $f_n(\cdot)$을 비교 검정하는 강력검정과 조건부밀도함수에 대응하는 조건부 누적분포함수로 표현되는 ROC 곡선에서 정확도와 진실율을 이용하여 최적분류점을 발견하는 세가지 방법을 소개한다. 3절에서는 $f_d(\cdot)$와 $f(\cdot)$을 비교검정하는 일반화가능도비검정과 이 밀도함수에 대응하는 누적분포함수 $F_d(\cdot)$와 $F(\cdot)$로 표현되는 CAP 곡선에서 정확도와 진실율을 이용하여 최적분류점을 발견하는 세가지 방법을 소개하고 비교한다. 4절에서는 다양한 평균과 분산을 갖는 정규분포의 경우에 대하여 최적분류점을 구하고 각각의 방법으로 구한 제 I 종과 제 II 종의 오류합을 비교하면서 각 방법들의 효율성을 토론한다. 마지막으로 5절은 이를 종합하여 결론을 유도한다.

2. ROC 함수와 최적분류점

2.1. 강력검정을 이용하는 방법

Tasche (2006)의 연구에서도 본 연구와 유사하게 단순 가설을 설정하였고 네이만-피어슨 정리(Neymann - Pearson lemma)를 이용한 가능도비검정(likelihood ratio test) 방법 그리고 제 I 종과 제 II 종 오류에 대하여 언급했지만, 여기에서는 다음과 같은 방법으로 최적분류점을 추정하고자 한다.

확률변수 X는 분류점을 나타내는 변수이고, 확률변수 X값의 범위는 스코어 변수 S와 같이 실수이며, 확률밀도함수는 식 (1.1)과 동일하다고 가정한다. 단, 스코어 변수의 확률표본의 크기는 $P+N$이나 분류점을 나타내는 확률변수의 표본의 크기는 1이라고 가정한다. 모수공간 $\Theta = \{\theta_d, \theta_n\}$에 속한 모수에 대하여 다음과 같은 단순 가설을 고려하자.

$$H_0 : \theta = \theta_d \quad vs. \quad H_1 : \theta = \theta_n.$$

이 가설은 부도와 정상상태에서의 조건부 확률밀도함수 $f_d(x) = f(x \mid \theta_d)$와 $f_n(x) = f(x \mid \theta_n)$에 관한 가설로 변환하여 고려할 수 있다. 이 가설을 이용하여 최적분류점을 추정하면 다음과 같다.

정리 2.1 강력검정을 이용한 최적분류점

$$H_0 : f_d(x) \quad vs. \quad H_1 : f_n(x).$$

위 가설에 대하여 유의수준 α에서의 강력검정을 이용하여 구한 최적분류점 x_c은 다음을 만족한다.

$$x_c = F_d^{-1}(1-\alpha). \tag{2.1}$$

증명: 최적의 분류점은 제 I 종 오류와 제 II 종 오류합이 최소가 되는 분류점이어야 한다. 정리 2.1의 가설에 대하여 제 I 종 오류크기 α가 주어진 네이만-피어슨 정리를 이용하는 강력검정을 실시하면, 상수 c_α와 대응하는 임계값 x_c에 대하여 제 I 종 오류 α는 다음과 같다.

$$\alpha = P\left(\frac{f_d(x)}{f_n(x)} < c_\alpha \middle| \theta_d\right) = P\left(X > x_c | \theta_d\right),$$

그러므로 임계값 x_c는 주어진 제 I 종 오류 α에 대응하고 제 II 종 오류 β가 가장 적은 분류점이 된다. 제 I 종 오류를 고정시켰기 때문에 최적분류점은 $x_c = F_d^{-1}(1-\alpha)$이다. □

제 II 종 오류는 $\beta = P[X < x_c|\theta_n] = F_n(x_c)$이 된다. 따라서 강력검정을 이용하여 구한 최적분류점에 대응하는 오류합은 다음과 같다.

$$\alpha + F_n(x_c). \tag{2.2}$$

2.2. 정확도를 이용하는 방법

정확도를 정의한 식 (1.2)은 2.1절에서 다루는 확률밀도함수의 누적분포함수 $F_d(\cdot), F_n(\cdot)$의 함수로 나타나고, 이 누적분포함수로 ROC 곡선이 표현되고 있다. 정확도로부터 유도한 최적분류점은 다음과 같다.

정리 2.2 정확도를 이용한 최적분류점

정확도를 최대로 하는 분류점 x_c은 다음을 만족한다.

$$\gamma f_d(x_c) = (1-\gamma) f_n(x_c). \tag{2.3}$$

증명: 정확도는 스코어 s의 함수로 $\mathrm{AC}(s) = \gamma F_d(s) + (1-\gamma)(1-F_n(s))$이며, $u = F_n(s)$의 함수로 표현하면 다음과 같다.

$$\mathrm{AC}(u) = \gamma F_d\left(F_n^{-1}(u)\right) + (1-\gamma)(1-u),$$

여기서 $s = F_n^{-1}(u)$이다. 그리고 정확도를 ROC 함수인 $\mathrm{ROC}(u) = F_d(F_n^{-1}(u))$를 이용하여 나타내면 다음과 같다.

$$\mathrm{AC}(u) = \gamma \mathrm{ROC}(u) + (1-\gamma)(1-u).$$

$\mathrm{AC}(u)$가 최대가 되는 분류점을 찾기 위해 u에 대하여 미분하고

$$\frac{\partial}{\partial u}(\mathrm{AC}(u)) = \gamma\left[\frac{f_d\left(F_n^{-1}(u)\right)}{f_n\left(F_n^{-1}(u)\right)}\right] - (1-\gamma)$$

이를 0으로 놓으면, 최적의 분류점 $x_c = F_n^{-1}(u_c)$는 다음 식을 만족한다.

$$\frac{f_d(F_n^{-1}(u_c))}{f_n(F_n^{-1}(u_c))} = \frac{f_d(x_c)}{f_n(x_c)} = \frac{1-\gamma}{\gamma}.$$

참조: ROC 함수의 미분은 Pepe (2003), Tasche (2006, 2009) 등의 문헌에서 참조할 수 있다. □

정확도를 최대로 하는 분류점 x_c에 대하여 $F_d(x_c) = 1-\alpha$이고 $F_n(x_c) = \beta$이므로 최적분류점에 대응하는 최대의 정확도 $\mathrm{AC}(x_c)$를 제 I 종 오류 α와 제 II 종 오류 β와의 관계식으로 나타내면

$$\mathrm{AC}(x_c) = \gamma(1-\alpha) + (1-\gamma)(1-\beta) = 1 - [\gamma\alpha + (1-\gamma)\beta]$$

된다. 따라서 $\gamma f_d(x_c) = (1-\gamma) f_n(x_c)$를 만족하는 분류점 x_c에서 정확도가 최대값을 가지며 이 분류점에 대응하는 제 I 종 오류와 제 II 종 오류는 다음과 같다.

$$\gamma\alpha + (1-\gamma)\beta = 1 - \mathrm{AC}(x_c).$$

2.3. 진실율을 이용하는 방법

진실율을 정의한 식 (1.3)도 2.2절에서와 같이 누적분포함수 $F_d(\cdot), F_n(\cdot)$의 함수로 나타나며 진실율로부터 유도한 최적분류점은 다음과 같다.

정리 2.3 진실율을 이용한 최적분류점

진실율을 최대로 하는 분류점 x_c은 다음을 만족한다.

$$f_d(x_c) = f_n(x_c). \tag{2.4}$$

증명: 스코어 s의 함수로 표현한 진실율은 $2\mathrm{TR}(s) = F_d(s) + (1 - F_n(s))$이며, 이를 u의 함수로 변환하고 ROC 함수로 표현하면 다음과 같다.

$$2\mathrm{TR}(u) = F_d\left(F_n^{-1}(u)\right) + 1 - u = \mathrm{ROC}(u) + 1 - u.$$

$\mathrm{TR}(u)$가 최대가 되는 분류점을 찾기 위해 u에 대하여 미분하고

$$\frac{\partial}{\partial u}(2\mathrm{TR}(u)) = \left[\frac{f_d\left(F_n^{-1}(u)\right)}{f_n\left(F_n^{-1}(u)\right)}\right] - 1$$

이를 0으로 놓으면, 최적의 분류점 x_c는 다음 식을 만족한다.

$$\frac{f_d\left(F_n^{-1}(u_c)\right)}{f_n\left(F_n^{-1}(u_c)\right)} = \frac{f_d(x_c)}{f_n(x_c)} = 1.$$

□

최적분류점에서의 진실율을 제 I 종 오류 α와 제 II 종 오류 β와의 관계식으로 나타내면 $2\mathrm{TR}(x_c) = 2 - (\alpha+\beta)$이므로 $f_d(x_c) = f_n(x_c)$을 만족하는 분류점 x_c에서 진실율이 최대값을 가지며 이때 제 I 종과 제 II 종 오류합 $\alpha+\beta$가 최소가 되며 그 최소값은 다음과 같다.

$$\alpha + \beta = 2 - 2\mathrm{TR}(x_c).$$

3. CAP 함수와 최적분류점

3.1. 반화가능도비검정을 이용한 방법

모수공간은 Θ에 속한 모수에 대하여 다음과 같은 복합 가설을 고려하자.

$$H_0 : \theta = \theta_d \quad vs. \quad H_1 : \theta \neq \theta_d.$$

두 개의 원소만으로 구성된 모수공간 $\Theta = \{\theta_d, \theta_n\}$에 대하여 강력검정방법 이외의 일반화가능도비검정 방법을 고려하면, $\Theta_0 = \{\theta_d\}$에 대하여 위의 가설을 $H_0 : \theta \in \Theta_0$ *vs.* $H_1 : \theta \in \Theta$으로 변환하여 고려할 수 있다. 이때 일반화가능도비검정 통계량 Λ은 다음과 같은 가능도비로 표현된다.

$$\Lambda = \frac{\mathrm{Max}_{\Theta_0} f(x \mid \theta)}{\mathrm{Max}_{\Theta} f(x \mid \theta)} = \frac{f(x \mid \theta_d)}{f(x \mid \theta_d \cup \theta_n)} = \frac{f_d(x)}{f(x)},$$

여기서 $f(x) = \gamma f(x \mid \theta_d) + (1-\gamma) f(x \mid \theta_n)$이다. 그러므로 이 가설은 CAP 곡선을 구성하는 누적분포함수 $F(x)$, $F_d(x)$에 대응하는 각각의 확률밀도함수 $f(x)$와 $f_d(x)$에 관한 가설로 변환하여 고려할 수 있다. 이 가설을 이용하여 최적분류점을 추정하면 다음과 같다.

정리 3.1 일반화가능도비검정을 이용한 최적분류점

$$H_0 : f_d(x) \quad vs. \quad H_1 : f(x).$$

위 가설에 대하여 유의수준 α에 대한 일반화가능도비검정을 이용하여 구한 최적분류점 x_c은 다음을 만족한다.

$$x_c = F_d^{-1}(1-\alpha).$$

증명: 정리 3.1의 가설에 대하여 일반화가능도비검정방법을 실시하면, 상수 c_α와 대응하는 임계값 x_c에 대하여 제 I 종 오류 α는 다음과 같다.

$$\alpha = P[\Lambda < c_\alpha \mid \theta_d] = P[X > x_c \mid \theta_d].$$

그러므로 α가 고정된 값이므로 최적분류점을 임계값 $x_c = F_d^{-1}(1-\alpha)$으로 추정한다. 이 최적분류점은 정리 2.1에서 구한 식 (2.1)의 최적분류점과 동일하다. □

제 II 종 오류는 $\beta = P[\Lambda > c_\alpha] = F(x_c) = \gamma F_d(x_c) + (1-\gamma) F_n(x_c)$이므로 일반화가능도비검정을 이용하여 추정한 분류점에 대응하는 제 I 종과 제 II 종 오류합은 다음과 같다.

$$\gamma + (1-\gamma)\alpha + (1-\gamma) F_n(x_c). \tag{3.1}$$

일반화가능도비검정에 의한 오류합 식 (3.1)과 강력검정에 의한 오류합 식 (2.2)의 크기를 비교하면 다음과 같다.

$$\begin{aligned} &[\gamma + (1-\gamma)\alpha + (1-\gamma) F_n(x_c)] - [\alpha + F_n(x_c)] \\ &= \gamma(1-\alpha) - \gamma F_n(x_c) \\ &= \gamma[1 - (\alpha + \beta)] > 0, \end{aligned}$$

왜냐하면 오류합이 1보다 작기때문다. 그러므로 일반화가능도비검정을 이용하여 추정한 최적분류점에 대응하는 제 I 종과 제 II 종 오류합은 강력검정을 이용하여 추정한 최적분류점에 대응하는 오류합보다 항상 크다는 것을 파악할 수 있다.

3.2. 정확도를 이용한 방법

정확도를 정의한 식 (1.2)을 3.1절에서 다루는 확률밀도함수의 누적분포함수 $F_d(\cdot), F(\cdot)$의 함수로 나타낼 수 있고, 이 누적분포함수로 CAP 곡선이 표현되고 있다. $F_d(\cdot), F(\cdot)$의 함수로 정의되는 정확도로부터 유도한 최적분류점은 다음과 같다.

정리 3.2 정확도를 이용한 최적분류점

정확도를 최대로 하는 분류점 x_c은 다음을 만족한다.

$$2\gamma f_d(x_c) = f(x_c). \tag{3.2}$$

증명: 정확도를 스코어 s의 함수로 다음과 같이 표현하고,

$$\begin{aligned} \mathrm{AC}(s) &= \gamma F_d(s) + (1-\gamma)(1-F_n(s)) \\ &= 2\gamma F_d(s) - F(s) + 1 - \gamma, \end{aligned}$$

$s = F^{-1}(u)$이므로 u의 함수로 변환하고 CAP 함수인 $\mathrm{CAP}(u) = F_d(F^{-1}(u))$를 이용하면 다음과 같다.

$$\begin{aligned} \mathrm{AC}(u) &= 2\gamma F_d\left(F^{-1}(u)\right) - u + 1 - \gamma \\ &= 2\gamma \mathrm{CAP}(u) - u + 1 - \gamma, \end{aligned}$$

$\mathrm{AC}(u)$가 최대가 되는 분류점을 찾기 위해 u에 대하여 미분하고

$$\frac{\partial}{\partial u}(\mathrm{AC}(u)) = 2\gamma \left[\frac{f_d\left(F^{-1}(u)\right)}{f\left(F^{-1}(u)\right)}\right] - 1$$

이를 0으로 놓으면, 최적의 분류점 x_c는 다음 식을 만족한다.

$$\frac{f_d(F^{-1}(u))}{f(F^{-1}(u))} = \frac{f_d(x_c)}{f(x_c)} = \frac{1}{2\gamma}.$$

□

즉, $2\gamma f_d(x_c) = f(x_c)$할 때의 분류점에서 정확도가 최대가 되면서 제 I 종과 제 II 종 오류합 $\alpha + \beta$가 최소가 되는 최적분류점이 된다.

3.3. 진실율을 이용한 방법

진실율을 정의한 식 (1.3)도 3.2절에서와 같이 누적분포함수 $F_d(\cdot), F(\cdot)$의 함수로 표현 가능하며, $F_d(\cdot), F(\cdot)$의 함수로 정의되는 진실율로부터 유도한 최적분류점은 다음과 같다.

정리 3.3 진실율을 이용한 최적분류점

진실율을 최대로 하는 분류점 x_c은 다음을 만족한다.

$$f_d(x_c) = f(x_c). \tag{3.3}$$

증명: 진실율을 스코어 S의 함수로 정리하면서 $F_d(s)$와 $F(s)$로 표현하면,

$$2\mathrm{TR}(s) = \frac{1}{1-\gamma} F_d(s) - \frac{1}{1-\gamma} F(s) + 1$$

이고, CAP 함수를 이용하여 u의 함수식으로 표현하면 다음과 같다.

$$\begin{aligned} 2\mathrm{TR}(u) &= \frac{1}{1-\gamma} F_d\left(F^{-1}(u)\right) - \frac{u}{1-\gamma} + 1 \\ &= \frac{1}{1-\gamma} \mathrm{CAP}(u) - \frac{u}{1-\gamma} + 1. \end{aligned}$$

TR(u)가 최대가 되는 분류점을 찾기 위해 u에 대하여 미분하고 0으로 놓으면

$$\frac{1}{1-\gamma}\frac{f_d\left(F^{-1}(u)\right)}{f\left(F^{-1}(u)\right)}-\frac{1}{1-\gamma}=0.$$

최적의 분류점 x_c는 다음을 만족한다.

$$\frac{f_d\left(F^{-1}(u)\right)}{f\left(F^{-1}(u)\right)}=\frac{f_d(x_c)}{f(x_c)}=1.$$

□

두 확률밀도함수 $f_d(\cdot)$와 $f(\cdot)$의 값이 동일할 때의 분류점에서 진실율이 최대가 되면서 제 I 종과 제 II 종 오류합 $\alpha+\beta$가 최소가 되는 최적분류점이 된다.

정확도를 최대로 하는 최적분류점은 정리 2.2의 식 (2.3)과 정리 3.2의 식 (3.2)에서와 같이 다른 함수식으로 나타나고 있으나 동일하다. 정리 3.2에서 정확도를 최대로 분류점은 $f(x_c)=2\gamma f_d(x_c)$을 만족하는데, $f(x_c)=\gamma f_d(x_c)+(1-\gamma)f_n(x_c)$이므로 이를 다시 표현하면 $\gamma f_d(x_c)=(1-\gamma)f_n(x_c)$이 되어 정리 2.2의 식 (2.3)과 일치한다. 그리고 정리 2.3의 식 (2.4)와 [정리 3.3]의 식 (3.3)에서의 진실율을 이용한 최적분류점의 경우도 동일하다. 따라서 조건부밀도함수 $f_d(x)$를 알고 있다면, $f_n(x)$혹은 $f(x)$중 하나만 알아도 $\alpha+\beta$을 최소로 하는 최적분류점을 발견할 수 있다. 정리 2.2와 정리 3.2 그리고 정리 2.3과 정리 3.3에서는 추정하는 과정에서 서로 다른 종류의 분포함수를 사용했다는 점에 유의한다.

그러므로 2절에서의 정리 2.1부터 정리 2.3에서 구한 최적분류점은 3절에서의 정리 3.1부터 정리 3.3까지에서 구한 최적분류점과 각각 일치한다. 다만, 정리 2.1과 정리 3.1에서는 최적분류점과 제 I 종 오류는 동일하나 제 II 종 오류만이 다르다(식 (2.2)와 (3.1)의 차이). 따라서 본 연구에서는 최적분류점을 추정하는 방법으로 가설검정, 정확도 그리고 진실율을 이용하는 세가지 방법을 제안한다.

4. 정규분포에서의 최적분류점과 오류율 비교

4.1. 정규분포에서 최적분류점

조건부 확률밀도함수 $f_d(x)$와 $f_n(x)$가 정규분포인 경우인 $\phi(x\,|\,\mu_d,\sigma_d^2)$와 $\phi(x\,|\,\mu_n,\sigma_n^2)$로 각각 가정하자. 즉 부도차주의 스코어 분포는 평균과 분산이 μ_d,σ_d^2인 정규분포를 따르고, 정상차주의 분포는 평균과 분산이 μ_n,σ_n^2인 정규분포를 따른다고 가정하고 최적분류점을 구해보자. 일반적인 신용평가 연구에서와 유사하게 $\mu_d<\mu_n$ 그리고 $\sigma_d\le\sigma_n$으로 설정하면서 정규혼합(normal mixture)분포를 고려한다.

정리 2.1과 정리 3.1에서 유도한 최적분류점은 보조정리 4.1에, 정리 2.2와 정리 3.2에서 유도한 최적분류점은 보조정리 4.2에 그리고 정리 2.3과 정리 3.3에서 유도한 최적분류점은 보조정리 4.3에서 유도하였다.

보조정리 4.1 정규분포에서 가설검정을 이용한 최적분류점 스코어가 정규분포 따르는 경우에 강력검정과 일반화 가능도비검정을 이용한 분류점 x_c은 다음과 같다.

$$x_c=\Phi^{-1}\left(1-\alpha\,|\,\mu_d,\sigma_d^2\right)=\mu_d+z_{1-\alpha}\sigma_d,$$

여기서 $z_{1-\alpha}$는 표준정규분포에서 $1-\alpha$백분위수이다.

보조정리 4.2 정규분포에서 정확도를 이용한 최적분류점

스코어의 확률밀도함수가 정규분포를 따르는 경우에 정확도가 최대인 분류점 x_c은 다음과 같다:

i) $\sigma_d = \sigma_n = \sigma$ 일 경우:

$$x_c = \frac{\sigma^2}{\mu_d - \mu_n} \ln\left(\frac{1-\gamma}{\gamma}\right) + \frac{1}{2}(\mu_d + \mu_n). \tag{4.1}$$

ii) $\sigma_d \neq \sigma_n$ 인 경우:
$(1-\gamma)/\gamma \leq \exp[-(\mu_d - \mu_n)^2/2(\sigma_d^2 - \sigma_n^2)] \times \sigma_n/\sigma_d$인 조건하에서,

$$x_c = \sqrt{\frac{2\sigma_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2} \ln\frac{\sigma_d(1-\gamma)}{\sigma_n\gamma} + \frac{\sigma_d^2\sigma_n^2}{(\sigma_d^2 - \sigma_n^2)^2}(\mu_d - \mu_n)^2} - \frac{\mu_d\sigma_n^2 - \mu_n\sigma_d^2}{\sigma_d^2 - \sigma_n^2}. \tag{4.2}$$

증명: 정리 2.2의 식 (2.3)에서 $f_d(x_c)/f_n(x_c) = (\sigma_n/\sigma_d)\exp[-(x_c - \mu_d)^2/2\sigma_d^2 + (x_c - \mu_n)^2/2\sigma_n^2] = (1-\gamma)/\gamma$이다.

i) $\sigma_d = \sigma_n = \sigma$ 일 경우

$$\frac{f_d(x_c)}{f_n(x_c)} = \exp\left[\frac{x_c(\mu_d - \mu_n)}{\sigma^2} - \frac{(\mu_d^2 - \mu_n^2)}{2\sigma^2}\right] = \frac{1-\gamma}{\gamma}$$

이므로 정확도 AC를 최대로 하는 최적분류점은 다음과 같다.

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

ii) $\sigma_d \neq \sigma_n$ 인 경우

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

이므로

$$\frac{2\sigma_d^2\sigma_n^2}{\sigma_d^2 - \sigma_n^2} \ln\frac{\sigma_d(1-\gamma)}{\sigma_n\gamma} + \frac{\sigma_d^2\sigma_n^2}{(\sigma_d^2 - \sigma_n^2)^2}(\mu_d - \mu_n)^2 \geq 0$$

또는

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

의 조건하에서, 최적분류점은 다음과 같이 구한다.

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

□

보조정리 4.3 정규분포에서 진실율을 이용한 최적분류점
스코어가 정규분포 따르는 경우에 진실율이 최대인 분류점 x_c은 다음과 같다:

i) $\sigma_d = \sigma_n = \sigma$ 일 경우

$$x_c = \frac{1}{2}(\mu_d + \mu_n). \tag{4.3}$$

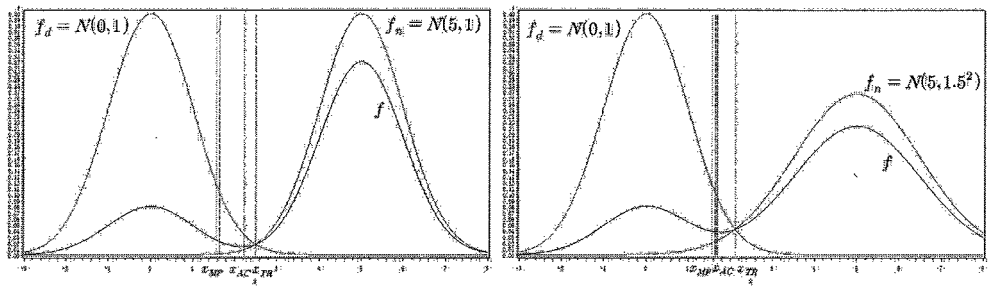

그림 **4.1.** 최적분류점(γ=0.2)

II) $\sigma_d \neq \sigma_n$ 인 경우

$$x_c = \sqrt{\frac{2\sigma_d^2\sigma_n^2}{\sigma_d^2-\sigma_n^2}\ln\frac{\sigma_d}{\sigma_n} + \frac{\sigma_d^2\sigma_n^2}{(\sigma_d^2-\sigma_n^2)^2}(\mu_d-\mu_n)^2} - \frac{\mu_d\sigma_n^2-\mu_n\sigma_d^2}{\sigma_d^2-\sigma_n^2}. \tag{4.4}$$

증명: 식 (4.3)과 (4.4)는 보조정리 4.2의 식 (4.1)과 (4.2)에서 $\gamma = 1/2$인 경우이다. □

본 연구에서 제안한 가설검정, 정확도 그리고 진실율을 이용하여 최적분류점을 추정하기 위해 다음과 같은 두 종류의 정규혼합(normal mixture)분포를 고려하자.

$$\text{(i)}\ f(x) = 0.2\,\phi(x\,|\,0,\,1) + 0.8\,\phi(x\,|\,5,\,1),$$
$$\text{(ii)}\ f(x) = 0.2\,\phi(x\,|\,0,\,1) + 0.8\,\phi(x\,|\,5,\,1.5^2).$$

부도차주의 분포는 표준정규분포이며 부도율총합 γ는 0.2로 설정한다. 그리고 첫번째 정규혼합분포에서 정상차주의 분포는 분산은 동일하나 평균이 다른 경우이고, 두번째 정규혼합분포에서 정상차주의 분포는 평균과 분산 모두 다른 경우이다. 보조정리 4.1부터 보조정리 4.3까지에서 유도한 식으로부터 가설검정, 정확도 그리고 진실율에 의한 최적분류점 x_{MP}, x_{AC}, x_{TR}은 다음과 같이 구한다.

$$\text{(i)}: x_{MP} = 1.645,\ x_{AC} = 2.223,\ x_{TR} = 2.50,$$
$$\text{(ii)}: x_{MP} = 1.645,\ x_{AC} = 1.698,\ x_{TR} = 2.12.$$

(i)과 (ii)의 경우를 그림 4.1에 나타내었다. 각 그림에 정규혼합분포 $f(x)$, 부도와 정상차주 분포 $f_d(x), f_n(x)$ 그리고 세 종류의 최적분류점 x_{MP}, x_{AC}, x_{TR}을 표현하였다. 그림 4.1을 살펴보면, 가설검정법에 의한 최적분류점은 오직 귀무가설 분포에만 의존하므로 x_{MP}는 두 경우 모두 제 I 종 오류에 의존하기 때문에 동일한 위치에 있다. 진실율에 의한 최적분류점 x_{TR}은 $f(x)$와 $f_d(x)$ 그리고 $f_n(x)$이 모두 만나는 스코어 값임을 확인할 수 있다. 이것은 정리 2.3에서 최적분류점 x_c는 $f_d(x_c) = f_n(x_c)$를 만족해야되고 정리 3.3에서 $f_d(x_c) = f(x_c)$를 만족해야 되기 때문에 세 확률밀도함수 모두가 동일한 스코어가 최적분류점이 되는 것이다.

4.2. 효율성 비교

모수가 두 개인 정규분포의 경우 평균에 대한 가설검정을 고려하는 경우의 모수공간 Θ_μ과 분산에 대한 가설검정을 고려하는 경우의 모수공간 Θ_σ으로 구분하여 살펴보자. 우선, 모수공간 Θ_μ을 $\{\theta_d = \mu_d,\ \theta_n = \mu_n\}$로 설정하고 $f_d(s)$는 표준정규분포이며 $f_n(s)$은 모평균이 μ_n이고 모분산은 1이라고 가

표 4.1. 대립가설의 평균변화에 따라 제 I 종 오류와 제 II 종 오류(γ=0.3)

		μ_n				
		1	2	3	4	5
MPT	α	0.0500	0.0500	0.0500	0.0500	0.0500
	β	0.7405	0.3613	0.0877	0.0093	0.0004
	오류합	0.7905	0.4113	0.1377	0.0593	0.0504
(GLRT)		0.8534	0.5879	0.3964	0.3415	0.3353
AC	α	0.6358	0.2822	0.1117	0.0369	0.0099
	β	0.0889	0.0773	0.0373	0.0135	0.0038
	오류합	0.7247	0.3595	0.1490	0.0504	0.0137
TR	α	0.3085	0.1587	0.0668	0.0228	0.0062
	β	0.3085	0.1587	0.0668	0.0228	0.0062
	오류합	0.6170	0.3174	0.1336	0.0456	0.0124

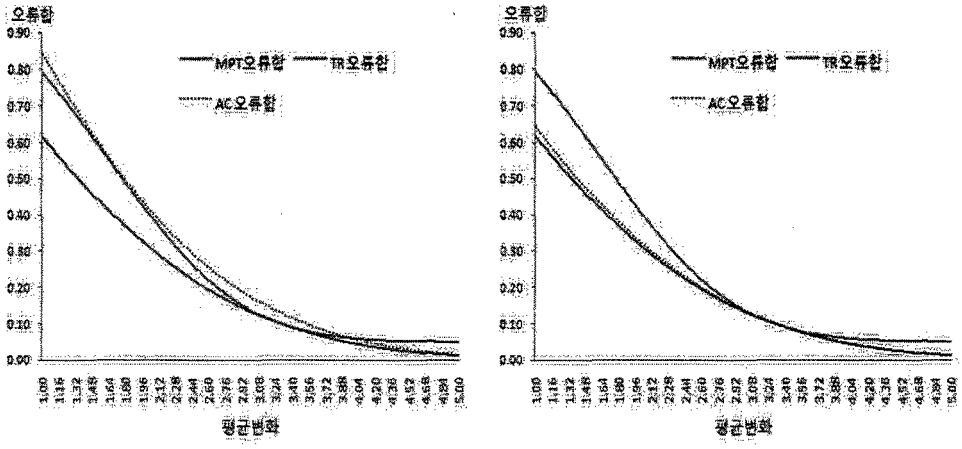

그림 4.2. 평균변화에 따른 오류합 변화(γ=0.2, 0.4)

정하자. 여기서 대립가설의 μ_n이 1부터 5까지 변하고 γ가 0.2부터 0.4까지 변할 때 강력검정과 일반화 우도비검정으로 최적분류점을 얻는 방법 그리고 정확도와 진실율을 이용하여 최적분류점을 얻는 방법을 통하여 제 I 종 오류와 제 II 종 오류 각각과 그 합을 구하여 표 4.1에 정리하였다. 표 4.1의 첫 번째 열에 강력검정방법, 정확도와 진실율에 대하여 각각 MPT, AC, TR로 나타내고, 일반화가능도비검정방법에 대응하는 제 I 종 오류는 강력검정의 오류와 동일하므로 오류합만을 MPT 행의 마지막(점선 아래)에 GLRT로 나타내었다(표 4.2에서도 동일한 형식을 취했음). γ가 0.3일때만의 결과를 표 4.1에 나타내고, γ가 0.2와 0.4일때의 결과를 그림 4.2에 구현하였다.

대립가설의 모평균 μ_n이 1부터 5까지 증가하는 표 4.1을 전반적으로 살펴보면, 강력검정법(MPT)과 진실율(TR)을 이용하여 구한 최적분류점으로부터의 오류합은 γ의 값에 의존하지 않으며 일반화가능도비검정법(GLRT)과 정확도(AC)를 이용한 방법에서는 γ의 값에 의존한다. TR을 이용한 최적분류점은 두 가설분포의 만나는 스코어에서 결정되고 분산을 동일하게 두었기 때문에 제 I 종 오류와 제 II 종 오류 값이 동일하게 나타난다. AC는 γ값이 증가할수록 제 I 종 오류가 감소하고 제 II 종 오류는 증가하지만 두 오류합은 점차 감소한다. 세가지 방법 모두 모평균이 증가할수록 오류합은 감소한다. 그 중에서도 제 I 종 오류만을 설명하는 GLRT에 대응하는 오류합이 항상 제일 크고, TR을 이용하여 구한 최적분류

표 **4.2.** 대립가설의 분산변화에 따라 제 I 종 오류와 제 II 종 오류(γ=0.3)

		σ_n				
		1	1.5	2	2.5	3
MPT	α	0.0500	0.0500	0.0500	0.0500	0.0500
	β	0.3613	0.4065	0.4296	0.4435	0.4529
	오류합	0.4113	0.4565	0.4796	0.4935	0.5029
(GLRT)		0.5879	0.6195	0.6357	0.6455	0.6520
AC	α	0.2822	0.3293	0.3077	0.2565	0.2084
	β	0.0773	0.1495	0.2270	0.2952	0.3461
	오류합	0.3595	0.4788	0.5347	0.5517	0.5545
TR	α	0.1587	0.1385	0.1079	0.0843	0.0679
	β	0.1587	0.2714	0.3515	0.4015	0.4328
	오류합	0.3174	0.4099	0.4594	0.4858	0.5007

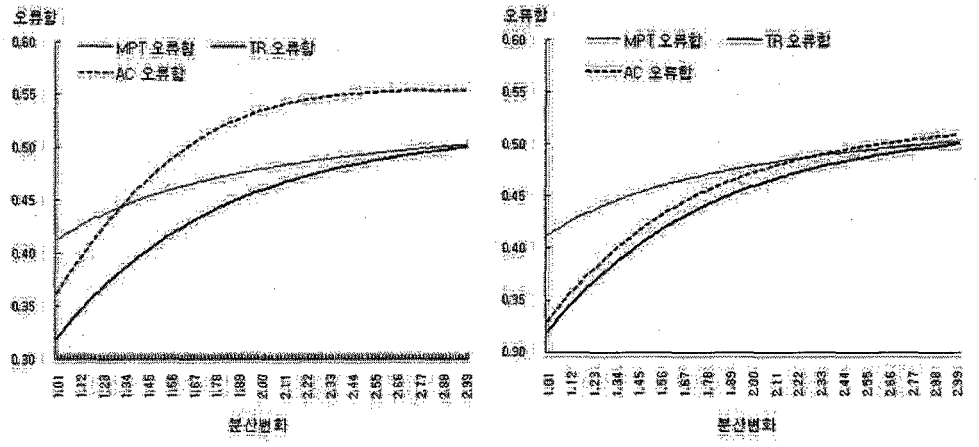

그림 **4.3.** 분산변화에 따른 오류합 변화(γ=0.3, 0.4)

점으로부터의 오류합이 가장 작다. MPT에 대응하는 오류합은 TR에 대응하는 오류합보다 항상 크지만, μ_n이 최적분류점값의 두배인 $\mu_n = 2 \times z_{0.95} = 3.29$에서는 동일한 오류합을 갖는다. AC에 대응하는 오류합은 TR에 대응하는 오류합보다 크지만 γ값이 증가할수록 TR에 대응하는 오류합에 접근한다. 그리고 MPT에 대응하는 오류합보다 작아지는 경향이 있다.

다음으로는 모수공간 Θ_σ을 $\{\theta_d = {\sigma_d}^2,\ \theta_n = {\sigma_n}^2\}$로 설정하고 $f_d(s)$는 표준정규분포이며 $f_n(s)$은 모분산이 ${\sigma_n}^2$이며 모평균은 2인 정규분포로 가정하자. 여기서 대립가설의 σ_n이 1부터 3까지 0.5씩 증가하고 γ가 0.3부터 0.4까지 변할 때 강력검정과 일반화우도비검정으로 최적분류점을 얻는 방법과 정확도와 진실율을 이용하여 최적분류점을 얻는 방법을 통하여 제 I 종 오류와 제 II 종 오류 각각과 그 합을 구하여 표 4.2에 정리하였다. 여기서 γ값의 변화범위를 0.3부터 0.4까지로 설정한 것은 AC의 최적분류점이 보조정리 4.2의 ii)의 조건을 만족해야 하기 때문이다. γ가 0.3일때만의 결과를 표 4.2에 나타내고, γ가 0.3과 0.4일때의 결과를 그림 4.3에 구현하였다.

대립가설의 모분산 ${\sigma_n}^2$이 증가하는 표 4.2를 전반적으로 살펴보면, GLRT와 AC를 이용한 방법에서는 γ의 값에 의존하지만 MPT와 TR을 이용하여 구한 최적분류점으로부터의 오류합은 γ의 값에 의존하지 않는다. 모분산이 커질수록 세가지 방법에 의한 오류합은 증가한다. 그 중에서도 제 I 종 오류만을 설명

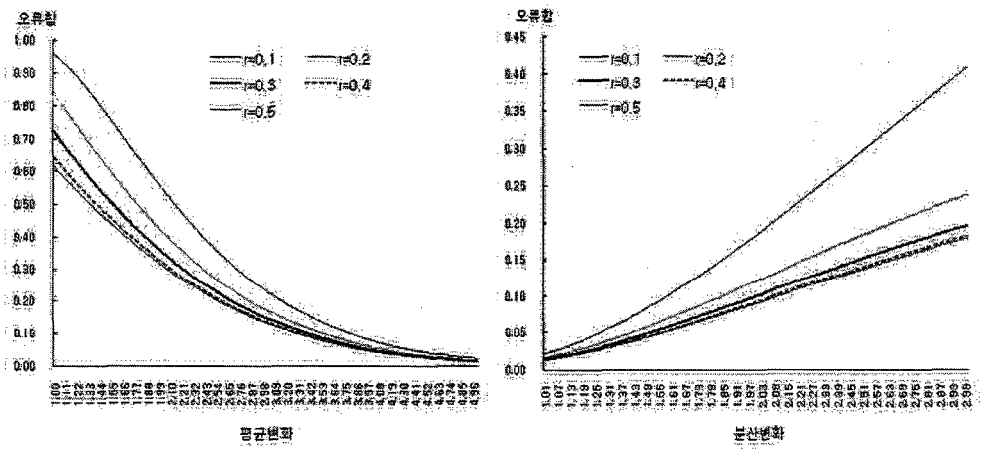

그림 **4.4.** 부도율종합과 AC 오류합: 평균 변화, 분산 변화

하는 GLRT에 대응하는 오류합이 항상 제일 크고 TR을 이용하여 구한 최적분류점으로부터의 오류합이 가장 작다. MPT에 대응하는 오류합은 TR에 대응하는 오류합보다 항상 크며 σ_n가 증가할수록 유사한 값으로 접근한다. AC에 대응하는 오류합은 TR에 대응하는 오류합보다 크지만 γ값이 증가할수록 TR에 대응하는 오류합에 접근한다. γ값이 0.3일때와 0.4일때 분산변화에 따른 오류합의 변화는 그림 4.3에 제시하였다.

AC를 이용한 최적분류점으로 부터 구한 오류의 합이 γ와 모평균, 모분산이 증가하면서 변화하는 현상을 그림 4.4에 나타내었다. 그림 4.4의 왼쪽을 살펴보면, 대립가설의 모평균이 1부터 5까지 증가할수록 오류합은 감소하고 γ의 값이 0.5까지 커질수록 감소하는 속도는 작아지는 것을 발견할 수 있다. 이런 현상은 표 4.1을 설명할 때와 유사하다.

다음으로 모분산의 변화 현상을 나타낸 그림 4.4의 오른쪽을 살펴보자. 여기에서 모분산이 증가하면 AC의 최적분류점이 보조정리 4.2의 ii)의 조건을 만족하지 못하기 때문에, 대립가설의 평균은 5로 두었다. 모분산이 커질수록 오류합은 증가하고 γ의 값이 0.5까지 커질수록 증가하는 속도는 작아지는 것을 발견할 수 있다. 이는 표 4.2를 설명할 때 발견한 현상이다.

이와 같은 결과를 요약해보면 세가지 방법 중 진실율을 이용하여 얻은 최적분류점으로부터 구한 제 I 종 오류와 제 II 종 오류합이 가장 작다는 것을 파악할 수 있다. 특히 진실율 방법은 제 I 종 오류를 고정시키고 가장 작은 제 II 종 오류를 생성하는 강력검정 방법보다 작은 오류합을 보여준다. 제 I 종 오류를 고정시킨 강력검정 방법을 이용하여 구한 최적분류점으로부터의 제 II 종 오류는 모평균이 증가할수록 작아지지만, 진실율을 이용해 얻은 최적분류점으로부터 구한 제 I 종 오류와 제 II 종 오류합보다는 크게 나타난다. 즉, 세가지 방법으로 추정한 최적분류점 중에서 진실율을 이용한 최적분류점은 제 I 종 오류와 제 II 종 오류의 합을 가장 작게하는 방법이라고 결론내릴 수 있다.

5. 결론

최근 신용평가 연구에서 부도와 정상차주를 판별하기 위한 최적분류점에 대한 연구 중에서 Vuk과 Curk (2006)은 최적분류점을 찾기 위해 동일한 정확도로 표현되는 선형식과 ROC 곡선(또는 CAP 곡선)의 접점이 정확도를 최대로 하는 최적분류점이 됨을 보였다. 또한 홍종선과 최진수 (2009)는 정상차주보다

부도차주의 수가 크게 적을 경우 기존의 정확도 통계량이 적절하지 않다고 지적하고 정확도의 대안적인 측도로서 진실율을 제안하면서 Vuk과 Curk (2006)의 방법과 유사하게 최적분류점을 구하였다.

본 연구는 제 I 종 오류와 제 II 종 오류합을 최소로 하는 최적분류점을 추정하는 방법으로 세가지 이론을 제안하였다. 첫 번째 방법은 통계적 가설검정방법으로 정리 2.1과 정리 3.1를 통해 각각 강력검정과 일반화가능도비검정을 이용하여 최적분류점을 구했다. 두 번째 방법은 정확도를 이용하는 방법으로, 정리 2.2와 정리 3.2를 통해 정확도를 최대로 하는 최적분류점을 찾는 이론을 제시했고 이를 증명했다. 세 번째 방법으로 진실율을 이용하는 방법으로 정리 2.3과 정리 3.3를 통해 진실율을 최대로 하는 최적분류점을 찾는 방법을 제시했다. 세가지 방법을 통해 추정한 최적분류점에 대응하는 제 I 종과 제 II 종의 오류를 정의하였다. 스코어가 정규분포를 따른다고 가정한 후 다양한 상황에서 오류합을 구하고 그 크기를 비교하였다.

첫 번째로는 귀무가설의 확률밀도함수 $f_d(s)$를 표준정규분포로 두고, 대립가설의 확률밀도함수 $f_n(s)$의 모평균 μ_n를 1-5까지 변화하면서(모분산$\sigma_n{}^2$을 1로 고정) 나타나는 현상을 살펴보았다. 그리고 부도율총합 γ이 0.2-0.4까지 변화시키면서 최적분류점을 구한 후 이 분류점에 대응하는 오류합을 비교분석하였다. 세가지 방법으로 구한 오류합은 전반적으로 모평균과 부도율총합이 증가할수록 감소하고 0으로 수렴한다. 그중에서도 진실율을 이용하여 구한 최적분류점에서 제 I 종 오류와 제 II 종 오류합이 가장 작았다. 강력검정의 최적분류점에 대응하는 오류합은 진실율을 이용하는 방법에 대응하는 오류합보다 항상 크지만 대립가설의 모평균 μ_n이 $1-\alpha$ 분위수에 대응하는 기각역의 두 배인 경우에는 일치했다. 정확도에 대응하는 오류합은 대립가설의 모평균이 증가할수록 그리고 γ값이 증가할수록 진실율에 대응하는 오류합에 접근한다는 것을 발견하였다. 참고로 제1종 오류만을 고려한 일반화가능도비검정법의 최적분류점에서 오류합이 가장 크게 나타났다.

두 번째는 대립가설의 모평균 μ_n을 2로 고정된 상태에서 σ_n를 1부터 3까지 그리고 γ값을 0.3부터 0.4까지 변화시키면서 구한 최적분류점에 대응하는 오류합을 비교분석하였다. 세가지 방법으로 구한 오류합은 전반적으로 모분산이 증가할수록 증가한다. 그리고 γ값이 증가하면 세가지 방법의 오류합은 일치한다(가설방법에서 GLRT방법 제외). 그중에서도 진실율을 이용해 추정한 최적분류점에 대응하는 오류합은 항상 다른 방법에 대응하는 오류합보다 작은 값을 갖는다. 정확도를 이용하여 구한 최적분류점과 강력검정으로 구한 최적분류점에 대응하는 오류합들은 모분산과 γ가 증가할수록 진실율에 대응하는 오류합에 수렴한다. 반면, 일반화가능도비검정에 대응하는 오류합은 다른 방법에 비해 항상 크게 나타났고 γ값이 클수록 크게 나타났다.

결론으로 제 I 종 오류가 고정된 강력검정방법에서 모평균이 증가할수록 제 II 종 오류가 작아지지만, 진실율을 이용해 얻은 최적분류점으로부터 구한 제 I 종 오류와 제 II 종 오류합보다는 크게 나타난다. 정확도에 대응하는 오류합은 γ값의 증가에 따라 진실율에 대응하는 오류합과 비슷하게 접근하지만 정확도에 대응하는 오류합이 진실율에 비교하여 크거나 같았다. 즉, 진실율을 이용하여 구한 최적분류점이 세가지 방법 중 가장 작은 오류합을 갖는 최상의 방법이라고 결론내릴 수 있다.

참고문헌

홍종선, 최진수 (2009). ROC와 CAP 곡선에서의 최적 분류점, <응용통계연구>, **22**, 911-922.

Berry, M. J. A. and Linoff, G. (1999). *Data Mining Techniques: For Marketing, Sales, and Customer Support*, Morgan Kaufmann Publishers.

Drummond, C. and Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance, *Machine Learning*, **65**, 95-130.

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

Fawcett, T. (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, *HP Laboratories*, 1501 page Mill Road, Palo Alto, CA 94304.

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

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

Provost, F. and Fawcett, T. (1997). Analysis and visualization of classifier performance comparison under imprecise class and cost distributions, In: *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Menlo park, CA, 43–48.

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.

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

Stein, R. M. (2005). The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing, *Journal of Banking and Finance*, **29**, 1213–1236.

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.*

Tasche, D. (2009). Estimating discriminatory power and PD curves when the number of defaults is small, *arXiv.org, eprint arXiv:0905.3928v1.*

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

Zou, K. H. (2002). Receiver Operating Characteristic Literature Research, On-line bibliography available from: http://www.spl. harvard.edu/pages/ppl/zou/roc.html.

Optimal Thresholds from Mixture Distributions

Chong Sun Hong[1] · Jae Seon Joo[2] · Jin Soo Choi[3]

[1]Department of Statistics, Sungkyunkwan University
[2]Statistics and Panel Center, Korean Women's Development Institute
[3]Research Institute of Applied Statistics, Sungkyunkwan University

(Received September 2009; accepted November 2009)

Abstract

Assuming a mixture distribution for credit evaluation studies, we discuss estimating threshold methods to minimize errors that default borrowers are predicted as non defaults or non defaults are regarded as defaults. A method by using statistical hypotheses tests, the most powerful test and generalized likelihood ratio test, for the probability density functions which are defined with the score random variable and the parameter space consisted of only two elements such as the default and non default states is proposed to estimate a threshold. And anther optimal thresholds to maximize classification accuracy measures of the accuracy and the true rate for ROC and CAP curves are estimated as equations related with these probability density functions. Three kinds of optimal thresholds in terms of the hypotheses testing, the accuracy and the true rate are obtained from normal random samples with various means and variances. The sums of the type I and type II errors corresponding to each optimal threshold are obtained and compared. Finally we discuss about their efficiency and derive conclusions.

Keywords: Accuracy, CAP, default, discriminatory, error, likelihood ratio, most powerful, ROC, score, threshold, true rate.

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