응용통계연구 (2010)
23(5), 943–953
DOI: 10.5351/KJAS.2010.23.5.943

비정규 혼합분포에서의 최적분류점

홍종선[1] · 주재선[2]

[1]성균관대학교 통계학과, [2]한국여성정책연구원 통계패널센터

(2010년 6월 접수, 2010년 8월 채택)

요 약

신용평가연구에서 확률변수 스코어와 정상과 부도상태의 모수공간으로 정의된 혼합분포에서 확률밀도함수의 관계식으로 최적분류점을 추정하고 이에 대응하는 오류합의 크기를 비교하는 연구가 정규분포의 가정하에 이루어져있는데 본 연구에서는 비정규분포인 와이블, 로지스틱 그리고 감마분포로 확장하여 가설검정을 이용하는 방법과 전체정확도와 진실율을 최대화하는 방법에 의한 최적분류점을 각각 구하고 최적분류점에 대응하는 제 I 종과 제 II 종 오류합의 크기를 비교하여 효율성을 비교 토론한다.

주요용어: 신용, 부도, 판별, 평가, 최적분류점, 전체정확도, 진실율.

1. 서론

두 분포함수의 혼합분포로부터 판별력을 극대화하는 분류점(threshold, cut-off)을 추정하는 최근의 연구는 차주(borrower)의 미래상태인 부도상태(default) 혹은 정상상태(non-default)에 대한 예측력을 최대화하는 신용평가모형에서 많이 활용되고 있다. 신용평가모형에서 차주의 신용가치를 기준으로 대출상환능력에 따라 부도와 정상상태를 판별하는 문제를 고려하자. 확률변수 X는 스코어 변수로 연속형 실수값이다. 모수공간은 부도와 정상상태로 가정하여 $\Theta = \{\theta_d, \theta_n\}$로 정의한다. $f_d(x)$와 $f_n(x)$을 각각 차주의 부도와 정상상태에서 스코어의 조건부 확률밀도함수 $f(x|\theta_d)$와 $f(x|\theta_n)$로 정의하며, 스코어 확률밀도함수 $f(x)$는 다음과 같이 가정한다.

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

여기서 γ는 전체부도율(total probability of default)이다.

신용평가모형에서 최적분류점 연구는 ROC와 CAP곡선을 기반으로 활용된다 (Hanley와 McNeil, 1982; Berry과 Linoff, 1999; Provost와 Fawcett, 1997; Sobehart와 Keenan, 2001; Zou, 2002; Engelmann 등, 2003; Drummond와 Holte, 2006; Tasche, 2006). ROC와 CAP곡선은 최적의 예측력(prediction power)을 탐색하기 위한 그래픽적인 방법으로, 최적분류점은 이들을 구성하는 누적분포함수인 $F_n(x)$, $F_d(x)$ 그리고 $F(x)$를 이용하여 분류정확도(classification accuracy)를 최대로하는 방법으로 탐색되어진다. 그리고 분류정확도 통계량은 전체정확도(Total Accuracy; TA), 홍종선과 최진수 (2009)와 홍종선 등 (2010)의 진실율(True Rate; TR), Youden (1950)의 유덴지수(Youden Index), Bairagi와 Suchindran (1989)의 SSS측도(Sum of Sensitivity and Specificity), Perkins과 Schisterman

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

(2006)의 수정된 $(0,1)$에서의 최단거리(Amended Closest to $(0,1)$) 측도 등의 기준이 많이 사용된다. 이들 기준 중 전체정확도는 분류정확도를 측정하는 방법으로 오랫동안 활용되고 있고, 진실율은 유덴지수, SSS측도, 수정된 $(0,1)$에서의 최단거리 측도와 선형관계를 가지고 있다. 홍종선 등 (2010)은 누적분포함수 $F_n(x)$, $F_d(x)$ 그리고 $F(x)$를 바탕으로 고정된 유의수준에 대한 가설검정을 이용하는 방법과 대표적인 분류정확도인 전체정확도 또는 진실율을 최대화하는 최적분류점(optimal threshold)을 추정하는 방법을 제안하였고, 정규분포에 적용하여 최적분류점과 최적분류점에 대응하는 제 I 종 오류(α)와 제 II 종 오류(β)를 구하였다. 실제로 사용되는 스코어 변수의 분포함수는 정규분포와 다르게 양수의 왜도계수와 높은 첨도계수를 가진 비대칭인 분포가 많다. 이에 본 연구는 홍종선 등 (2010)의 연구를 비정규분포인 와이블, 로지스틱, 감마분포로 확장하여 비정규분포 혼합에서의 최적분류점을 추정하고 각 최적분류점에 대응하는 오류합을 비교 토론하고자 한다. 본 연구의 구성은 다음과 같다. 2절에서는 전체정확도와 진실율을 누적분포함수를 이용하여 설명하고 이 두가지 기준을 통하여 최적분류점을 추정하는 방법을 소개하며 강력검정과 일반화가능도비검정의 가설검정 방법과 같이 요약한다. 3절은 제안한 최적분류점 추정방법을 통해 세 종류의 비정규 분포인 와이블, 로지스틱, 감마분포의 혼합분포에서의 최적분류점을 추정한 다음 이에 대응하는 오류합을 구하고 비교 토론한다. 마지막 4절에서는 이를 종합하여 결론을 유도한다.

2. 최적분류점 추정방법

전체정확도(TA)와 균등정확도(balanced accuracy; Velez 등, 2007)로 불리는 진실율(TR)은 식 (2.1)과 (2.2)와 같이 정의한다.

$$\begin{aligned} \text{TA} &= \gamma F_d(x) + (1-\gamma)(1-F_n(x)) \\ &= 2\gamma F_d(x) - F(x) + 1 - \gamma, \\ \text{TR} &= \frac{1}{2}[F_d(x) + (1-F_n(x)] \\ &= \frac{1}{2(1-\gamma)}[F_d(x) - F(x)] + \frac{1}{2}. \end{aligned} \tag{2.1}$$

분류정확도인 전체정확도 또는 진실율을 최대로하는 분류점 즉, max(TA) 또는 max(TR)이 되는 분류점이 최적분류점(optimal threshold)이 된다. Vuk과 Curk (2006)과 홍종선과 최진수 (2009)은 각각 전체정확도와 진실율을 이용한 식 (2.1)과 (2.2)의 선형식과 ROC와 CAP 곡선의 접점으로부터 최적분류점을 발견하였고, 홍종선 등 (2010)은 X와 Y축 좌표가 각각 $(F_n(x), F_d(x))$와 $(F(x), F_d(x))$로 구성된 ROC와 CAP 곡선에 대해 가설검정을 이용하고, 전체정확도와 진실율을 최대로하는 새로운 방법을 제안하였다.

새로 제안한 방법 중 첫째는 가설검정을 이용하는 방법으로, 확률변수 X를 분류점을 나타내는 변수이며 확률밀도함수를 스코어 변수의 확률밀도함수인 식 (1.1)과 동일하다고 가정하고 다음과 같은 가설을 고려하였다.

$$H_0 : f_d(x) \quad vs. \quad H_1 : f_n(x), \tag{2.2}$$

$$H_0 : f_d(x) \quad vs. \quad H_1 : f(x). \tag{2.3}$$

유의수준 α에서 가설 (2.3)과 (2.4)에 대해 각각 강력검정(most powerful test)과 일반화가능도비검정(generalized likelihood ratio test)을 이용하여 구한 최적분류점 x_o는 식 (2.5)과 같음을 보였다.

$$x_o = F_d^{-1}(1-\alpha). \tag{2.4}$$

두 번째로 ROC와 CAP 곡선에 대하여 전체정확도를 이용한 최적분류점 x_o는 $\gamma f_d(x_o) = (1-\gamma)f_n(x_o)$의 조건을 만족한다. 그리고 세번째로 ROC와 CAP 곡선에 대하여 진실율 기준에서 최적분류점 x_o는 $\gamma f_d(x_o) = f_n(x_o)$과 같이 확률밀도함수의 관계로 나타난다. $f(x_o)$가 $f_d(x_o)$와 $f_n(x_o)$의 혼합분포이므로, 위의 두 식은 다음과 같이 하나의 관계식으로 정리된다.

$$2\gamma f_d(x_o) = f(x_o). \tag{2.5}$$

3. 주요 혼합분포에서의 최적분류점과 오류합

홍종선 등 (2010)은 부도차주의 분포가 $N(\mu_d, \sigma_d^2)$이고 정상차주의 분포가 $N(\mu_n, \sigma_n^2)$인 정규혼합에서 제 I 종 오류와 제 II 종 오류의 합을 최소화하는 분류점을 세가지 방법에 대해 제시하고 이들 방법으로 구한 최적분류점에 대응하는 오류합의 효율성을 토론하였다. 신용평가에서 스코어 분포는 한쪽으로 치우쳐서 대칭이 아니거나, 부도와 정상 스코어 분포의 형태가 서로 다른 경우들이 대부분이다. 본 연구에서는 스코어 분포가 정규분포가 아닌 와이블분포, 로지스틱분포, 감마분포일 경우에 2절에서 토론한 최적분류점을 구하고 오류합을 최소로 하는 효율성이 높은 기준을 토론한다.

3.1. 와이블분포

와이블분포(Weibull distribution)는 모수의 크기에 따라 다양한 형태의 분포에 접근할 수 있는 장점이 있어 모형 적합에 자주 사용된다. 2-모수 와이블분포의 조건부 분포함수와 조건부 밀도함수는 다음과 같이 주어진다.

$$F(x; b, \lambda) = 1 - \exp\left[-\left(\frac{x}{\lambda}\right)^b\right], \qquad f(x; b, \lambda) = \frac{bx^{b-1}}{\lambda^b}\exp\left[-\left(\frac{x}{\lambda}\right)^b\right],$$

여기서 b는 형태모수이고 λ는 척도모수이며 $x > 0$, $b > 0$, $\lambda > 0$이다. 척도모수를 $\lambda_d < \lambda_n$로 설정하고 가설검정, 전체정확도, 진실율을 이용하여 구한 와이블 혼합분포에서의 최적분류점은 다음과 같다.

보조정리 3.1. 와이블분포에서 가설검정을 이용한 최적분류점

유의수준 α에서의 강력검정과 일반화가능도비검정을 이용한 분류점은 다음과 같다.

$$x_o = F_d^{-1}(1-\alpha) = \lambda_d[1 - \ln(\alpha)]^{\frac{1}{b_d}}.$$

와이블분포의 역함수는 $F^{-1}(u) = \lambda[-\ln(1-u)]^{1/b}$이므로 식 (2.5)를 이용하여 구한다. 2-모수 와이블분포는 매우 다양한 형태의 분포를 보여 주지만, 최적분류점을 수식적으로 표현하는 것은 쉽지 않으므로 분포의 형태모수(shape parameter)를 1로 설정($b = 1$)하여 지수분포로 가정할 때 전체정확도(TA)와 진실율(TR)을 최대로하는 최적분류점을 다음과 같이 제시하였다.

보조정리 3.2. $b = 1$인 와이블분포에서 전체정확도를 최대로하는 최적분류점

전체정확도가 최대인 분류점은 다음과 같다.

$$x_o = \frac{\lambda_d\lambda_n}{\lambda_n - \lambda_d}\ln\frac{\lambda_n}{\lambda_d}\frac{\gamma}{1-\gamma}.$$

보조정리 3.3. $b = 1$인 와이블분포에서 진실율을 최대로하는 최적분류점

표 **3.1.** 와이블분포에서 λ의 변화와 오류합($\gamma = 0.3$)

$\alpha/\beta/$ 오류합	WEI$(1,1)$ *vs.* WEI$(3,1)$	WEI$(1,1)$ *vs.* WEI$(4,1)$	WEI$(1,1)$ *vs.* WEI$(5,1)$
MPT	0.05000	0.05000	0.05000
	0.63160	0.52713	0.45072
	0.68160	0.57713	0.50072
TA	0.68594	0.48740	0.38571
	0.11808	0.16445	0.17348
	0.80402	0.65185	0.55919
TR	0.19245	0.15749	0.13375
	0.42265	0.37004	0.33126
	0.61510	0.52753	0.46501

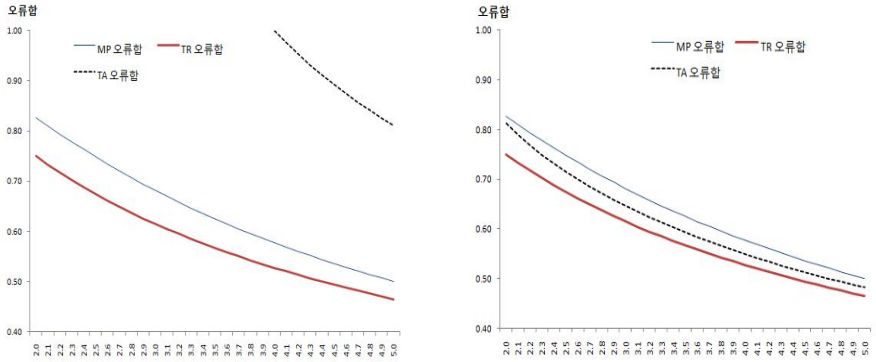

그림 **3.1.** 와이블분포에서 λ의 변화와 오류합($\gamma = 0.2, 0.4$)

진실율0이 최대인 분류점은 다음과 같다.

$$x_o = \frac{\lambda_d \lambda_n}{\lambda_n - \lambda_d} \ln \frac{\lambda_n}{\lambda_d}.$$

식 (2.6)을 이용하여 $f(x_o; 1, \lambda_d)/f(x_o; 1, \lambda_n) = (\lambda_n/\lambda_d)\exp(x_o/\lambda_n - x_o/\lambda_d)$을 각각 $(1-\gamma)/\gamma$과 1로 놓고 정리하면 얻을 수 있다.

와이블분포에서 형태모수가 $b = 1$인 경우, 즉 와이블분포에서 귀무가설은 $\lambda = 1$로 두고 대립가설을 가 2부터 5까지 변할 때 세가지 방법으로 구한 최적분류점에 대해 제 I 종 오류와 제 II 종 오류 각각과 그 합을 구하여 표 3.1에 정리하고 이를 그림 3.1에 표현하였다. 일반적으로 신용평가연구에서의 부도율 γ은 0.1 이하를 갖게 되나, γ가 0.1 이하에서는 TR과 TA의 오류합 차이가 매우 크게 나타났다. 이에 γ는 0.2, 0.4, 0.5인 경우에 대해서 각각 표와 그림으로 제시하였다. 표 3.1에서 전체부도율이 $\gamma = 0.3$일 때 척도모수(scale parameter) λ가 증가할수록 오류합은 감소하였다. 세가지 방법 중 TA에 대응하는 오류합이 제일 크고, TR을 이용하여 구한 오류합이 가장 작게 나타났으며, MPT에 대응하는 오류합은 TA보다는 작지만 TR보다는 크게 나타났다. 하지만 λ의 증가는 평균과 분산을 동시에 증가시킴에 따라, 오류합은 큰 폭의 하락은 보이지 않았다. γ에 의존하는 TA는 γ에 따라 값의 클수록 오류합이 낮게 나타났다.

그림 3.1에서 γ에 의존하지 않는 MPT와 TR은 동일한 오류합을 보이지만, TA는 $\gamma = 0.2$와 $\gamma = 0.4$일

때 오류합의 차이가 매우 크게 나타난다. 형태모수가 $b = 1$인 와이블분포에서 세가지 방법으로 구한 최적분류점에 대응하는 오류합은 진실율이 항상 가장 작았고 γ에 의존하는 전체정확도는 γ에 따라 오류합의 변화가 매우 크게 나타났다.

3.2. 로지스틱분포

로지스틱분포(logistic distribution)의 조건부 분포함수와 조건부 밀도함수는 각각 다음과 같이 주어진다.

$$F(x;\mu,\lambda) = \frac{\exp((x-\mu)/\lambda)}{1+\exp((x-\mu)/\lambda)}, \qquad f(x;\mu,\lambda) = \frac{\exp((x-\mu)/\lambda)}{\lambda[1+\exp((x-\mu)/\lambda)]^2},$$

여기서 $\lambda > 0$이다. 로지스틱 혼합분포에서 세가지 방법에 의한 최적분류점은 다음과 같다.

보조정리 3.4. 로지스틱분포에서 가설검정을 이용한 최적분류점

유의수준 α에서의 강력검정과 일반화가능도비검정을 이용한 분류점은 다음과 같다.

$$x_o = \mu_d + \lambda_d \ln\left[\frac{(1-\alpha)}{\alpha}\right].$$

로지스틱분포의 역함수는 $F^{-1}(u) = \mu + \lambda \ln[u/(1-u)]$이므로 식 (2.5)에 의해 최적분류점 x_o는 쉽게 구해진다. 로지스틱분포에서 전체정확도와 진실율을 최대로하는 최적분류점은 형태모수를 $\lambda = 1$로 두고 추정하면 보조정리 3.4와 보조정리 3.5와 같이 정리된다.

보조정리 3.5. 로지스틱분포에서 전체정확도를 최대로하는 최적분류점

전체정확도가 최대인 분류점은 다음과 같다.

$$x_o = \mu_d + \ln\left[\frac{\sqrt{\frac{\gamma}{1-\gamma}\exp(\mu_n-\mu_d)}-1}{\exp(\mu_n-\mu_d)-\sqrt{\frac{\gamma}{1-\gamma}\exp(\mu_n-\mu_d)}}\right],$$

여기서 $\exp(\mu_n - \mu_d) > \gamma/(1-\gamma)$을 만족해야 한다.

보조정리 3.6. $\lambda = 1$인 로지스틱분포에서 진실율을 최대로하는 최적분류점

진실율이 최대인 최적분류점은 다음과 같다.

$$x_o = \frac{\mu_d + \mu_n}{2}.$$

보조정리 3.5와 보조정리 3.6은 $f_d(x_o;\mu_d,1)/f_n(x_o;\mu_n,1)$을 각각 $(1-\gamma)/\gamma$와 1로 전개하면 얻을 수 있다.

로지스틱분포에서 얻은 최적분류점에 대응하는 제 I 종 오류와 제 II 종 오류 각각과 그 합을 계산하여 표 3.2과 그림 3.2에 정리하였다. 귀무가설은 분산을 고정하고 평균이 0인 로지스틱분포로 가정한다. 그리고 분산을 동일하게 고정하고 대립가설을 2부터 5까지 변할 때 세가지 방법으로 구한 최적분류점에서의 오류합을 비교하였다. 부도율 γ는 와이블 분포와 오류합의 비교가 가능하도록 0.2, 0.4, 0.5인 경우에 대해서 각각 표와 그림으로 제시하였다. $\gamma = 0.3$일 때 세가지 방법으로 구한 최적분류점에서의 오

표 **3.2.** 로지스틱분포에서 평균변화와 오류합($\gamma = 0.3$)

$\alpha/\beta/$ 오류합	$L(0,1)$ *vs.* $L(2,1)$	$L(0,1)$ *vs.* $L(3,1)$	$L(0,1)$ *vs.* $L(4,1)$	$L(0,1)$ *vs.* $L(5,1)$
MPT	0.0500	0.05000	0.05000	0.05000
	0.7200	0.48611	0.25816	0.11349
	0.7700	0.53611	0.30816	0.16349
TA	0.49338	0.30630	0.19193	0.11945
	0.12201	0.10133	0.07159	0.04732
	0.61539	0.40763	0.26352	0.16677
TR	0.26894	0.18243	0.11920	0.07586
	0.26894	0.18243	0.11920	0.07586
	0.53788	0.36485	0.23841	0.15172

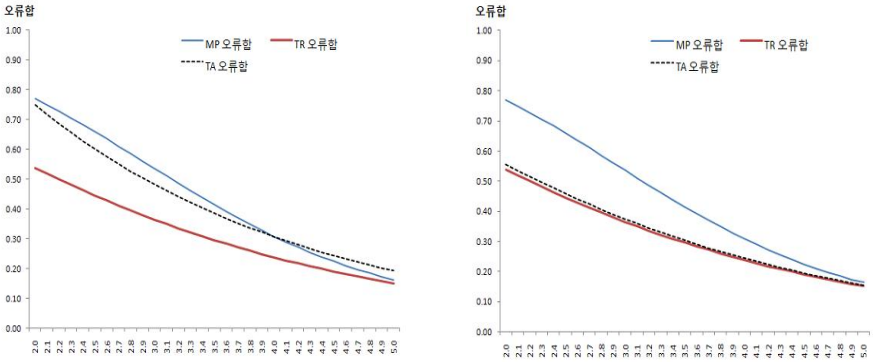

그림 **3.2.** 로지스틱분포분포에서의 평균변화와 오류합($\gamma = 0.2, 0.4$)

류합을 보면 평균차이가 2일 경우 TR < TA < MPT 순으로 작게 나타나지만, 평균차이가 5일 경우 TR < MPT < TA 순으로 나타난다. 즉 두 분포의 평균차이가 작을 때는 TA에서 구한 최적분류점에서의 오류합이 MPT에서 보다 작지만, 두 분포의 평균차이가 클 때는 TA보다 MPT에서 구한 최적분류점에서의 오류합이 더 작게 나타난다. 하지만 세가지 방법 중 TR 기준을 최대로하는 최적분류점은 다른 기준에서 구한 최적분류점보다 항상 작게 난다. γ에 의존하는 전체정확도(TA)는 γ가 0.5에 가까울수록 TA에서 구한 최적분류점의 오류합에 빠른 속도로 근접하고 있다. 그림 4.2에서 TA는 $\gamma = 0.2$와 $\gamma = 0.4$일 때 오류합의 차이가 매우 크게 나는 것을 볼 수 있다. $\gamma = 0.2$일 경우 TA의 오류합은 평균차이가 4인 지점에서 MPT의 오류합에 비해 커지지만 두 방법에서의 오류합 차이는 크지 않다. 하지만 $\gamma = 0.4$일 경우 TA의 오류합은 TR의 오류합에 근접하고 MP에 비해 항상 작은 것을 알 수 있다. 한편 평균차이가 클수록 TA, MPT, TR의 최적분류점에서 오류합 차이는 작아지지만, γ과 관계없이 TR의 최적분류점에서의 오류합이 항상 작게 나타난다.

3.3. 감마분포

감마분포(gamma distribution)는 스코어 분포가 양의 왜도를 가질 때 정규분포의 합리적인 대안으로 활용될 수 있다. 조건부 확률밀도함수가 감마분포로 가정하자.

$$F(x;k,\lambda) = \frac{1}{\lambda^k \Gamma(k)} \int_0^x t^{k-1} \exp\left(\frac{-t}{\lambda}\right) dt, \qquad f(x;k,\lambda) = \frac{x^{k-1}}{\lambda^k \Gamma(k)} \exp\left(\frac{-x}{\lambda}\right),$$

여기서 k와 λ는 각각 형태모수와 척도모수이며 $x,\ k,\ \lambda \geq 0$이다. 부도차주와 정상차주의 분포를 각각 GAM($k_d,\ \lambda_d$)와 GAM($k_n,\ \lambda_n$)로 두고 전체정확도와 진실율을 최대로하는 분류점은 다음과 같다.

보조정리 3.7. 감마분포에서 가설검정을 이용한 최적분류점

유의수준 α에서의 강력검정과 일반화가능도비검정을 이용한 분류점은 다음과 같다.

$$x_o = \frac{\lambda^{k_d}}{\Gamma(k_d)} \int_0^{1-\alpha} t^{k_d-1} \exp\left(\frac{-\lambda_d}{t}\right) dt.$$

보조정리 3.7은 로지스틱분포의 역함수가 $F^{-1}(u) = \lambda^k/\Gamma(k) \int_0^u t^{-k-1} \exp(-\lambda/t)dt$이므로 쉽게 증명된다.

보조정리 3.8. 감마분포에서 전체정확도를 최대로하는 최적분류점

전체정확도가 최대인 분류점은 다음과 같다.

$$\text{i)}\ \ \lambda_n = \lambda_d = \lambda, \qquad x_o = \lambda \times \left(\frac{\Gamma(k_d)}{\Gamma(k_n)} \times \frac{1-\gamma}{\gamma}\right)^{\frac{1}{k_d - k_n}},$$

$$\text{ii)}\ \ k_n = k_d = k, \qquad x_o = \left(k \times \ln\left(\frac{\lambda_d}{\lambda_n}\right) + \ln\left(\frac{1-\gamma}{\gamma}\right)\right)\left(\frac{1}{\lambda_n} - \frac{1}{\lambda_d}\right)^{-1}.$$

식 (2.6)을 이용하여

$$\frac{f(x_o; k_d, \lambda_d)}{f(x_o; k_n, \lambda_n)} = \frac{\lambda_n^{k_n}\Gamma(k_n)}{\lambda_d^{k_d}\Gamma(k_d)} x_o^{k_d - k_n} \exp\left(\frac{x_o}{\lambda_n} - \frac{x_o}{\lambda_d}\right) = \frac{1-\gamma}{\gamma} \tag{3.1}$$

이 된다. 여기서 척도모수 또는 형태모수를 동일하다고 가정하면, 최적분류점은 보조정리 3.8과 같이 얻는다.

보조정리 3.9. 감마분포에서 진실율을 최대로하는 최적분류점

진실율이 최대인 분류점은 다음과 같다.

$$\text{i)}\ \ \lambda_n = \lambda_d = \lambda, \qquad x_o = \lambda \times \left(\frac{\Gamma(k_d)}{\Gamma(k_n)}\right)^{\frac{1}{k_d - k_n}},$$

$$\text{ii)}\ \ k_n = k_d = k, \qquad x_o = k \times \ln\left(\frac{\lambda_d}{\lambda_n}\right)\left(\frac{1}{\lambda_n} - \frac{1}{\lambda_d}\right)^{-1}.$$

보조정리 3.9는 식 (3.1)에서의 우변을 1로 설정하면 쉽게 구할 수 있다. 감마분포에서 척도모수와 형태모수 둘 중 하나의 모수를 고정한 후 최적분류점을 얻고 있다. 이에 최적분류점에서의 오류합 비교도이를 기준으로 작성하였다. 오류율 γ는 앞에서 제시한 분포들과의 오류합 비교가 가능하도록 0.2, 0.4, 0.5인 경우에 대해서 각각 표와 그림으로 제시하였다.

$\lambda = 0.3$일 때 최적분류점에서의 오류합 비교는 표 3.3에 제시하였고 $\lambda = 0.2$와 $\lambda = 0.4$에서의 오류합 비교는 그림 3.3과 3.4에 제시하였다. $\lambda = 0.3$일 때 형태모수 k가 커질수록 평균과 분산이 함께 커지지만 오류합은 세 방법에서 구한 최적분류점 모두에서 작아지는 것으로 나타났다. 형태모수 k의 크기와 상관없이 오류합의 크기는 항상 TR < TA < MPT 순으로 나타났다. 즉 진실율이 가장 낮은 오류합을 보이고 강력검정을 이용한 분류점에서 가장 큰 오류합을 보였다.

표 **3.3.** 감마분포에서 척도모수(λ) 변화와 오류합($\gamma = 0.3$)

α /β /오류합	Gam(1, 2) *vs.*			Gam(2, 1) *vs.*		
	Gam(2, 2)	Gam(3, 2)	Gam(4, 2)	Gam(2, 2)	Gam(2, 3)	Gam(2, 4)
MPT	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
	0.8002	0.5759	0.3518	0.6854	0.4690	0.3323
	0.8502	0.6259	0.4018	0.7354	0.5190	0.3823
TA	0.6514	0.3962	0.2541	0.7071	0.3993	0.2738
	0.0694	0.0672	0.0504	0.1023	0.1472	0.1358
	0.7208	0.4634	0.3045	0.8094	0.5465	0.4096
TR	0.3679	0.2431	0.1625	0.2358	0.1591	0.1165
	0.2642	0.1699	0.1115	0.4034	0.3005	0.2364
	0.6321	0.4130	0.2740	0.6392	0.4596	0.3529

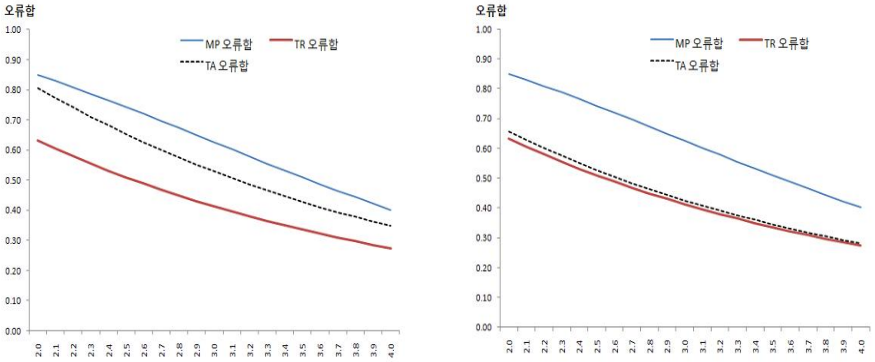

그림 **3.3.** 감마분포에서 형태모수(k)변화와 오류합($\gamma = 0.2, 0.4$)

표 3.3의 오른쪽 부분과 그림 3.4는 형태모수가 동일하다고 가정($k_n = k_d = k$)한 후 세가지 방법에 의해 얻은 최적분류점에 대응하는 오류합을 제시하고 있다. 귀무가설은 형태모수 k를 2로 하고 척도모수 λ을 1로 가정한다. 그리고 대립가설은 척도모수인 λ를 2에서 4까지 변화시키면서 오류합을 비교하였다. λ가 커질수록 평균과 분산이 함께 커지지만 세가지 방법에서 구한 최적분류점에 대응하는 오류합은 λ가 커질수록 작아졌다. $\gamma = 0.3$의 경우 λ의 크기 차이와 관계없이 오류합의 크기는 TR $<$ MPT $<$ TA 순으로 나타났다. 하지만 γ가 커질수록 TA의 오류합은 점차 작아져서 $\gamma = 0.4$일 경우 세가지 방법에서 구한 최적분류점에서의 오류합은 TR $<$ MPT $<$ TA 순으로 나타났다. 즉 스코어 분포를 감마분포로 가정하고 오류합을 최소하는 최적분류점은 진실율이 항상 가장 낮은 것으로 나타나고, γ에 의존하는 TA는 γ가 0.5에 가까울수록 진실율의 오류합에 근접하고 강력검정보다 작은 오류합을 갖는다.

4. 결론

본 연구는 스코어 분포가 와이블분포, 로지스틱분포, 감마분포일 때 홍종선 등 (2010)이 제안한 세가지 방법을 이용하여 각 분포에서의 최적분류점을 제시하였고, 이 분류점에 대응하는 제 I 종 오류와 제 II 종 오류합의 크기를 비교하고 토론하였다.

와이블분포에서는 형태모수를 $b = 1$로 고정한 후 귀무가설을 $\lambda = 1$로 두고 대립가설을 λ가 2부터 5까지 변할 때 세가지 방법으로 구한 최적분류점에서의 오류합을 비교하였다. 전체부도율이 $\gamma = 0.3$일 때

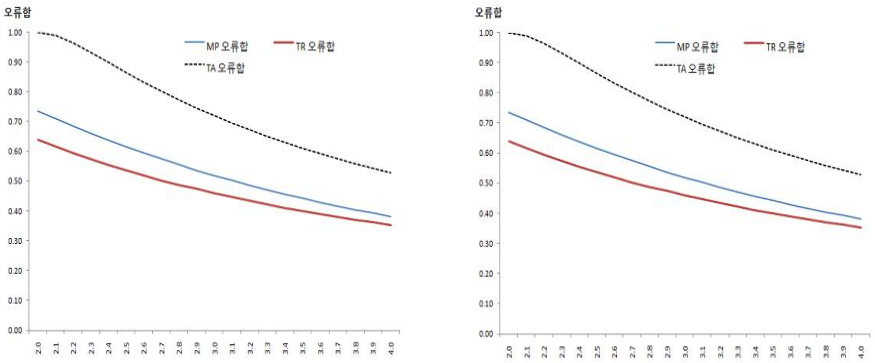

그림 3.4. 감마분포에서 척도모수(λ) 변화와 오류합 ($\gamma = 0.2, 0.4$)

척도모수 λ가 증가할수록 오류합은 감소하였지만 λ의 증가는 평균과 분산을 동시에 증가시킴에 따라, 오류합의 크기는 크게 감소하지 않았다. 하지만 오류합의 크기는 세가지 방법 중 TA에 대응하는 오류합이 가장 크고, TR을 이용하여 구한 오류합이 항상 가장 작게 나타났다. 로지스틱분포에서는 분산을 고정하고 평균을 2부터 5까지 변할 때 세가지 방법으로 구한 최적분류점에서의 오류합을 비교하였다. $\gamma = 0.3$일 때 세가지 방법으로 구한 최적분류점에서의 오류합은 두 분포의 평균차이가 작을 때는 TA에서 구한 최적분류점에서의 오류합이 MPT에서 보다 작지만, 두 분포의 평균차이가 클 때는 TA보다 MPT에서 구한 최적분류점에서의 오류합이 더 작게 나타난다. 하지만 로지스틱분포에서 세가지 방법 중 TR은 항상 작게 나타났다. 마지막 감마분포에서 척도모수와 형태모수 둘 중 하나의 모수를 고정한 후 구한 최적분류점에서 세가지 방법으로 비교하였다. $\gamma = 0.3$일 때 척도모수를 고정한 경우 오류합의 크기는 TR $<$ TA $<$ MPT 순으로 크게 나타났고 형태모수를 고정할 경우 오류합의 크기는 TR $<$ MPT $<$ TA 순으로 나타났다. 한편 γ에 의존하는 TA는 $\gamma\,(< 0.5)$가 커질수록 각 분포의 최적분류점에서의 오류합은 TR의 오류합에 근접하였다. 이에 세가지 방법에서 구한 최적분류점에 대응하는 오류합 크기는 전체부도율(γ)과 모수 변화에 따라 약간의 순서 변화는 보지만, TR을 최대로하는 최적분류점이 세가지 분포에서 항상 오류합을 가장 작게 하는 기준으로 나타났다.

와이블분포, 로지스틱분포, 감마분포는 신용평가에서 스코어 분포가 높은 왜도를 가지거나, 부도와 정상 스코어 분포의 형태가 다른 경우 많이 활용되는 분포이다. 본 연구는 이들 분포에서 최적분류점을 제시하고 모수의 다양한 변화에서 제 I 종 오류와 제 II 종 오류합을 가장 작게 하는 분류정확도는 진실율임을 밝혔다. 실제 신용평가에서의 스코어 분포가 정규성을 갖지 못하는 경우가 많음을 고려하면, 본 연구에서 제시한 세 분포에서의 최적분류점과 다양한 조건에서 오류합을 최소로 하는 기준은 신용평가모형에서 유용하게 활용될 것이다.

참고문헌

홍종선, 주재선, 최진수 (2010). 혼합분포에서의 최적분류점, <응용통계연구>, **23**, 13–28.

홍종선, 최진수 (2009). ROC와 CAP 곡선에서의 최적분류점, <응용통계연구>, **22**, 911–921.

Bairagi, R. and Suchindran, C. M. (1989). An estimator of the cutoff point maximizing sum of sensitivity and specificity, *The Indian Journal of Statistics*, **51**, 263–269.

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*.

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

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.

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

Sobehart, J. R. and Keenan, S. C. (2001). Measuring default accurately, Credit Risk Special Report, *Risk*, **14**, March, 31–33.

Tasche, D. (2006). Validation of internal rating systems and PD estimates, on-line bibliography available from: http://arxiv.org/abs/physics/0606071.

Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Willianms, 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, *Metodoloki zvezki,* **3**, 89–108.

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

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 Non-Normal Mixture

Chong Sun Hong[1] · Jae Seon Joo[2]

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

(Received June 2010; accepted August 2010)

Abstract

From a mixture distribution of the score random variable for credit evaluation, there are many methods of estimating optimal thresholds. Most the research news is based on the assumption of normal distributions. In this paper, we extend non-normal distributions such as Weibull, Logistic and Gamma distributions to estimate an optimal threshold by using a hypotheses test method and other methods maximizing the total accuracy and the true rate. The type I and II errors are obtained and compared with their sums. Finally we discuss their efficiency and derive conclusions for non-normal distributions.

Keywords: Credit, default, discriminatory, evaluation, optimal threshold, total accuracy, 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