

AROC 곡선과 최적분류점

홍중선¹ · 이희정²

¹성균관대학교 통계학과, ²성균관대학교 응용통계연구소

(2010년 8월 접수, 2010년 10월 채택)

요약

혼합분포를 가정한 신용평가 연구에서 ROC 곡선은 부도와 정상 차주의 판별력을 탐색하는데 유용한 그림이다. ROC 곡선을 개선하여 스코어를 파악할 수 있는 AROC 곡선을 수리적으로 분석하고, 정규분포를 적용하여 다양한 곡선의 형태를 파악한다. 최적분류점을 발견하는 다양한 분류정확도 통계량과 AROC 곡선의 관계를 발견하고, 두 분포의 분산이 동일한 경우에 AROC 곡선의 극소점으로 최적의 분류점을 추정할 수 있음을 발견한다.

주요용어: 부도, 분류정확도, 스코어, 신용, 오즈비, 판별력.

1. 서론

차주(borrower)는 확률변수 X 와 모수공간 Θ 에 의해서 특성을 나타낸다고 가정하자. 확률변수 X 는 대출기관에서 차주의 신용가치를 예상하기 위해 차주에게 부여한 연속형 값을 갖는 스코어(score)이다. 그리고 모수공간 Θ 는 차주가 대출상환기간 내에 대출금을 납부할 수 있는지 혹은 체납할 것인지를 나타내는 상태(state)이다. 이 상태 모수는 부도(default; θ_d) 혹은 정상(non-default; θ_n) 상태의 원소로 구성된다. 스코어 변수 X 를 통한 대출기관의 의도는 궁극적으로 차주의 신용가치에 관한 정보에 의거하여 차주의 미래상태 Θ 를 예상하는 것이다. 추가적으로 차주의 모집단은 두 개의 부모집단으로 구성된다고 가정한다. 부모집단은 미래시점에 대출상환능력이 없는 부도 상태와 대출상환능력이 있는 정상 상태로 구분된다. 다시 말해서, 차주의 모수 Θ 가 θ_d 일 때($\Theta = \theta_d$) 부도차주의 모집단에 속하고, 차주의 Θ 가 θ_n 일 때($\Theta = \theta_n$) 정상차주의 모집단에 속한다. 그리고 주어진 모수공간 Θ 에서 스코어 변수의 조건부 누적분포함수를 각각 $F_d(\cdot)$ 와 $F_n(\cdot)$ 으로 나타낼 때, 스코어 X 의 분포는 다음과 같이 표현할 수 있다.

$$F(x) = \gamma F_d(x) + (1 - \gamma) F_n(x), \quad (1.1)$$

여기서 γ 는 전체부도율(total probability of default)이다. 즉 $\gamma = P[\Theta = \theta_d]$. 이러한 가정 하에 차주의 모집단에서 미래시점의 부도와 정상을 얼마나 잘 판별할 수 있는지를 신용평가모형(credit evaluation model)을 통해서 평가하고자 한다.

신용평가모형의 타당성을 검증하는 대표적인 방법으로 CAP과 ROC 곡선이 있다. 자세한 성질은 Sobehart 등 (2000), Hanley와 McNeil (1982), Sobehart와 Keenan (2001), Vuk과 Curk (2006) 등의 문헌을 참조. 이차원 그래프로 표현되는 CAP과 ROC 곡선은 각 절단점(cut-off value)의 스코어에

¹교신저자: (110-745) 서울시 종로구 명륜동 3-53, 성균관대학교 경제학부 통계학전공, 교수.

E-mail: cshong@skku.ac.kr

서 얻는 비율들로 구성되어 있는데 각 곡선의 점을 통해서 스코어를 알 수 없다는 문제점이 있다. 홍중선 등 (2009)은 ROC 함수를 응용하여 식 (2.1)의 AROC(Adjusted ROC) 함수를 제안하고 이를 수직축에 나타내며 이에 대응하는 스코어를 수평축에 나타내어 AROC 곡선을 제안하였다. 그리고 많은 자료를 생성해서 AROC 곡선을 구현하였고 이를 통해 최적분류점(optimal threshold)을 추정하였다. 그러나 생성된 자료가 다양하지 않은 이유로 특히 정상과 부도 상태의 표본에서 평균차이가 크거나 두 상태의 표본이 중복되는 경우가 적은 경우가 대부분이라 유도된 결론이 일반적이지 않다는 것을 발견하였다. 따라서 본 연구에서는 홍중선 등 (2009)이 제안한 AROC 함수를 수리적으로 접근하여 AROC 함수의 성격을 파악하고 다양한 정상과 부도 상태의 확률밀도함수를 적용하여 AROC 곡선을 구현하고자 한다. 또한 여러 종류의 분류정확도(classification accuracy) 통계량을 설명하고 이를 최대화하는 최적분류점과 AROC함수와의 관계를 수리적으로 증명하고자 한다.

본 연구의 구성은 다음과 같다. 2절에서는 AROC 함수를 이론적으로 접근하고 이 함수의 성격을 연구한다. 그리고 확률밀도함수가 정규분포인 경우에 다양한 평균과 분산의 조합으로 AROC 곡선을 구현하여 AROC 함수의 형태를 3절에서 토론한다. 4절에서는 여러 종류의 분류정확도 통계량 중에서 많이 사용하는 전체정확도(total accuracy)와 SSS(sum of sensitivity and specificity), 유덴지수(Youden index), (0, 1)에 최단거리(Closest-to-(0, 1)) 통계량과 홍중선과 최진수 (2009) 그리고 홍중선 등 (2010)이 제안한 진실율(true rate)을 소개하고 이 통계량들의 관계를 간략히 설명하면서 이 분류정확도를 최대로 할 때의 최적분류점과 AROC 함수와의 관계를 증명한다. 마지막 5절에서 결론을 맺는다.

2. AROC 곡선

ROC 곡선의 좌표는 $(F_n(x), F_d(x))$ 또는 $(u, ROC(u))$, 여기서 $ROC(u) = F_d(F_n^{-1}(u))$ 으로 표현하는데, 홍중선 등 (2009)이 제안한 AROC 곡선의 수평축은 스코어를, 수직축은 $F_d(\cdot)$ 와 $F_n(\cdot)$ 의 오즈비(odds ratio)로 표현하며 이를 다음과 같이 AROC 함수로 나타낸다.

$$AROC(x) = \frac{F_d(x)(1 - F_n(x))}{F_n(x)(1 - F_d(x))}. \quad (2.1)$$

AROC 함수의 성격을 파악하기 위해서 우선 AROC 함수를 일차 미분하고, 이를 0으로 놓으면 다음과 같다. 즉 $dAROC(x)/dx = 0$ 은 다음을 유도한다.

$$(1 - F_n(x))f_d(x)F_n(x) = (1 - F_d(x))f_n(x)F_d(x). \quad (2.2)$$

AROC함수의 성격을 정리 2.1에서 요약한다.

정리 2.1 AROC 함수가 극소값(LOCAL MINIMUM VALUE)을 갖는 $X = x_0$ 에서 다음을 만족한다.

$$\frac{f_d(x_0)}{f_n(x_0)} = \frac{F_d(x_0)(1 - F_d(x_0))}{F_n(x_0)(1 - F_n(x_0))}.$$

이 때 $f_d(x_0) = f_n(x_0)$ 이면, 다음과 같은 조건이 된다.

$$\frac{F_d(x_0)}{1 - F_n(x_0)} = \frac{F_n(x_0)}{1 - F_d(x_0)}. \quad (2.3)$$

증명은 식 (2.2)에서 쉽게 유도되므로 생략한다.

식 (2.3)의 의미를 그림 2.1에서 살펴보고자 한다. 평균은 다르지만 분산이 동일한 두 확률밀도함수 $f_d(x)$ 와 $f_n(x)$ 를 나타내고, 이에 대응하는 누적분포함수 $F_d(x)$ 와 $F_n(x)$ 그리고 $1 - F_d(x)$ 와

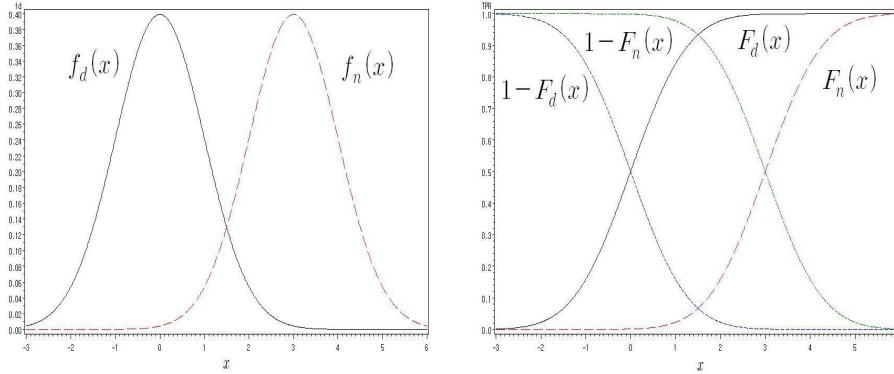


그림 2.1. 확률밀도함수와 누적분포함수

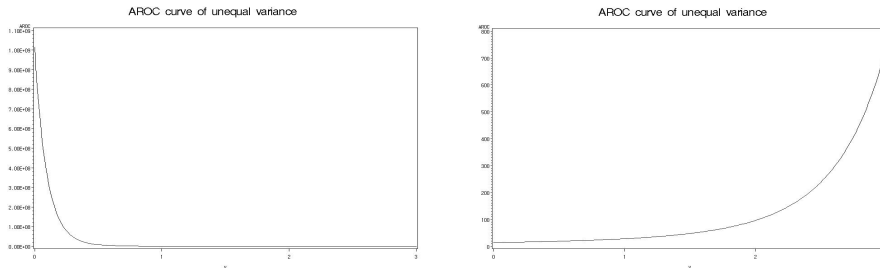


그림 3.1. AROC 곡선(좌: $\mu_n=3, \sigma_n^2=0.5$, 우: $\mu_n=3, \sigma_n^2=2$)

$1 - F_n(x)$ 를 겹쳐 표현하였다. 두 확률밀도함수가 교차되어 같은 값을 갖는 $X = x_0$ 에서 $F_d(x)$ 와 $1 - F_n(x)$ 이 교차되어 같은 값을 갖고 또한 $F_n(x)$ 과 $1 - F_d(x)$ 이 같은 값을 가져 식 (2.3)이 성립되는 것을 확인할 수 있다. 따라서 두 확률밀도함수가 분산이 동일하다면, 식 (2.3)을 만족하는 스코어 또는 절단점에서 AROC 함수가 극소값을 갖는다는 결론을 유도한다.

만약 두 확률밀도함수가 분산이 동일하지 않다면, $f_d(x_0) = f_n(x_0)$ 을 만족하는 특정한 $X = x_0$ 에서 식 (2.3)이 성립하지 않는다. 따라서 이분산인 경우에는 AROC 함수가 극소값을 갖지 않음을 파악할 수 있다.

다음 절에서는 확률밀도함수가 다양한 정규분포인 경우에 AROC 곡선을 표현하여 함수 형태와 특징을 살펴본다.

3. AROC 함수의 형태

부도 상태의 확률밀도함수 $f_d(x)$ 를 표준정규분포 $\phi(x|0, 1)$ 그리고 정상 상태의 확률밀도함수 $f_n(x)$ 를 평균과 분산이 μ_n, σ_n^2 인 정규분포 $\phi(x|\mu_n, \sigma_n^2)$ 로 설정하자. 우선 정상 상태의 분산이 다른 경우 ($\sigma_n^2 \neq 1$) AROC 곡선의 형태에 대해 살펴보자. 부도분포의 분산이 정상분포의 분산보다 큰 경우, 즉 $\sigma_n^2 < 1$ 을 고려하자. 그림 3.1의 왼쪽과 같이 AROC 함수는 감소함수이며 극소값이 존재하지 않는다. 따라서 이 경우에는 식 (2.3)의 조건이 성립하지 않기 때문에 단조감소함수 형태를 갖는 AROC 함수이다. 부도분포의 분산이 정상분포의 분산보다 작은 경우($\sigma_n^2 > 1$)의 그림 3.1의 오른쪽을 살펴보자. 여기에서는 AROC 함수가 증가함수이다. 따라서 식 (2.3)의 조건이 성립하지 않기 때문에 극소값을 갖지 않는 단조증가함수 형태의 AROC 함수이다. 그리고 분산이 다른 경우에 대한 식 (2.2)에 대하여 수리

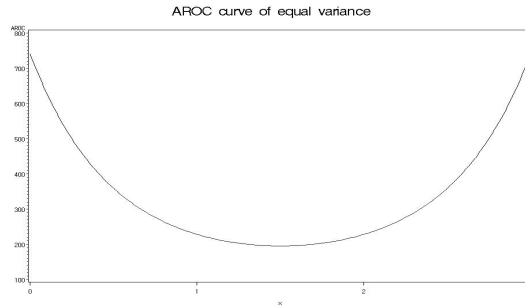


그림 3.2. AROC 곡선($\mu_n=3 \sigma_n^2=1$)

적 분석은 더 이상 불가능하다.

다음으로 정상 상태의 분산이 부도 상태의 분산과 동일한 경우 ($\sigma_n^2 = 1$)인 그림 3.2를 살펴보자. 다양한 정상분포의 평균 μ_n 과 전체부도율에서 AROC 함수는 감소하다 증가하는 형태를 나타내고 극소값을 갖는 스코어는 두 분포의 평균의 가운데(mid point)임을 확인할 수 있다. 이 경우는 그림 2.1에서의 확률밀도함수와 누적분포함수를 통해 식 (2.3)의 관계가 성립한다. 그리고 σ_n^2 이 0.9 또는 1.2와 같은 경우에도 유사한 결과를 유도할 수 있다.

그러므로 정상 상태의 분산이 부도 상태보다 작은(큰) 경우에는 감소(증가)하는 AROC 함수를 나타내며, 정상과 부도 상태의 분산이 동일하거나 유사한 경우에는 AROC 함수가 극소값을 갖는다는 현상을 발견하였다.

4. AROC 곡선과 최적분류점

분류정확도를 측정하는 다양한 통계량이 존재한다. 많이 사용하는 전체정확도(total accuracy; TA), $TA(x) = \gamma F_d(x) + (1 - \gamma)(1 - F_n(x))$, Bairagi와 Suchindran (1989)의 SSS(the sum of sensitivity and specificity) 통계량, $SSS(x) = \text{Max}(F_d(x) + 1 - F_n(x))$, Youden (1950)이 제안한 유덴지수(Youden index; J) $J(x) = \text{Max}(F_d(x) - F_n(x))$ 와 Perkins와 Schisterman (2006)의 (0, 1)에 최단거리 통계량(the closest-to-(0, 1) criterion; CC), $CC(x) = \text{Min}(1 - F_d(x) + F_n(x))$, 그리고 홍종선과 최진수 (2009)와 홍종선 등 (2010)이 제안한 진실율(true rate; TR), $TR(x) = (F_d(x) + 1 - F_n(x))/2$ 등이 있다. 본 연구에서는 전체정확도 TA보다도 진실율 TR은 SSS, J 그리고 CC 통계량과 다음과 같은 선형 관계를 갖고 있으므로 여러 분류정확도 통계량 중 진실율을 사용하여 AROC 곡선과의 관계를 살펴보고자 한다.

$$\begin{aligned} SSS(x) &= \max(2 \text{TR}(x)), \\ J(x) &= \max(2 \text{TR}(x) - 1), \\ CC(x) &= 2 - \max(2 \text{TR}(x)). \end{aligned}$$

진실율 함수를 AROC 함수와 같이 비교하기 위해서 그림 4.1에 함께 나타내었다. 그리고 부도와 정상 상태의 확률밀도함수 $f_d(x)$, $f_n(x)$ 와 두 함수의 혼합분포함수 $f(x)$ 도 그림 4.1의 아래에 나타내었으며 비교를 위해 수평축인 스코어의 값을 동일하게 설정하였다.

홍종선 등 (2010)은 식 (1.1)에서의 $f_d(x)$, $f_n(x)$ 그리고 $f(x)$ 의 확률밀도함수가 동일한 값을 갖는 스코어에서 진실율을 최대로 하는 최적분류점이라고 제안하였는데, 본 연구에서는 진실율이 최대일 때의 스

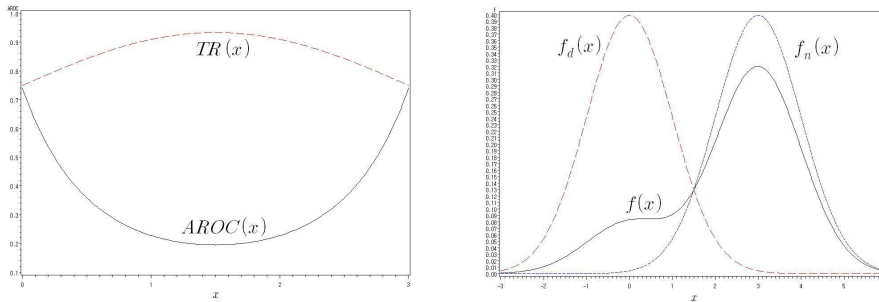


그림 4.1. 진실율과 확률밀도함수

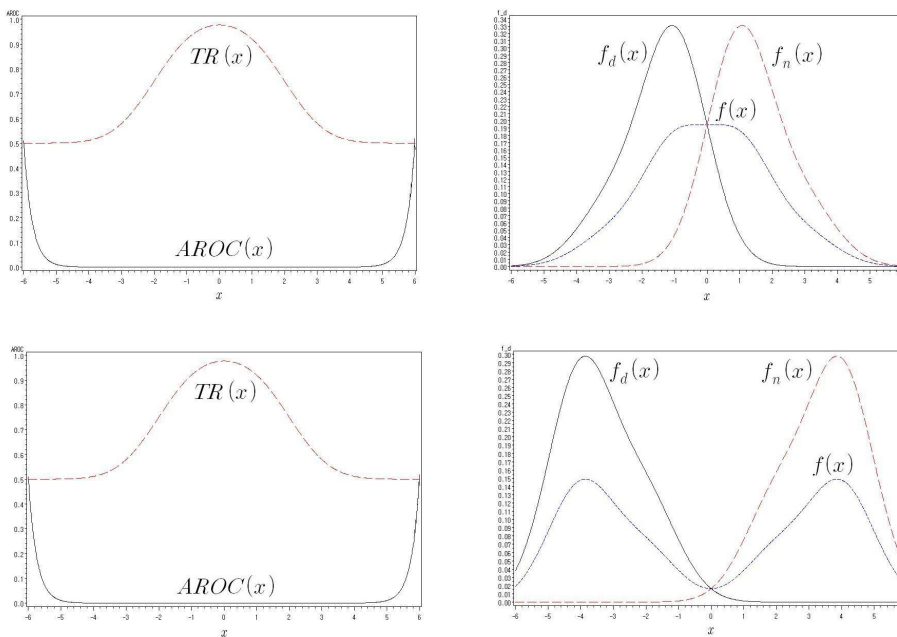


그림 4.2. 진실율과 확률밀도함수들

코어에서 AROC 함수가 극소값을 갖는다고 파악하였다. 이에 대하여 정리 4.1에서 이론적으로 설명한다. 그러므로 AROC 함수가 극소값을 갖는 경우에 이에 대응하는 스코어에서 TR, SSS 그리고 J 가 최대이고 CC가 최소인 최적분류점을 추정한다.

정리 4.1 AROC 함수가 이차함수 형태인 경우에, 진실율, SSS 통계량, 유덴지수가 최대이고 (0,1)에 최단거리가 최소인 최적분류점에서 AROC 함수가 극소값을 갖는다.

증명: 확률밀도함수가 $f_d(x_0) = f_n(x_0) = f(x_0)$ 을 만족하는 특정한 $X = x_0$ 에서 진실율(TR)이 최대(SSS, J 통계량이 최대이고 CC 통계량이 최소)인데, 이때 $X = x_0$ 에서 $F_d(x_0) = 1 - F_n(x_0)$ 그리고 $F_n(x_0) = 1 - F_d(x_0)$ 을 만족한다. 그러므로 정리 2.1에서의 AROC 함수의 조건식 (2.3)을 만족한다.□

그러므로 AROC 함수가 극소값을 갖는 경우에 이에 대응하는 스코어에서 TR, SSS 그리고 J 가 최대이고 CC가 최소인 최적분류점을 추정한다.

AROC 함수가 감소하다가 증가해서 극소값을 갖는 경우는 두 확률밀도함수 $f_d(x)$, $f_n(x)$ 의 분산이 동일하다고 언급하였는데, 본 연구의 2절에서는 분산이 동일하면서 대칭적인 확률밀도함수를 고려하였다. 여기에서는 분산은 동일하지만 대칭이 아니며 왼쪽 또는 오른쪽으로 치우친(skewed) 확률밀도함수인 경우에 AROC 함수와 진실율과의 관계를 살펴보자. 부도와 정상 상태의 확률밀도함수를 정규분포의 혼합분포로 각각 $f_d(x) = 0.2 \phi(x|-3, 1) + 0.8 \phi(x|-1, 1)$ 와 $f_n(x) = 0.8 \phi(x|1, 1) + 0.2 \phi(x|3, 1)$ 인 치우친 분포로 설정하고 전체부도를 $\gamma = 0.5$ 인 경우에 AROC 곡선과 진실율 확률밀도함수를 그림 4.2의 왼쪽 두개의 그래프로 표현하였다. 또한 다른 형태의 혼합분포인 부도와 정상 상태의 확률밀도함수를 각각 $f_d(x) = 0.7 \phi(x|-4, 1) + 0.3 \phi(x|-2, 1)$ 와 $f_n(x) = 0.7 \phi(x|2, 1) + 0.3 \phi(x|4, 1)$ 이며 전체부도를 $\gamma = 0.5$ 인 경우를 그림 4.2의 오른쪽에 구현하였다.

5. 결론

ROC 곡선에서는 절단점인 스코어가 주어졌을 때 ROC 함수를 구하고 표현하지만 ROC 곡선으로부터 스코어를 알 수 없으나, 본 연구에서 제안한 AROC 곡선의 수평축은 스코어를 그리고 수직축은 오즈비로 나타낸 AROC 함수로 표현하여 스코어와 일대일 대응 관계를 갖는다.

스코어의 혼합 확률밀도함수를 구성하는 부도와 정상 상태의 확률밀도함수가 동일한 분산을 가질 경우에는 AROC 함수가 극소값을 가지며, 이에 대응하는 스코어는 진실율, SSS 통계량, 유덴지수가 최대이고 (0, 1)에 최단거리가 최소인 최적분류점이라는 사실을 발견하였다. 그리고 AROC 함수가 단조 형태를 나타낼 때에는 각각의 확률밀도함수의 분산이 다른 경우이며, 부도 상태의 분산이 정상 상태보다 클 경우는 AROC 함수가 단조감소 함수이고, 부도 상태의 분산이 정상 상태보다 작을 경우는 AROC 함수가 단조증가 함수라는 것을 파악할 수 있으나 최적분류점을 추정할 수 없다는 단점을 갖고 있다.

참고문헌

- 홍종선, 김지훈, 최진수 (2009). 조정된 ROC와 CAP 곡선, <응용통계연구>, **22**, 29-39.
- 홍종선, 주재선, 최진수 (2010). 혼합분포에서 최적분류점, <응용통계연구>, **23**, 13-28.
- 홍종선, 최진수 (2009). ROC와 CAP 곡선에서의 최적 분류점, <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.
- Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic(ROC) curve, *Diagnostic Radiology*, **143**, 29-36.
- Perkins, N. J. and Schisterman, E. F. (2006). The inconsistency of Optimal cutpoints obtained using two criteria and based on the receiver operating characteristic curve, *American Journal of Epidemiology*, **163**, 670-675.
- 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*, *Moody's Investors Service*.
- Vuk. M. and Curk, T. (2006). Roc curve, lift chart and calibration plot, *Metodološki Zvezki*, **3**, 89-108.
- Youden, W. J. (1950). An index for rating diagnostic tests, *Cancer*, **3**, 32-35.

AROC Curve and Optimal Threshold

Chong-Sun Hong¹ · Hee-Jung Lee²

¹Department of Statistics, Sungkyunkwan University

²Research Institute of Applied Statistics, Sungkyunkwan University

(Received August 2010; accepted October 2010)

Abstract

In the credit evaluation study with the assumption of mixture distributions, the ROC curve is a useful method to explore the discriminatory power of default and non-default borrowers. The AROC curve is an adjusted ROC curve that can be identified with the corresponding score and is mathematically analyzed in this work. We obtain patterns of this curve by applying normal distributions. Moreover, the relationship between the AROC curve and many classification accuracy statistics are explored to find the optimal threshold. In the case of equivalent variances of two distributions, we obtain that the local minimum of the AROC curve is estimated at the optimal threshold to maximize certain classification accuracies.

Keywords: Classification accuracy, credit, default, discriminatory power, odds ratio, score.

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