응용통계연구 (2009)
22(5), 911-921

ROC와 CAP 곡선에서의 최적 분류점

홍종선[1] · 최진수[2]

[1]성균관대학교 통계학전공, [2]성균관대학교 응용통계연구소

(2009년 3월 접수, 2009년 6월 채택)

요 약

신용평가 연구에서 부도와 정상차주에 대한 판별력을 평가하는 방법으로 Receiver Operating Characteristic(ROC)와 Cumulative Accuracy Profile(CAP) 곡선을 사용한다. ROC 곡선에서 최적의 분류정확도를 갖는 분류점과 CAP 곡선에서 최대의 이익을 나타내는 분류점은 일반적인 정확도의 개념으로 정의된 동일한 성과를 가진 접선을 사용하여 구한다. 본 연구에서는 정확도의 대안적인 측도로 진실율을 제안하고, 이 진실율을 이용하여 ROC와 CAP 곡선에서 대안적인 최적의 분류점을 구한다. 대부분 실제 차주의 모집단에서 부도차주는 정상차주보다 훨씬 수가 적다. 이러한 경우에 진실율은 정확도보다 비용함수의 측면에서 더욱 효율적일 수 있다. 진실율을 이용하여 최적의 분류정확도를 나타내는 분류점과 최대의 이익을 의미하는 분류점에 대응하는 스코어는 동일하다는 것을 보였으며, 이 스코어는 부도와 정상 차주의 분포함수의 동일성을 검정하는 Kolmogorov-Smirnov 통계량에 대응하는 스코어와도 일치하는 것을 발견하였다.

주요용어: 정확도, 신용평가, 절단점, 부도, 판별력, 스코어.

1. 서론

차주(borrower)는 두 종류의 확률변수 S와 Z에 의해서 특성을 나타난다고 가정하자. 첫번째 변수 S는 대출기관에서 차주의 신용가치를 예상하기 위해 차주에게 부여한 연속형 값을 갖는 스코어이다. 그리고 두번째 변수 Z는 차주가 대출상환기간내에 대출금을 납부할 수 있는지 혹은 체납할 것인지를 평가하는 변수를 나타낸다. 이 평가변수는 부도(default; D) 혹은 정상(non-default; N)으로 나타난다. 스코어 변수 S를 통한 대출기관의 의도는 궁극적으로 차주의 신용가치에 관한 정보에 의거하여 차주의 미래상태 Z를 예상하는 것이다. 따라서 이러한 관점에서 스코어변수와 평가변수는 이항분류와 관련이 있다.

추가적으로 차주의 모집단은 두 개의 부모집단으로 구성되어 있다고 가정한다. 부모집단은 미래시점에 대출상환능력이 없는 부도와 대출상환능력이 있는 정상으로 구분된다. 다시 말해서, 차주의 Z가 D값을 갖을때($Z = D$) 부도차주의 모집단에 속하고, 차주의 Z가 N값을 갖을때($Z = N$) 정상차주의 모집단에 속한다. 그리고 두 개의 Z값이 주어진 스코어 변수 S의 조건부 분포를 각각 $F_D(\cdot)$와 $F_N(\cdot)$으로 나타낼 때, 스코어 S의 분포는 다음과 같이 표현할 수 있다.

$$F(s) = \alpha F_D(s) + (1 - \alpha)F_N(s),$$

여기서 α는 부도율총합(total probability of default)이다. 즉 $\alpha = P(Z = D)$.

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

이러한 가정 하에 차주의 모집단에서 미래시점의 부도와 정상을 얼마나 잘 판별할 수 있는지를 신용평가모형을 통해서 평가하고자 한다. 신용평가모형의 타당성을 검정하는 대표적인 방법으로 Cumulative Accuracy Profile(CAP)이 가장 많이 사용된다. 두번째로는 CAP과 유사한 Receiver Operation Characteristic(ROC)이 있다. ROC와 CAP 곡선의 성질을 연구한 연구문헌들은 Sobehart 등 (2000), Hanley와 McNeil (1982), Sobehart와 Keenan (2001), Vuk과 Curk (2006) 등에서 찾아볼 수 있다.

ROC와 CAP 곡선은 성과(performance)를 기반으로 한 분류모형(classification model) 또는 분류자(classifiers)를 시각화할 수 있고, 조직화하여 향상시킬 수 있으며, 그리고 평가할 수 있는 유용한 방법이다. ROC 곡선은 분류자의 '*hit rate*'(이익)와 '*false alarm rate*'(비용) 사이에 교환(trde-off)을 나타내는 신호탐지이론에서 오랫동안 사용되어졌다 (Egan, 1975; Centor, 1991; Swets 등, 2000). 또한 의사결정과 의학진단의 체계에서 폭넓게 사용되어졌다 (Hanley와 McNeil, 1982; Swets, 1988; Zou, 2002). ROC 곡선의 특성에 관한 설명과 실제 연구에서 ROC 분석을 응용하는데 관련된 정보는 Fawcett (2003)과 Provost와 Fawcett (1997, 2001)에서 발견할 수 있다. CAP 곡선 또는 Lift chart는 마케팅과 판매에 관한 데이터 마이닝에서 잘 알려진 도구이다 (Berry과 Linoff, 1999). 그러나 CAP 곡선은 사용빈도에 비하여 많은 연구가 되어있지 않은 상태라고 Vuk과 Curk (2006)는 언급하였다.

본 연구의 초점은 ROC와 CAP 곡선의 통계적 성질을 연구하고 최적의 분류정확도(optimal classification accuracy) 또는 최대의 이익(maximal profit)을 나타내는 최적 분류점 또는 절단점(threshold, cut-off)을 발견하는데 목적이 있다. Vuk과 Curk (2006)은 *accuracy*(정확도)의 개념을 이용하여 최적 분류점을 발견하는 방법을 개발하였다. 그러나 정상차주보다 부도차주의 수가 적을 때 *accuracy*라는 측도는 적절하지 않을 수 있다. 본 연구에서는 최적의 분류정확도 또는 최대의 이익을 나타내는 분류점을 발견하기 위하여 *accuracy* 측도의 대안적인 *true rate*(진실율)를 제안한다.

2절에서는 ROC와 CAP 곡선에 대해 설명하면서, 분류자의 개념과 이항 분류자와 확률 분류자로 구분하여 소개한다. 3절에서는 *accuracy*의 정의를 설명하고 *accuracy*를 이용하여 동일한 성과를 나타내는 선형식을 얻은 후에, ROC와 CAP 곡선과 이 선형식의 접점을 통해 최적의 분류점을 구하는 과정을 설명한다. 그리고 *accuracy*의 대안으로 *true rate*라는 측도를 제안하고, *true rate*를 기반으로 동일한 성과를 갖는 두 종류의 접선을 생성하고 이 접선을 이용하여 최적의 분류정확도를 가진 대안적인 분류점을 제시한다. 4절에서는 기존의 *accuracy*를 이용하여 구한 최적의 분류점과 *true rate*를 이용하여 구한 최적의 분류점을 모의실험과 실증예제를 통하여 비교하고, 이 분류점을 비용함수와 K-S 통계량에 대응하는 스코어와 함께 토론한다. 5절에서는 *true rate*에 근거한 최적의 분류정확도의 분류점에 대해 결론을 유도한다.

2. ROC와 CAP 곡선

ROC 와 CAP 함수식은 다음과 같이 정리된다.

$$\mathrm{ROC}(u) = F_D\left(F_N^{-1}(u)\right), \quad u \in (0,1),$$
$$\mathrm{CAP}(u) = F_D\left(F^{-1}(u)\right), \quad u \in (0,1).$$

ROC와 CAP 곡선은 다음의 식으로 각각 표현되거나

$$(u, \mathrm{ROC}(u)), \qquad (u, \mathrm{CAP}(u)), \quad u \in (0,1),$$

또는 다음의 점들로 각각 이루어진다 (상세한 정보는 Tasche (2006) 참조).

$$(F_N(s), F_D(s)), \qquad (F(s), F_D(s)), \quad s \in (-\infty, \infty).$$

표 **2.1.** 혼동행렬

	Actual default	Actual non-default
Bad Predicted default	TP True Positive	FP False Positive
Good Predicted non-default	FN False Negative	TN True Negative
합	P	N

고정된 스코어 s보다 작거나 같은 스코어를 갖는 모든 차주들이 부도로 간주될 때, $F_N(s)$는 분류점으로 스코어 s를 가질 때 부도로 잘못 예측된 정상차주의 모집단의 비율로서 정의하고, "*false alarm rate*" 또는 "*false positive rate*"라 한다. 반대로 $F_D(s)$는 "*hit rate*" 또는 "*true positive rate*"라고 부르고 이러한 절차로 정확하게 평가된 부도차주의 비율을 나타낸다. "*alarm rate*"라고 부르는 $F(s)$는 분류점 s에서 전체 모집단에서 부도로 예측되는 비율을 나타낸다.

부도 예측모형의 성과를 평가하는 가장 기본적인 접근 방법은 예측된 부도(또는 정상)차주 수를 고려하고 이를 실제로 발생한 부도(또는 정상)차주 수와 비교한다. 이를 표현하는 기본적인 방법은 표 2.1과 같이 분할표 또는 혼동행렬(confusion matrix) 형태로 표현된다.

가장 간단한 경우로는 bad와 good의 두 종류로 평가한다. 이러한 평가는 부도 또는 정상차주의 실제 결과와 비교된다. 이 상황에서 정확하게 분류된 부도와 정상차주의 수를 각각 TP와 TN이라고 정의한다. 반대로 FP는 부도로 예측되었는데 정상으로 평가된 수이며, FN은 실제 부도인데 정상으로 예측된 수이다. 모형의 오류들로 행렬에서 비대각의 항인 FN은 제 I종 오류로 나타나며 FP는 제 II종 오류를 나타낸다. 그리고 P는 모집단에서 전체 부도차주의 수이고 N은 정상차주의 수이며, 사전에 알고있다고 가정한다 (Stein, 2005).

*false positive rate*인 $F_N(s)$와 *true positive rate*인 $F_D(s)$는 각각 FP/N과 TP/P로 주어지고, *alarm rate*인 $F(s)$는 $(\mathrm{TP}+\mathrm{FP})/(P+N)$으로 나타난다. FP_{rate}, TP_{rate} 그리고 Alarm_{rate}는 다음과 같이 요약된다.

$$\begin{aligned}
\mathrm{FP}_{rate} &= \hat{F_N}(s) = \frac{\mathrm{FP}}{N}, \\
\mathrm{TP}_{rate} &= \hat{F_D}(s) = \frac{\mathrm{TP}}{P}, \\
\mathrm{Alarm}_{rate} &= \hat{F}(s) = \hat{\alpha}\hat{F_D}(s) + (1-\hat{\alpha})\hat{F_N}(s) \\
&= \left(\frac{P}{P+N}\right)\mathrm{TP}_{rate} + \left(\frac{N}{P+N}\right)\mathrm{FP}_{rate} \\
&= \frac{\mathrm{TP}+\mathrm{FP}}{P+N}.
\end{aligned}$$

그러므로 $(F_N(s),\, F_D(s))$와 $(F(s),\, F_D(s))$, $s \in (-\infty, \infty)$의 점으로 구성되는 ROC와 CAP 곡선은 각각 추정값인 $(\mathrm{FP}_{rate}, \mathrm{TP}_{rate})$와 $(\mathrm{Alarm}_{rate}, \mathrm{TP}_{rate})$으로 구현된다. 다시 말해 ROC 곡선의 추정값은

$$\left(\hat{F_N}(s), \hat{F_D}(s)\right) = (\mathrm{FP}_{rate}, \mathrm{TP}_{rate})$$

으로 이루어지며, CAP 곡선은 다음과 같은 추정값으로 구성된다.

$$\left(\hat{F}(s), \hat{F_D}(s)\right) = (\mathrm{Alarm}_{rate}, \mathrm{TP}_{rate}).$$

$F_N(s)$, $F_D(s)$ 그리고 $F(s)$는 확률 분류자이고 FP_{rate}, TP_{rate} 그리고 Alarm_{rate}는 이항 분류자이다. 또한 모집단에서 부도차주의 수인 P를 파악하지 못하는 경우에는 TP_{rate}를 계산할 수 없으며, 따라서 ROC 곡선을 사용할 수 없다. 그러므로 이러한 경우에는 CAP 곡선이 유용하게 사용될 수 있다.

3. ROC와 CAP 곡선의 최적의 분류점

ROC와 CAP 곡선의 모든 점은 분류정확도와 다른 질적인 측도로 계산되는 이항 분류자에 대응된다. 분류정확도를 계산하기 위해 우선 $P:N$의 비율을 알아야 한다. 우선 *accuracy*(정확도)는 다음과 같이 정의한다.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{P + N}. \tag{3.1}$$

*accuracy*는 다음과 같이 표현할 수 있다.

$$\begin{aligned}
\text{Accuracy} &= \left(\frac{P}{P+N}\right)\text{TP}_{rate} + \left(\frac{N}{P+N}\right)\text{TN}_{rate} \\
&= \left(\frac{P}{P+N}\right)\text{TP}_{rate} + \left(\frac{N}{P+N}\right)(1 - \text{FP}_{rate}) \\
&= \hat{\alpha}\hat{F_D}(s) + (1-\hat{\alpha})\left(1 - \hat{F_N}(s)\right) \qquad (3.2) \\
&= 2\hat{\alpha}\hat{F_D}(s) - \hat{F}(s) + (1-\hat{\alpha}). \qquad (3.3)
\end{aligned}$$

Vuk과 Curk (2006)은 식 (3.2)와 (3.3)을 사용하여 Fawcett (2003)가 제안한 동일한 분류정확도를 가지는 두 개의 선형식(동일한 성과를 나타내는 선)을 얻었다.

$$\hat{F_D}(s) = \frac{1-\hat{\alpha}}{\hat{\alpha}}\left(\hat{F_N}(s)\right) + \frac{1}{\hat{\alpha}}\left[\text{Accuracy} - (1-\hat{\alpha})\right], \tag{3.4}$$

$$\hat{F_D}(s) = \frac{1}{2\hat{\alpha}}\left[\hat{F}(s) + \text{Accuracy} - (1-\hat{\alpha})\right]. \tag{3.5}$$

또한 Vuk과 Curk (2006)은 최적의 분류정확도를 나타내는 분류점은 ROC 곡선과 동일한 성과 직선인 식 (3.4)와의 접점이며 그리고 최대의 이익을 나타내는 점은 CAP 곡선과 동일한 성과 직선인 식 (3.5)와의 접점인 것을 보였다. 그리고 ROC와 CAP 곡선에서 접점이외의 다른 점들은 일반적으로 낮은 분류정확도와 이익을 갖는다는 것으로 파악된다.

본 연구에서는 2절에서 정의된 *true positive rate*와 *true negative rate*의 합으로 표현되는 *true rate*(진실율)를 제안한다.

$$\text{True}_{rate} = \frac{\text{TP}}{P} + \frac{\text{TN}}{N}. \tag{3.6}$$

이 *true rate*는 식 (3.1)에서 정의된 *accuracy*의 대안적 측도로서 분류모형의 정확도를 측정하지만, 다음 4절에서 보여주듯이 $P:N$의 비율이 너무 작거나 또는 너무 클 때 *accuracy*의 정의보다 더 효율적일 수 있다. 실제로 차주의 모집단 대부분에서는 부도차주의 수(P)가 정상차주의 수(N)보다 적은 상황임을 상기한다.

식 (3.6)의 *true rate*는 다음과 같이 $F_N(s)$와 $F_D(s)$로 나타낼 수 있다.

$$\begin{aligned}
\text{True}_{rate} &= \text{TP}_{rate} + (1 - \text{FP}_{rate}) \\
&= \hat{F_D}(s) - \hat{F_N}(s) + 1.
\end{aligned}$$

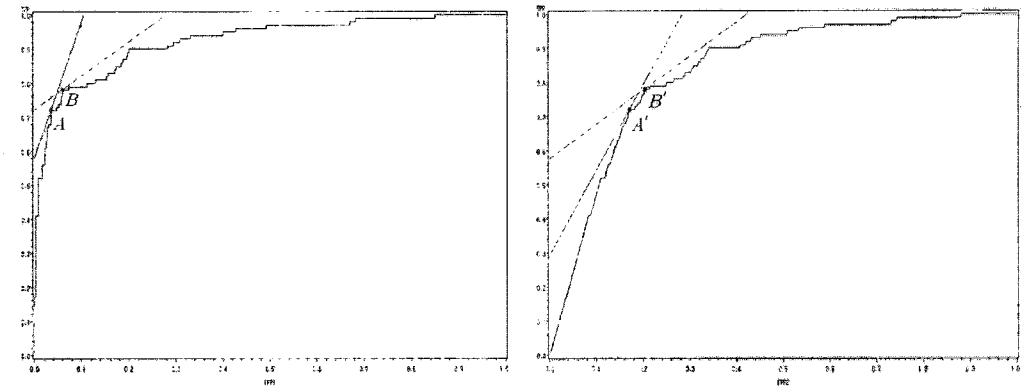

그림 **4.1.** ROC와 CAP 곡선

따라서 동일한 *true rate*를 갖는 다음과 같은 선형식을 얻는다.

$$\hat{F_D}(s) = \hat{F_N}(s) + (\text{True}_{rate} - 1). \tag{3.7}$$

그리고 식 (3.6)은 $F(s)$와 $F_D(s)$로 다음과 같이 표현되므로

$$\text{True}_{rate} = \frac{1}{1-\hat{\alpha}}\left[\hat{F_D}(s) - \hat{F}(s)\right] + 1,$$

동일한 *true rate*로 표현되는 두번째 선형식을 다음과 같이 얻는다.

$$\hat{F_D}(s) = \hat{F}(s) + (1-\hat{\alpha})(\text{True}_{rate} - 1). \tag{3.8}$$

식 (3.4)와 (3.5)와 유사하게, 식 (3.7)과 ROC 곡선의 접점은 최적의 분류정확도의 분류점을 결정하는 대안적인 방법을 제공하고, 식 (3.8)과 CAP 곡선의 접점은 최대의 이익을 나타내는 대안적인 분류점을 제공한다. 식 (3.4)와 (3.5) 선형식의 기울기는 동일하지 않으며 $\hat{\alpha} < 0.5$일 때 1보다 큰 값을 갖지만 식 (3.7)과 (3.8)의 기울기는 동일한 값을 가진다. 따라서 대안적인 분류점은 기울기가 1인 식 (3.7)과 (3.8)의 접선으로 결정되므로 ROC와 CAP 곡선에서 가장 북서쪽에 위치한 점이 최적의 분류점이 된다.

4. 모의실험과 실증예제

4.1. 모의실험

평균이 0이고 분산이 1인 정규분포로부터 표본크기가 100인 확률표본을 생성하여 부도차주의 역할을 부여한다. 그리고 평균이 2이고 분산이 1인 정규분포로부터 400개의 확률표본을 생성하여 정상차주의 역할을 부여한다. 두 표본을 합하여 표본크기가 500인 하나의 표본으로 만든다 ($P : N = 100 : 400$이므로 $\hat{\alpha} = 1/5$). 확률적 분류자의 분류점을 변화시키면서 이항적인 분류의 집합을 얻을 수 있다. 이항 분류의 집합을 사용함으로써 그림 4.1에서와 같이 ROC와 CAP 곡선을 작성한다.

그림 4.1의 왼쪽 그림에 나타난 ROC 곡선의 A점은 식 (3.4)에서 기울기가 4 ($= (1-\hat{\alpha})/\hat{\alpha}$)인 직선과의 접점이다; 이 접점 A의 좌표는 (0.0350, 0.7200)이고, 이 점에 대응하는 최적의 절단점 또는 최적의 분류점의 스코어는 0.47695이다. 대안적인 최적의 분류점 B는 식 (3.7)에서 기울기가 1인 직선과의 접점이다; B의 좌표는 (0.0600, 0.7800)이고, 이때 대응되는 스코어는 0.75742이다.

표 4.1. 혼동 행렬

	Case I		Case II	
	Actual default	Actual non-default	Actual default	Actual non-default
Bad	72	14	78	24
Good	28	386	22	376

표 4.2. Accuracy와 True rate의 비교

	Case I		Case II
Accuracy	(72+386)/500 = 0.9160	>	(78+376)/500 = 0.9080
True rate	72/100 + 386/500 = 1.6850	<	78/100 + 376/400 = 1.7200

표 4.3. 총비용함수

	$c_1 : c_2$			
	2:1	3:1	4:1	5:1
Case I	70	98	126	154
Case II	68	90	112	134

그림 4.1의 오른쪽 그림은 CAP 곡선으로써 A'과 B'이라는 유사한 형태로 나타나는데, CAP 곡선의 A'은 식 (3.5)에서 기울기가 2.5 (= $1/2\hat{\alpha}$)인 직선과의 접점이다; 좌표는 (0.1720, 0.7200)이고 대응되는 스코어는 그림 4.1의 ROC 곡선의 A점의 스코어와 동일하다. B'은 기울기가 1인 식 (3.8)과의 접점이다; 좌표는 (0.2040, 0.7800)이고 대응되는 스코어는 ROC 곡선의 B점의 스코어와 동일하다.

ROC와 CAP 곡선에서 최적의 분류점을 발견하기 위하여 *accuracy*를 이용하는 방법을 Case I이라 하고, *true rate*를 사용하는 방법을 Case II라고 하자. Case I과 Case II에서 각각 구한 최적의 분류점에 대응하는 혼동행렬은 표 4.1에 나타내었다. 이 혼동행렬을 통해서 Case I과 Case II의 *accuracy*와 *true rate*를 표 4.2에 구하였다. Case I에서 정확하게 분류된 TP와 TN의 합은 $72 + 386 = 458$로서 Case II인 $78 + 376 = 454$보다 크므로 Case I의 *accuracy*는 Case II보다 높은 값을 갖는다. 그러나 $P : N$의 비율 때문에 Case I의 *true rate*는 1.6850으로 Case I인 1.7200보다 낮은 것을 볼 수 있다.

제I종 오류인 FN과 제II종 오류인 FP에 대응하는 비용 c_1과 c_2를 가진 다음과 같은 비용함수를 고려하자.

$$\text{Cost} = c_1 \times \text{FN} + c_2 \times \text{FP}. \tag{4.1}$$

신용평가연구에서 실제 부도를 정상으로 예측하는 제I종 오류에 대한 비용(손실)은 부도로 예측되었으나 정상으로 평가되는 제II종 오류에 대한 비용보다 크기 때문에 현실적인 비용함수를 위하여 c_1과 c_2의 비용비율을 1:1이 아닌 2:1 이상으로 가정하자. Case I과 Case II의 총비용은 여러 종류의 c_1과 c_2의 비율을 사용하여 표 4.3에 요약하였다. 표 4.3으로부터 c_1과 c_2의 비율이 증가할수록 Case I에 대한 비용이 Case II에 대한 비용보다 점점 높게 증가한다는 것을 파악할 수 있다. 그러므로 c_1과 c_2의 비용비율이 높은 경우의 비용함수를 고려하면, *accuracy*를 이용하여 구한 최적의 분류점보다 *true rate*를 이용한 최적의 분류점이 더 선호된다고 주장할 수 있다.

다음의 가설을 검정하는 Kolmogorov-Smirnov(K-S) 통계량을 구하기 위하여 부도차주의 분포함수

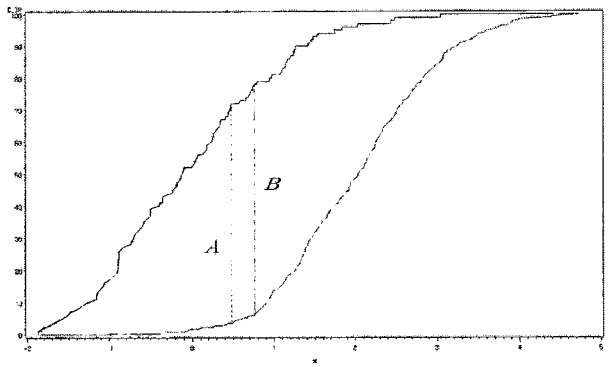

그림 **4.2.** Kolmogorov-Smirnov 통계량

$\hat{F_D}(s)$와 정상차주의 분포함수 $\hat{F_N}(s)$를 구해 보자.

$$H_0 : F_D(s) = F_N(s) \quad vs. \quad H_1 : F_D(s) > F_N(s).$$

$\hat{F_D}(s) = \mathrm{TP}_{rate}$ 그리고 $\hat{F_N}(s) = \mathrm{FP}_{rate}$이므로 두 분포함수는 그림 4.2에 구현하였다. 놀랍게도 K-S 통계량에 대응되는 스코어는 0.75742로 ROC와 CAP 곡선에서 *true rate*를 이용하여 구한 최적의 분류점 B(또는 B')에 대응하는 스코어와 일치한다는 것을 확인할 수 있다. K-S 통계량은 Case II에 해당하는 B(또는 B')에 대응하는 스코어 0.75742에서 실선으로 나타나고, Case I의 A(또는 A')에 대응되는 스코어는 0.47695에서 점선으로 나타난다.

표본크기가 500인 예제 4.1과 같은 확률표본을 여러번 발생시키고 *accuracy*와 *true rate*에 대응하는 두 종류의 최적의 분류점을 구하여 비교한 모의실험에서 표 4.2와 4.3과 같이 동일한 결과를 얻었으며, *true rate*에 대응하는 최적의 분류점은 K-S 통계량에 대응하는 스코어와 일치함을 확인하였다.

그러므로 현실적인 비용함수와 K-S 통계량에 기반하는 관점에서 살펴볼 때, ROC와 CAP 곡선에서 *accuracy*를 이용한 최적의 분류점보다 *true rate*를 이용한 최적의 분류점이 판별력이 좋다고 말할 수 있다.

4.2. 실증예제

1994년부터 2005년까지 외감기업 중 총 매출액이 4,500억원 이상인 1,009건의 한국 기업(58건의 부도기업과 951건의 정상기업)을 고려하자. 분류점을 변동시키면서 이항분류의 집합을 얻을 수 있으며 이를 바탕으로 ROC와 CAP 곡선을 그림 4.3에 나타내었다.

그림 4.3의 왼쪽 그림은 ROC 곡선으로 점 A는 기울기가 951/58인 식 (3.4)와의 접점이다; A의 좌표는 (0.0053, 0.1552)이고, 대응되는 스코어는 13.6819이다. 대안적인 최적의 점 B는 *true rate*를 적용하여 기울기가 1인 식 (3.7)과의 접점이다; B의 좌표는 (0.1546, 0.7586)이고 대응되는 스코어는 29.5233으로 나타난다.

그림 4.3의 오른쪽 그림은 CAP 곡선으로 점 A'은 기울기가 1009/116인 식 (3.5)와의 접점이다; 좌표는 (0.0139, 0.1552)이고 대응되는 스코어는 그림 4.3에서 ROC 곡선의 A의 스코어와 동일하다. B'은 기울기가 1인 식 (3.8)과의 접점이다; 좌표는 (0.1893, 0.7586)이며 대응되는 스코어는 그림 4.3에서 ROC 곡선의 B의 스코어와 동일하다.

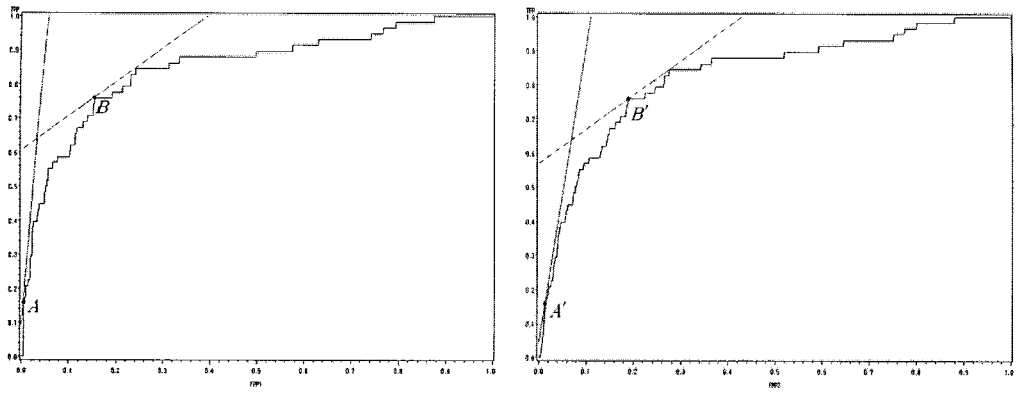

그림 **4.3.** ROC와 CAP 곡선

표 **4.4.** 혼동 행렬

	Case I		Case II	
	Actual default	Actual non-default	Actual default	Actual non-default
Bad	9	5	44	147
Good	49	946	14	804

표 **4.5.** Accuracy와 True rate의 비교

	Case I		Case II
Accuracy	$(9+946)/1009$ $= 0.9465$	$>$	$(44+804)/1009$ $= 0.8404$
True rate	$9/58+946/951$ $= 1.1499$	$<$	$44/58+804/951$ $= 1.6041$

표 **4.6.** 총비용함수

	$c_1 : c_2$			
	2:1	3:1	4:1	5:1
Case I	103	152	201	250
Case II	175	189	203	217

이 예제에서도 4.1절처럼 *accuracy*를 이용한 Case I과 *true rate*를 이용한 Case II로 구분하여 살펴보자. Case I과 Case II에서 구한 최적의 분류점 각각에 대응하는 혼동행렬은 표 4.4에 나타내었고 Case I과 Case II에 관한 *accuracy*와 *true rate*를 계산하여 표 4.5에 나타내었다. Case I의 *accuracy*가 Case II보다 더 높은 반면에, Case I의 *true rate*는 Case II보다 더 낮은 것을 식별할 수 있다. 비록 Case I의 TP와 TN의 합이 Case II보다 매우 큼에도 불구하고 예제 4.1에서와 동일한 결과를 보여주는 것을 파악할 수 있다.

식 (4.1)에서 정의된 동일한 비용함수를 고려하자. Case I과 Case II에서의 총비용은 다양한 c_1과 c_2의 비율로 표 4.6에 표현하였다. 표 4.6으로부터 c_1과 c_2의 비율이 높지 않을 때에는 Case I에 대한 비용은 Case II 보다 낮지만, c_1과 c_2의 비율이 증가할수록 Case I에 대한 비용은 Case II를 초과한다는 것을 식별할 수 있다. 따라서 c_1과 c_2의 비용비율이 높은 경우에는 *accuracy*를 이용하는 것보다 *true rate*를 이용하여 구한 최적의 분류점이 선호된다고 판단할 수 있다.

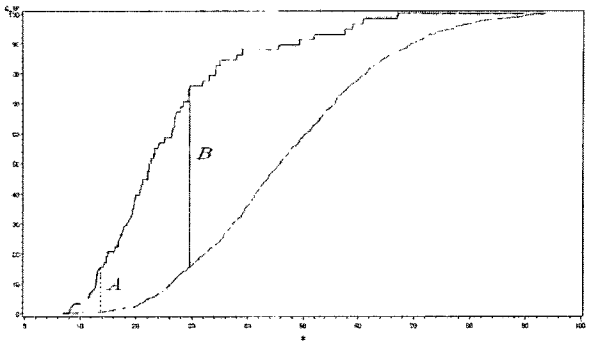

그림 **4.4.** Kolmogorov-Smirnov 통계량

부도차주와 정상차주의 분포함수인 $\hat{F}_D(s)$와 $\hat{F}_N(s)$를 구하고 그림 4.4에 구현하였다. 놀랍게도 K-S 통계량에 대응되는 스코어는 29.5233이며, 이것은 그림 4.3의 B (또는 B')의 대안적인 최적의 분류점에 대응하는 스코어와 일치한다는 것을 발견하였다. K-S 통계량은 그림 4.3에 실선으로 나타냈으며, A (또는 A')에 대응하는 스코어 13.6819에서는 점선으로 표현하였다.
그러므로 비용함수와 K-S 통계량에 근거하여 살펴볼 때, ROC와 CAP 곡선에서 *accuracy*를 이용한 최적의 분류점보다 *true rate*를 이용한 최적의 분류점이 보다 좋은 판별력을 갖는다고 결론내릴 수 있다.

5. 결론

Vuk과 Curk (2006)은 동일한 분류정확도를 갖는 점들로 이루어진 식 (3.4)와 (3.5)의 두 직선을 정의하였다. Fawcett (2003)은 이 직선을 동일한 성과를 나타내는 선(iso-performance line)이라 하였다. 그리고 Vuk과 Curk (2006)은 식 (3.4)와 ROC 곡선의 접점을 최적의 분류정확도의 분류점으로 나타내고, 식 (3.5)와 CAP 곡선의 접점을 최대의 이익을 나타내는 분류점을 나타낸다고 주장하였다. 이 최적의 분류점은 *accuracy*에 기반하고 최적의 분류점에 대응하는 스코어는 동일하였다.

본 연구에서는 *accuracy*의 대안적인 측도로 *true rate*를 제안하였다. *true rate*를 사용하여 두 종류의 동일한 성과를 나타내는 직선으로 식 (3.7)과 (3.8)을 구하였다. 식 (3.7)과 ROC 곡선의 접점은 대안적인 최적의 분류정확도를 나타내는 분류점을 의미하고, 식 (3.8)과 CAP 곡선의 접점은 대안적인 최대의 이익을 나타내는 분류점을 의미한다. 그리고 대안적인 두 종류의 최적의 분류점에 대응하는 스코어는 항상 동일하고, *true rate*를 이용한 최적의 분류점에 대응하는 스코어는 K-S 통계량에 대응되는 스코어와 일치함을 발견하였다. 두 종류의 최적 분류점 그리고 K-S 통계량에 대응하는 스코어가 모두 일치하는것을 모의실험과 예제를 통해 탐색적으로 살펴보았고 수리적인 증명은 향후 연구과제로 남겨둔다.

모의실험과 실증예제에 근거하여 현실적인 비용함수와 K-S 통계량에 대응되는 스코어를 살펴볼 때, *accuracy* 보다 *true rate*를 이용하여 구한 최적의 분류점이 판별력이 더 높다는 것을 보였다. 따라서 *true rate*를 사용하면 *accuracy* 보다 최적의 분류점을 발견하는데 효율적이라는 결론내릴 수 있다.

본 연구에서 제안한 *true rate*를 이용하여 ROC와 CAP 곡선으로부터 각각 대안적인 최적의 분류점을 얻었다; ROC 곡선에서는 동일한 성과를 나타내는 직선인 식 (3.7)과의 접점이고, CAP 곡선에서는 동일한 성과를 나타내는 직선인 식 (3.8)과의 접점이다. ROC와 CAP 곡선과의 접점은 기울기가 모두 1인 접선과 각각 만나기 때문에 곡선의 가장 북서쪽에 해당하는 점이다. 그러므로 *true rate*를 이용하면, ROC와 CAP 곡선의 가장 북서쪽에 위치한 점에 대응하는 스코어가 최적의 분류점이 된다고 판단할 수 있다.

참고문헌

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

Centor, R. M. (1991). Signal detectability: *The use of ROC curves and their analyses*, Medical Decision Making.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*, Academic Press, New York.

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 (ROC) curve, *Diagnostic Radiology*, **143**, 29–36.

Provost, F. and Fawcett, T. (1997). Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions, KDD-97.

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

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

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

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. (1988). Measuring the accuracy of diagnostic systems, *Science* **240**, 1285–1293.

Swets, J. A., Dawes, R. M. and Monahan, J. (2000). Better decisions through science, *Scientific American*, **283**, 82–87.

Tasche, D. (2006). Validation of internal rating systems and PD estimates, arXiv:physics/0606071, **1**.

Vuk, M. and Curk, T. (2006). ROC Curve, Lift Chart and Calibration Plot, *Metodolo ški zvezki*, **3**, 89–108.

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

Optimal Threshold from ROC and CAP Curves

Chong Sun Hong[1] · Jin Soo Choi[2]

[1]Department of Statistics, Sungkyunkwan University
[2]Research Institute of Applied Statistics, Sungkyunkwan University

(Received March 2009; accepted June 2009)

Abstract

Receiver Operating Characteristic(ROC) and Cumulative Accuracy Profile(CAP) curves are two methods used to assess the discriminatory power of different credit-rating approaches. The points of optimal classification accuracy on an ROC curve and of maximal profit on a CAP curve can be found by using iso-performance tangent lines, which are based on the standard notion of accuracy. In this paper, we offer an alternative accuracy measure called the true rate. Using this rate, one can obtain alternative optimal threshold points on both ROC and CAP curves. For most real populations of borrowers, the number of the defaults is much less than that of the non-defaults, and in such cases the true rate may be more efficient than the accuracy rate in terms of cost functions. Moreover, it is shown that both alternative scores of optimal classification accuracy and maximal profit are the identical, and this single score coincides with the score corresponding to Kolmogorov-Smirnov statistic used to test the homogeneous distribution functions of the defaults and non-defaults.

Keywords: Accuracy, credit rating, cut-off point, default, discriminatory power, score.

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