The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.5.773
27(5), 773–786

Alternative Optimal Threshold Criteria: MFR

Chong Sun Hong[a,1] · Hyomin Alex Kim[a] · Dong Kyu Kim[a]

[a]Department of Statistics, Sungkyunkwan University

(Received August 12, 2014; Revised September 29, 2014; Accepted September 30, 2014)

Abstract

We propose the multiplication of false rates (MFR) which is a classification accuracy criteria and an area type of rectangle from ROC curve. Optimal threshold obtained using MFR is compared with other criteria in terms of classification performance. Their optimal thresholds for various distribution functions are also found; consequently, some properties and advantages of MFR are discussed by comparing FNR and FPR corresponding to optimal thresholds. Based on general cost function, cost ratios of optimal thresholds are computed using various classification criteria. The cost ratios for cost curves are observed so that the advantages of MFR are explored. Furthermore, the definition of MFR is extended to multi-dimensional ROC analysis and the relations of classification criteria are also discussed.

Keywords: Classification performance, confusion matrix, cost ratio, default, threshold.

1. 서론

두 분포함수의 혼합분포로부터 판별력을 극대화하는 분류점(절단점; threshold, cut-off)을 추정하는 최근의 연구는 신용평가와 의학통계 분야 등에서 많이 활용되고 있다. 본 연구에서는 신용평가에서 차주의 신용가치를 기준으로 대출상환능력에 따라 부도(default)와 정상(non-default)상태를 판별하는 문제를 고려하자. 확률변수 X는 스코어 변수로 연속형 실수값이다. 모수공간은 부도와 정상상태로 가정하여 $\Theta = \{\theta_d, \theta_n\}$로 정의한다. $F_d(x)$와 $F_n(x)$를 각각 차주의 부도와 정상상태에서 스코어의 조건부 누적분포함수 $P(X \le x|\theta_d)$와 $P(X \le x|\theta_n)$로 정의하며, 스코어 확률변수 X의 누적분포함수 $F(x)$는 다음과 같이 가정한다.

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

여기서 γ는 전체부도율이다.

ROC(Receiver Operating Characteristic) 곡선은 성과(performance)를 기반으로 분류모형(classification model) 또는 분류자(classifiers)를 시각화하며 평가할 수 있는 유용한 방법이다. 2000년 이후의 ROC 곡선에 관한 연구는 Provost와 Fawcett (1997), Sobehart와 Keenan (2001), Zhou 등 (2002), Engelmann 등 (2003), Fawcett (2003), Hong (2009), Hong 등 (2010) 이외의 많은 문헌에서 발견할 수 있다. ROC 곡선은 이항분류자(binary classifier)를 사용하여 각 분류점의 스코어에서 실제부도

[1]Corresponding author: Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea.
E-mail: cshong@skku.edu

Table 1.1. Confusion matrix

Rate		State of Nature	
		e_1	e_2
Decision	d_1	$F_d(x_0)$	$F_n(x_0)$
	d_2	$1 - F_d(x_0)$	$1 - F_n(x_0)$

를 부도로 정확히 예측하는 비율 TPR(true positive rate 또는 hit rate, recall, sensitivity), $F_d(x)$와 실제정상을 부도로 잘못 예측하는 비율 FPR(false positive rate 또는 false alarm rate, 1-specificity), $F_n(x)$를 각각 수직축과 수평축 좌표에 대응시킨 그래프로 표현된다 (상세한 정보는 Pepe (2003)와 Tasche (2006) 참조). 이항 분류모형에 대하여 임의의 x_0를 기준으로 부도 또는 정상으로 의사결정한 결과는 Table 1.1과 같이 2×2 분할표 또는 혼동행렬(confusion matrix) 형태로 표현된다.

ROC 곡선에서 정의된 최적분류기준들이 많이 존재하는데 이에 대한 연구로는 Cantor 등 (1999), Greiner과 Gardner (2000), Freeman과 Moisen (2008), Liu 등 (2009), Hong (2009) 등이 있다.

다양한 분류기준들 중에서 ROC 곡선으로부터 최적분류점을 구할 수 있는 대표적 세가지 전체정확도(Total Accuracy; TA (Lambert와 Lipkovich, 2008)), 진실율(True Rate; TR (Velez 등, 2007; Hong과 Joo, 2010)), 정확도면적(Accuracy Area; AA (Brasil, 2010))를 간략히 소개하면 다음과 같다.

$$\begin{aligned}
\mathrm{TA} &= \max\left\{\gamma F_d(x) + (1-\gamma)(1-F_n(x))\right\}, \\
\mathrm{TR} &= \max\left\{\frac{1}{2}\left[F_d(x) + (1-F_n(x))\right]\right\}, \\
\mathrm{AA} &= \max\left\{F_d(x)(1-F_n(x))\right\}.
\end{aligned}$$

최적분류기준 TA와 TR은 다음과 같이 일차식으로 표현할 수 있으며, Figure 1.1의 왼쪽 그림과 같이 기울기가 서로 다른 직선과 ROC 곡선과 만나는 각각의 접점 좌표 ($F_n(x)$, $F_d(x)$)에 대응하는 x를 최적분류점으로 구한다 (Vuk과 Curk, 2006).

$$\begin{aligned}
F_d(x) &= F_n(x) + (2\mathrm{TR} - 1), \\
F_d(x) &= \frac{1-\gamma}{\gamma} F_n(x) + \frac{1}{\gamma}(\mathrm{TA} + \gamma - 1).
\end{aligned}$$

또한 최적분류기준 TR과 AA는 Figure 1.1의 오른쪽 그림과 같이 ROC 곡선으로부터 구한 다각형과 사각형의 면적을 최대로 하는 ROC 곡선의 좌표로부터 최적분류점으로 구한다 (Brasil, 2010). 본 연구에서는 ROC 곡선으로부터 TR과 AA와는 다른 형태의 면적으로 정의하는 새로운 분류기준인 오분류율곱(multiplication of false rates; MFR)를 제안한다. 앞에서 언급한 분류기준 TA, TR, AA와 본 연구에서 제안한 MFR 기준으로부터 얻는 최적분류점들을 비교하면서 분류성과와 비용곡선(cost curve)에 대하여 토론하고자 한다. 분류성과를 비교하기 위해서 다양한 분포함수에 대하여 TA, TR과 MFR 기준에 기반하는 최적분류점을 구하고 이에 대응하는 제1종 오류, 제2종 오류, 두 오류의 합을 구하면서 비교 분석하여 어떤 경우에 MFR의 오류율이 작은지를 살펴보면서 MFR의 장점을 토론한다. 그리고 TA와 TR 기준을 비용곡선을 이용하여 비용비율을 비교 분석한 Hong과 Yoo (2010)의 연구를 확장하여 본 연구에서 제안한 MFR 기준과 TA, TR에 대한 비용비율을 다양한 정규분포에서 구하고 분석한다.

본 논문의 2절에서는 MFR 기준을 제안한다. MFR의 정의를 다양한 최적분류기준들 중에서 TA, TR, AA와 비교하면서 MFR을 소개한다. 다양한 분포의 분산의 조건 하에서 TA, TR, MFR 기준을 이용하

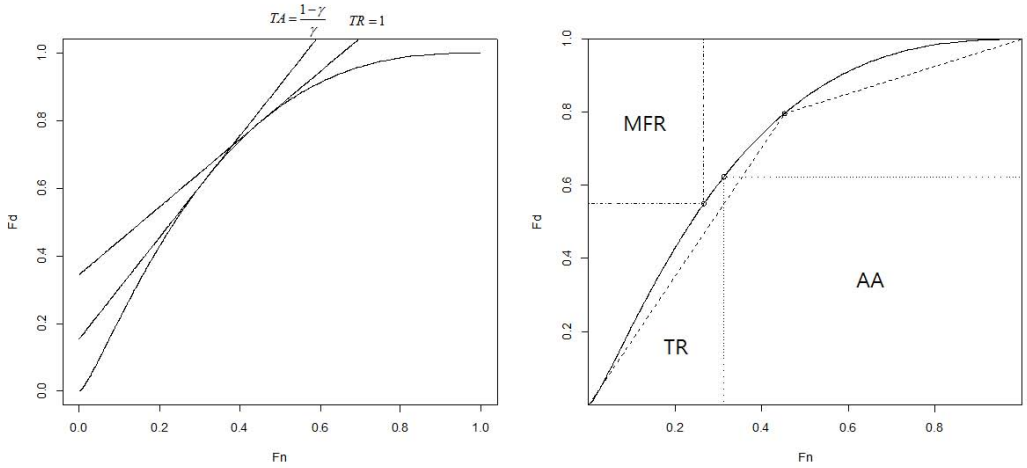

Figure 1.1. Optimal thresholds obtained from ROC curve

여 구한 분류점의 관계를 살펴본다. 그리고 정규분포에 대하여 TA, TR 그리고 MFR 기준에 대응하는 분류점을 구하고 각 분류점에 대응하는 제1종과 제2종 오류의 크기를 살펴보면서 MFR 기준의 특징을 토론한다. 3절에서는 일반적인 비용함수를 고려하면서 MFR 기준을 이용하여 구한 분류점에 대한 비용비율을 정의한다. 또한 비용곡선으로 세 기준으로부터 구한 분류점에 대한 비용비율의 관계를 정리하여 비용비율의 측면에서 MFR 기준의 장점을 유도한다. 4절에서는 이차원의 ROC 곡선을 확장한 다차원의 ROC 분석에서 MFR 기준을 정의하고 다른 기준들과의 관계를 토론하면서 다차원으로 확대 가능함을 설명하고, 마지막으로 결론은 5절에서 유도한다.

2. MFR

ROC 곡선으로부터 최적분류점을 찾은 다양한 기준들 중에 Figure 1.1의 왼쪽그림과 같이 ROC 곡선과의 접점을 이용하는 TA와 TR 기준이 있으며, Figure 1.1의 오른쪽 그림처럼 ROC곡선 아래에서 형성되는 면적을 이용하는 TR과 AA 기준이 있다. TR과 AA 기준은 ROC곡선을 이루는 평면에서 (1,0) 좌표를 포함하는 사각형의 면적형태로 표현된다. 본 연구에서는 분류 정확도를 측정하는 통계량으로 (0, 1) 좌표에서부터 ROC 곡선까지의 직선을 대각선으로 하는 직각사각형의 면적을 최대화하는 분류기준을 고려하자. 이 기준이 최대화될 때의 ROC 곡선상의 점에 대응하는 스코어를 분류점으로 설정할 수 있다. 이러한 분류기준을 ROC 곡선에서 표시하면 Figure 1.1의 오른쪽 그림과 같이 ROC 곡선의 윗부분의 사각형 면적을 나타낸다. 본 연구에서는 이러한 분류기준을 정의 2.1에 제안한다. 분류기준은 $1 - F_d(x)$와 $F_n(x)$의 곱을 최대화하는 x점을 분류점으로 설정하는데, 정의 2.1에서 $1 - F_d(x)$는 FNR(false negative rate)이고 $F_n(x)$는 FPR(false positive rate)으로 모두 오분류율(false rate)이므로 두 오분류율의 곱으로 정의한 분류 기준을 오분류율곱(multiplication of false rates; MFR)으로 정의한다.

정의 2.1 ROC 곡선에서 MFR 기준은 다음과 같이 정의한다.

$$\text{MFR} = \max\{(1 - F_d(x))F_n(x)\}. \tag{2.1}$$

부도와 정상상태의 분포함수의 평균들 사이에 존재하며 최대의 MFR을 만족하는 분류점 x_0는 조건 (2.2)를 만족한다.

$$\frac{f_d(x_0)}{f_n(x_0)} = \frac{1 - F_d(x_0)}{F_n(x_0)}, \tag{2.2}$$

여기서 $f_d(\cdot)$와 $f_n(\cdot)$는 누적분포함수 $F_d(\cdot)$와 $F_n(\cdot)$의 확률밀도함수이다. 조건 (2.2)는 식 (2.1)의 MFR을 x에 대해 미분하여 그 값을 0으로 놓으면 구할 수 있다. 참고로 TA와 TR 기준을 만족하는 분류점은 각각 $f_d(x_0)/f_n(x_0)$가 $(1-\gamma)/\gamma$과 1일 때이다 (Yoo와 Hong, 2011).

2.1. 분류점들의 관계와 오류

본 연구에서는 MFR 기준을 이용하여 얻은 분류점과 TA와 TR 기준으로 얻은 분류점의 관계를 살펴본다. 부도와 정상상태의 분포함수가 동일하다는 가정($F_d(\cdot) = F_n(\cdot)$) 하에서 TA, TR 그리고 MFR 기준에 대응하는 분류점을 구하여 비교 분석하면서 최적분류점의 변화와 제1종 오류 FNR과 제2종 오류 FPR의 크기를 살펴보면서 탐색하고자 한다.

정리 2.1 $F_d(\cdot)$와 $F_n(\cdot)$가 동일한 밀도함수이고 분산이 같은 경우 MFR과 TR 기준을 사용하여 구한 최적분류점은 동일하고, $F_d(\cdot)$의 분산이 $F_n(\cdot)$보다 작은 경우에는 MFR의 최적분류점은 TR보다 작으나, $F_d(\cdot)$의 분산이 $F_n(\cdot)$보다 큰 경우에는 MFR의 최적분류점은 TR보다 큰 값을 갖는다.

증명: σ_d^2와 σ_n^2은 분포함수 $F_d(\cdot)$와 $F_n(\cdot)$의 분산이라고 하자.

(경우 A) $\sigma_d^2 = \sigma_n^2 \equiv \sigma^2$인 경우에는 최적분류점이 $x_0 = (\mu_d + \mu_n)/2$이다 (Hong 등, 2010). 그리고 $1-F_d(x_0) = 1-\Phi((x_0-\mu_d)/\sigma) = \Phi((\mu_d-\mu_n)/2\sigma)$와 $F_n(x_0) = \Phi((x_0-\mu_n)/\sigma) = \Phi((\mu_d-\mu_n)/2\sigma)$는 동일하다. 따라서 최적분류점 x_0에 대응하는 조건은 $(1 - F_d(x_0))/F_n(x_0) = 1$이다.

(경우 B) $\sigma_d^2 < \sigma_n^2$인 경우에 $1 - F_d(x^*) = F_n(x^*)$를 만족하는 $x^* \in (\mu_d,\ (\mu_d + \mu_n)/2)$에 대하여 최적분류점 x_0는 구간 $(\mu_d,\ x^*)$사이에 존재하며 $(1 - F_d(x_0))/F_n(x_0) > 1$을 만족한다.

(경우 C) $\sigma_d^2 > \sigma_n^2$인 경우에 $1 - F_d(x^*) = F_n(x^*)$를 만족하는 $x^* \in ((\mu_d + \mu_n)/2,\ \mu_n)$에 대하여 최적분류점 x_0는 구간 $(x^*,\ \mu_d)$사이에 존재하며 $(1 - F_d(x_0))/F_n(x_0) < 1$을 만족한다. □

정리 2.1에서는 MFR과 TR의 최적분류점에 대하여만 설명하였는데 TA 기준을 이용해서 구한 분류점은 TR 기준으로부터 구한 분류점보다 항상 작은 값으로 나타난다 (Hong과 Yoo, 2010). TA의 최적분류점과의 관계와 각 분류점에 대응하는 제1종과 제2종 오류와의 관계에 대하여는 다음과 같이 설명할 수 있다.

경우 A에서 최적분류점은 $x_{\mathrm{TA}} < x_{\mathrm{MFR}} = x_{\mathrm{TR}}$순으로 나타나며, 제1종 오류 α는 $1 - F_d(x_{\mathrm{MFR}}) = 1 - F_d(x_{\mathrm{TR}}) < 1 - F_d(x_{\mathrm{TA}})$이고, 제2종 오류 β는 $F_n(x_{\mathrm{TA}}) < F_n(x_{\mathrm{MFR}}) = F_n(x_{\mathrm{TR}})$로 나타난다. 그러므로 경우 A에서는 TR과 MFR 기준으로 구한 분류점에 대응하는 제1종 오류가 TA 기준으로 구한 분류점의 제1종 오류보다 작으며, MFR 기준에 대응하는 제2종 오류가 TA 기준의 제2종 오류보다 크다는 것을 파악할 수 있다.

경우 B에서 최적분류점은 γ의 크기에 따라 $x_{\mathrm{TA}} < x_{\mathrm{MFR}} < x_{\mathrm{TR}}$ 또는 $x_{\mathrm{MFR}} < x_{\mathrm{TA}} < x_{\mathrm{TR}}$로 나타나므로 경우 B1과 경우 B2로 각각 구분하여 설명한다. 경우 B1 $x_{\mathrm{TA}} < x_{\mathrm{MFR}} < x_{\mathrm{TR}}$인 경우에 α는 $1 - F_d(x_{\mathrm{TR}}) < 1 - F_d(x_{\mathrm{MFR}}) < 1 - F_d(x_{\mathrm{TA}})$, β는 $F_n(x_{\mathrm{TA}}) < F_n(x_{\mathrm{MFR}}) < F_n(x_{\mathrm{TR}})$의 관계로 나타난다. 그러나 경우 B2 $x_{\mathrm{MFR}} < x_{\mathrm{TA}} < x_{\mathrm{TR}}$인 경우에 α는 $1 - F_d(x_{\mathrm{TR}}) < 1 - F_d(x_{\mathrm{TA}}) < 1 - F_d(x_{\mathrm{MFR}})$,

Table 2.1. Values of α, β and $\alpha+\beta$ when $\sigma_d^2=\sigma_n^2$

	α	β	$\alpha+\beta$
MFR	**0.3086**	0.3085	**0.6171**
TR	**0.3086**	0.3085	**0.6171**
TA($\gamma=0.4$)	0.4624	**0.1826**	0.6450

Table 2.2. Values of α, β and $\alpha+\beta$ when $\sigma_d^2<\sigma_n^2$ and $x_{\mathrm{TA}}<x_{\mathrm{MFR}}<x_{\mathrm{TR}}$

	α	β	$\alpha+\beta$
MFR	0.3808	0.2808	0.6616
TR	**0.2530**	0.3900	**0.6431**
TA($\gamma=0.4$)	0.4317	**0.2451**	0.6768

β는 $F_n(x_{\mathrm{MFR}})<F_n(x_{\mathrm{TA}})<F_n(x_{\mathrm{TR}})$의 관계로 나타난다. 그러므로 경우 B1에서는 MFR 기준에 대응하는 제1종과 2종 오류가 TA와 TR 기준으로 구한 분류점의 오류들 사이에 존재하며, 경우 B2에서는 MFR 기준에 대응하는 제1종 오류가 가장 크며, MFR 기준에 대응하는 제2종 오류가 가장 작다.

경우 C에서 최적분류점은 $x_{\mathrm{TA}}<x_{\mathrm{TR}}<x_{\mathrm{MFR}}$로 나타나며, α는 $1-F_d(x_{\mathrm{MFR}})<1-F_d(x_{\mathrm{TR}})<1-F_d(x_{\mathrm{TA}})$이고, β는 $F_n(x_{\mathrm{TA}})<F_n(x_{\mathrm{TR}})<F_n(x_{\mathrm{MFR}})$로 나타난다. 그러므로 경우 C에서는 MFR 기준으로 구한 분류점에 대응하는 제1종 오류가 다른 기준에 대응하는 오류보다 가장 작으며, MFR 기준에 대응하는 제2종 오류는 다른 기준에 대응하는 오류보다 가장 크다는 것을 파악할 수 있다.

2.2. 정규분포에서의 오류

$F_d(\cdot)$를 표준정규분포 $N(0,1)$로 고정하고 $F_n(\cdot)$를 평균이 1인 정규분포로 설정하고 세 가지 경우에 따라 분산을 변화시켜 TA, TR, MFR에 대한 제1종 오류 α, 제2종 오류 β, 그리고 오류의 합 $\alpha+\beta$를 구한다. 일반적으로 전체부도율 γ는 0.5보다 작으므로 0.5보다 작은 γ를 고려한다.

(경우 A) $\sigma_d^2=\sigma_n^2$

$F_n(\cdot)$를 분산이 동일하게 $N(1,\ 1)$로 설정하면, 분류점들은 $x_{\mathrm{TA}}<x_{\mathrm{MFR}}=x_{\mathrm{TR}}$의 관계를 가지고 $0.377<\gamma<0.5$일 때 TA 기준으로 구한 분류점이 적절한 구간에서 존재하므로 $\gamma=0.4$인 경우와 MFR, TR의 최적분류점에 대응하는 α, β, $\alpha+\beta$의 값은 Table 2.1과 같이 나타났다. 제1종 오류의 경우 MFR과 TR에서 0.3070으로 가장 작게 나타났으며, 제2종 오류의 경우 TA에서 0.1825로 가장 작게 나타났다. 두 오류의 합의 값은 0.6171로 MFR과 TR이 가장 작게 나타났다.

(경우 B) $\sigma_d^2<\sigma_n^2$

$F_d(\cdot)$의 분산이 $F_n(\cdot)$의 분산보다 작은 경우의 최적분류점은 γ의 크기에 따라 두 종류의 경우로 나눈다.

(경우 B1) $\sigma_d^2<\sigma_n^2$, $\gamma\in(0.371,\ 0.424)$

$F_n(\cdot)=N(1,\ 1.2^2)$, $\gamma=0.4$로 설정하면 α, β, $\alpha+\beta$의 값은 Table 2.2와 같이 나타났다. 분류점들은 $x_{\mathrm{TA}}<x_{\mathrm{MFR}}<x_{\mathrm{TR}}$의 관계를 가지므로 제1종 오류는 TR에서 0.2531로 가장 작게 나타났으며, 제2종 오류는 TA에서 0.2456으로 가장 작게 나타났다. 두 오류의 합은 0.6430으로 TR이 가장 작게 나타났다. $F_n(\cdot)=N(1,\ 1.4^2)$와 $N(1,\ 1.6^2)$을 따르는 분포에서의 실험을 더 해본 결과 $F_n(\cdot)=N(1,\ 1.2^2)$과 마찬가지로 α의 경우 TR에서 가장 작게 나타났으며, β의 경우 TA에서 가장 작게 나타났다. $\alpha+\beta$의 값은 TR이 가장 작게 나타났다.

Table 2.3. Values of α, β and $\alpha+\beta$ when $\sigma_d^2 < \sigma_n^2$ and $x_{\mathrm{MFR}} < x_{\mathrm{TA}} < x_{\mathrm{TR}}$

	α	β	$\alpha+\beta$
MFR	0.3808	**0.2807**	0.6616
TR	**0.2530**	0.3900	**0.6431**
TA($\gamma = 0.45$)	0.3327	0.3182	0.6508

Table 2.4. Values of α, β and $\alpha+\beta$ when $\sigma_d^2 > \sigma_n^2$

	α	β	$\alpha+\beta$
MFR	**0.1775**	0.4502	0.6277
TR	0.3606	0.1419	**0.5025**
TA($\gamma = 0.4$)	0.4284	**0.0860**	0.5144

(경우 B2) $\sigma_d^2 < \sigma_n^2$, $\gamma \in (0.424,\ 0.5)$

$F_n(\cdot) = N(1,\ 1.2^2)$, $\gamma = 0.45$로 설정하여 구한 α, β, $\alpha+\beta$의 값을 Table 2.3에 정리하였다. 분류점들은 $x_{\mathrm{MFR}} < x_{\mathrm{TA}} < x_{\mathrm{TR}}$의 관계를 가지므로 제1종 오류는 TR에서 0.2531로 가장 작게 나타났으며, 제2종 오류는 MFR에서 0.2810으로 가장 작게 나타났다. 두 오류의 합은 0.6430으로 TR이 가장 작게 나타났다. $F_n(\cdot) = N(1,\ 1.4^2)$와 $N(1,\ 1.6^2)$을 따르는 분포에서 추가적으로 얻은 결과도 $F_n(\cdot) = N(1,\ 1.2^2)$과 마찬가지로 α의 경우 TR에서 가장 작게 나타났으며, β의 경우 MFR에서 가장 작게 나타났다. $\alpha+\beta$의 값은 TR이 가장 작게 나타났다.

(경우 C) $\sigma_d^2 > \sigma_n^2$

$F_n(\cdot) = N(1,\ 0.6^2)$, $\gamma = 0.4$에 대한 α, β, $\alpha+\beta$의 값은 Table 2.4에 나타난다. 분류점들은 $x_{\mathrm{TA}} < x_{\mathrm{TR}} < x_{\mathrm{MFR}}$의 관계를 가지므로 제1종 오류는 MFR에서 0.1775로 가장 작게 나타났으며, 제2종 오류는 TA에서 0.0870으로 가장 작게 나타났다. 두 오류의 합은 0.5025로 TR이 가장 작게 나타났다. $F_n(\cdot) = N(1,\ 0.525^2)$와 $N(1,\ 0.8^2)$을 따르는 분포에서도 $F_n(\cdot) = N(1,\ 0.6^2)$과 마찬가지로 α의 경우 MFR에서 가장 작게 나타났으며, β의 경우 TA에서 가장 작게 나타났다. $\alpha+\beta$의 값은 TR이 가장 작게 나타났다.

제1종 오류는 FNR로서 신용평가 측면에서는 정상으로 예측했으나 실제로는 부도로 오분류하거나 또는 의학통계분야에서는 건강하다고 예측했으나 실제로는 질병에 걸린 오분류를 말한다. 그리고 부도로 예측했지만 실제로는 정상으로 나타난 경우 또는 질병에 걸렸다고 예측했으나 실제로는 건강한 상태를 제2종 오류로 간주한다. 제1종 오류 FNR은 제2종 오류인 FPR에 비해 더 심각하고 많은 금전적 손실 등이 발생한다고 볼 수 있다. 그러므로 제1종 오류를 최소화하는 것이 제2종 오류를 최소화하는 것보다 더 중요하다. 실제 상황에서는 대부분 경우 C와 같이 $F_d(\cdot)$의 분산이 $F_n(\cdot)$의 분산보다 큰 경우가 많다 (Hong과 Lee, 2011, p.273–275; Hong 등, 2011, p.989–991). 이런 상황을 모의실험한 경우 C를 살펴보면 MFR에 의한 최적분류점의 제1종 오류가 가장 작게 나타나므로 MFR 기준점이 TR이나 TA 기준으로 구한 최적분류점보다 제1종 오류를 줄이면서 심각한 오류를 최소화시킬 수 있으며 커다란 금전적 손실을 축소할 수 있으며 효율적이라고 판단할 수 있다.

3. 비용곡선에서의 비용비율

Table 3.1과 같은 비용행렬(cost matrix)이 주어졌을 때 일반적인 기대비용(expected cost; EC) 함수는 다음과 같다 (Metz, 1978; Zhou 등, 2002; Pepe, 2003; Kim, 2004).

$$\mathrm{EC} = C_0 + C_{\mathrm{TP}}\gamma F_d(x) + C_{\mathrm{FN}}\gamma(1-F_d(x)) + C_{\mathrm{FP}}(1-\gamma)F_n(x) + C_{\mathrm{TN}}(1-\gamma)(1-F_n(x)), \quad (3.1)$$

Table 3.1. Cost matrix

Rate		State of Nature	
		e_1	e_2
Decision	d_1	C_{TP}	C_{FP}
	d_2	C_{FN}	C_{TN}

여기서 C_0는 고정비용이다.

기대비용 식 (3.1)을 최소화하는 스코어를 찾는 방법은 최적분류점을 추정하는 문제로 간주하기 때문에 Jund 등 (2005), Hand (2009) 그리고 Hand와 Zhou (2009) 등은 기대비용 함수를 최소화하는 최적분류점 x_0는 $f_d(x_0)/f_n(x_0) = [(C_{\mathrm{FP}} - C_{\mathrm{TP}})/(C_{\mathrm{FN}} - C_{\mathrm{TN}})](1-\gamma)/\gamma$을 만족한다고 제안하였다. $(C_{\mathrm{FP}} - C_{\mathrm{TP}})/(C_{\mathrm{FN}} - C_{\mathrm{TN}})$를 비용비율(cost ratio; CR)로 정의하면, Hand (2009)는 기대비용 함수를 최소화할 때의 비용비율은 최적분류점 x_0에 대응하는 가능도비 $f_d(x_0)/f_n(x_0)$ 와 $\gamma/(1-\gamma)$의 곱으로 식 (3.2)를 유도하였다.

$$\mathrm{CR} = \left(\frac{f_d(x_0)}{f_n(x_0)}\right)\frac{\gamma}{1-\gamma}. \tag{3.2}$$

최근의 연구들은 최적분류점을 찾기 위해서 비용비율에 대한 결정문제로 접근하게 된다. Cantor 등 (1999)는 다양한 진단실험에서의 비용비율에 대한 추정값을 제시하였으며, Adams와 Hand (1999)는 비용비율의 최대가능도 추정값과 구간 문제를 언급하였다. 분류정확도 기준들의 결과를 그래픽적으로 살펴볼 수 있는 방법으로 performance graphs (Turney, 1995), regret graphs (Hilden과 Glasziou, 1996), ROC 곡선, loss difference plots (Adams와 Hand, 1999), precision-recall curve (Davis와 Goadrich, 2006), prevalence-value-accuracy plot (Antoni 등, 2006), DET curves (Liu와 Shriberg, 2007), skill plot (Brigg과 Zartzki, 2007) 그리고 비용곡선(cost curve) 등이 있다.

본 연구에서는 대표적인 비용곡선을 바탕으로 비용비율에 대하여 토론한다 (Drummond와 Holte, 2006; Holte와 Drummond, 2008; Hoshino 등, 2009 등의 문헌참고). Hong과 Yoo (2010)은 TA와 TR을 최대화하는 비용비율은 각각 $\mathrm{CR}_{\mathrm{TA}} = 1$과 $\mathrm{CR}_{\mathrm{TR}} = \gamma/(1-\gamma)$임을 보였고, 본 연구에서는 MFR을 최대화하는 비용비율을 정리 3.1에서 토론한다.

정리 3.1 MFR을 최대화하는 비용비율 CR은 다음과 같다.

$$\mathrm{CR}_{\mathrm{MFR}} = \left(\frac{1-F_d(x_{\mathrm{MFR}})}{F_n(x_{\mathrm{MFR}})}\right)\frac{\gamma}{1-\gamma}. \tag{3.3}$$

증명: 식 (3.2)에 MFR을 최대했을 때 분류점 조건 (2.2)를 대체시키면 구할 수 있다. □

정규화된 비용비율을 최소화하기 위하여 비용곡선을 이용하여 비용비율을 구하고자 한다. 비용곡선은 CR과 γ로 이루어진 확률비용함수(probability cost function; PCF)와 정규화된 기대비용(normalized expected cost; NEC)을 수평축과 수직축으로 나타난다 (Drummond와 Holte, 2000; Kim, 2004; Hong과 Yoo, 2010).

$$\mathrm{PCF} = \frac{(C_{\mathrm{FN}} - C_{\mathrm{TP}})\gamma}{(C_{\mathrm{FN}} - C_{\mathrm{TP}})\gamma + (C_{\mathrm{FP}} - C_{\mathrm{TN}})(1-\gamma)} = \frac{1}{1 + \mathrm{CR}(1-\gamma)/\gamma},$$
$$\mathrm{NEC} = F_n(x) + (1 - F_n(x) - F_d(x))\mathrm{PCF}.$$

비용곡선의 축을 나타내는 PCF와 NEC는 모두 0과 1사이의 값을 갖도록 표준화하였다. 비용곡선의 모든 점에 대응하는 접선을 비용직선(cost line)이라고 하는데, 실제 상황에서는 소수의 분류점에 대응하는 비용직선들의 최소덮개(minimum envelope 또는 lower envelope)를 연결하여 비용비율을 구하는데 사용한다. 만약 세 종류의 분류기준을 최대화하는 분류점의 접선인 세 비용직선 중에서 NEC가 가장 작은 선을 연결하면 비용곡선이 되며 세 종류의 분류기준 중에서 어느 구간의 PCF에서 선호되는 기준이 무엇인지 그리고 선호되는 분류기준에 따른 비용비율을 파악할 수 있다. 본 연구에서는 대표적인 분류기준인 TA와 TR 그리고 MFR로 구한 각각의 최적분류점에 대응하는 NEC와 PCF에 대한 비용비율을 살펴보자.

2절의 경우 A에서 두 분류점이 일치하므로 $x_{\mathrm{TA}} < x_{\mathrm{MFR}} = x_{\mathrm{TR}}$를 의미한다. Hong과 Yoo (2010)에서 구한 바와 같이 비용비율은 다음과 같다(경우 A: $\mathrm{NEC}_{\mathrm{MFR}} = \mathrm{NEC}_{\mathrm{TR}} < \mathrm{NEC}_{\mathrm{TA}}$).

$$\mathrm{CR}_A < \frac{F_d(x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TA}})}{F_d(x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TA}})} \frac{\gamma}{1-\gamma} = \frac{F_d(x_{\mathrm{TR}}) - F_d(x_{\mathrm{TA}})}{F_d(x_{\mathrm{TR}}) - F_d(x_{\mathrm{TA}})} \frac{\gamma}{1-\gamma}.$$

다음으로 $\sigma_d^2 < \sigma_n^2$일 때는 분류점 $x_{\mathrm{MFR}} < x_{\mathrm{TR}}$(이 때는 $x_{\mathrm{TA}} < x_{\mathrm{MFR}} < x_{\mathrm{TR}}$와 $x_{\mathrm{MFR}} < x_{\mathrm{TA}} = x_{\mathrm{TR}}$ 경우로 구분), 그리고 $\sigma_d^2 > \sigma_n^2$일 때는 분류점 $x_{\mathrm{TR}} < x_{\mathrm{MFR}}$ 관계가 있다. 따라서 분류점 x_{MFR}이 두 분류점 x_{TA}, x_{TR}과 서로 다를 때 정규화된 기대비용(NEC)가 서로 다른 경우에 대하여 비용비율을 살펴보자. 우선 분류점이 $x_{\mathrm{TA}} < x_{\mathrm{MFR}} < x_{\mathrm{TR}}$인 경우 B1에서는 $\mathrm{NEC}_{\mathrm{MFR}} < \mathrm{NEC}_{\mathrm{TA}}$ 그리고 $\mathrm{NEC}_{\mathrm{MFR}} < \mathrm{NEC}_{\mathrm{TR}}$를 동시에 만족하여야 하고, 분류점이 $x_{\mathrm{MFR}} < x_{\mathrm{TA}} < x_{\mathrm{TR}}$인 경우 B2에서는 $\mathrm{NEC}_{\mathrm{MFR}} < \mathrm{NEC}_{\mathrm{TA}}$을 만족하여야 한다. 분류점 $x_{\mathrm{TR}} < x_{\mathrm{MFR}}$인 경우 C에서는 $\mathrm{NEC}_{\mathrm{MFR}} < \mathrm{NEC}_{\mathrm{TR}}$을 만족하여야 하므로 각각의 경우에 비용비율이 갖을 수 있는 값의 범위는 다음과 같이 요약한다.

정리 3.2 (경우 B1) $x_{\mathrm{TA}} < x_{\mathrm{MFR}} < x_{\mathrm{TR}}$, (경우 B2) $x_{\mathrm{MFR}} < x_{\mathrm{TA}} < x_{\mathrm{TR}}$ 그리고 (경우 C) $x_{\mathrm{TR}} < x_{\mathrm{MFR}}$의 하에서의 비용비율은 각각 다음과 관계를 갖는다.

$$\frac{F_d(x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TR}})}{F_n(x_{\mathrm{MFR}}) - F_n(x_{\mathrm{TR}})} \frac{\gamma}{1-\gamma} < \mathrm{CR}_{\mathrm{B1}} < \frac{F_d, (x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TA}})}{F_n(x_{\mathrm{MFR}}) - F_n(x_{\mathrm{TA}})} \frac{\gamma}{1-\gamma}, \tag{3.4}$$

$$\mathrm{CR}_{\mathrm{B2}} > \frac{F_d(x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TA}})}{F_n(x_{\mathrm{MFR}}) - F_n(x_{\mathrm{TA}})} \frac{\gamma}{1-\gamma}, \tag{3.5}$$

$$\mathrm{CR}_C < \frac{F_d(x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TR}})}{F_n(x_{\mathrm{MFR}}) - F_n(x_{\mathrm{TR}})} \frac{\gamma}{1-\gamma}. \tag{3.6}$$

증명: 분류점이 $x_{\mathrm{TA}} < x_{\mathrm{MFR}} < x_{\mathrm{TR}}$인 경우 B1에서는 $\mathrm{NEC}_{\mathrm{MFR}} < \mathrm{NEC}_{\mathrm{TA}}$ 그리고 $\mathrm{NEC}_{\mathrm{MFR}} < \mathrm{NEC}_{\mathrm{TR}}$를 만족하여야 하고 각각의 경우에 PCF는 다음과 같이 하한값과 상한값을 얻는다.

$$\frac{1}{1 + \frac{F_d(x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TA}})}{F_n(x_{\mathrm{MFR}}) - F_n(x_{\mathrm{TA}})}} < \mathrm{PCF} < \frac{1}{1 + \frac{F_d(x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TR}})}{F_n(x_{\mathrm{MFR}}) - F_n(x_{\mathrm{TR}})}}.$$

이에 대한 비용비율의 하한과 상한값을 정리하면 식 (3.4)를 얻을 수 있다. 경우 B2와 경우 C에서는 각각 $\mathrm{NEC}_{\mathrm{MFR}} < \mathrm{NEC}_{\mathrm{TA}}$와 $\mathrm{NEC}_{\mathrm{MFR}} < \mathrm{NEC}_{\mathrm{TR}}$를 만족하여야 하므로 PCF는 다음과 같으므로 비용비율은 각각 식 (3.5)와 식 (3.6)을 얻는다.

$$\mathrm{PCF} < \frac{1}{1 + \frac{F_d(x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TA}})}{F_n(x_{\mathrm{MFR}}) - F_n(x_{\mathrm{TA}})}},$$

$$\mathrm{PCF} > \frac{1}{1 + \frac{F_d(x_{\mathrm{MFR}}) - F_d(x_{\mathrm{TR}})}{F_n(x_{\mathrm{MFR}}) - F_n(x_{\mathrm{TR}})}}.$$

□

Table 3.2. Cost ratio of case B1

$F_n(\cdot)$	γ	$\mathrm{CR_{TR}}$	$\mathrm{CR_{MFR}}$	$\mathrm{CR_{B1}}$
$N(1, 1.2^2)$	0.371	0.5898	0.8	(0.6896, 0.8954)
	0.4	0.6667	0.9047	(0.7795, 0.9512)
	0.42	0.7241	0.9827	(0.8467, 0.9911)
$N(1, 1.4^2)$	0.365	0.5748	0.9697	(0.7642, 0.9847)
	0.371	0.5898	0.9951	(0.7842, 0.9975)

Table 3.3. Cost ratio of case B2

$F_n(\cdot)$	γ	$\mathrm{CR_{TR}}$	$\mathrm{CR_{MFR}}$	$\mathrm{CR_{B2}}$
$N(1, 1.2^2)$	0.44	0.7857	1.0657	1.0325
	0.46	0.8518	1.1554	1.0578
	0.48	0.9231	1.2520	1.1212
$N(1, 1.4^2)$	0.44	0.7857	1.3251	1.1594
	0.46	0.8518	1.4367	1.2126
	0.48	0.9231	1.5567	1.2692

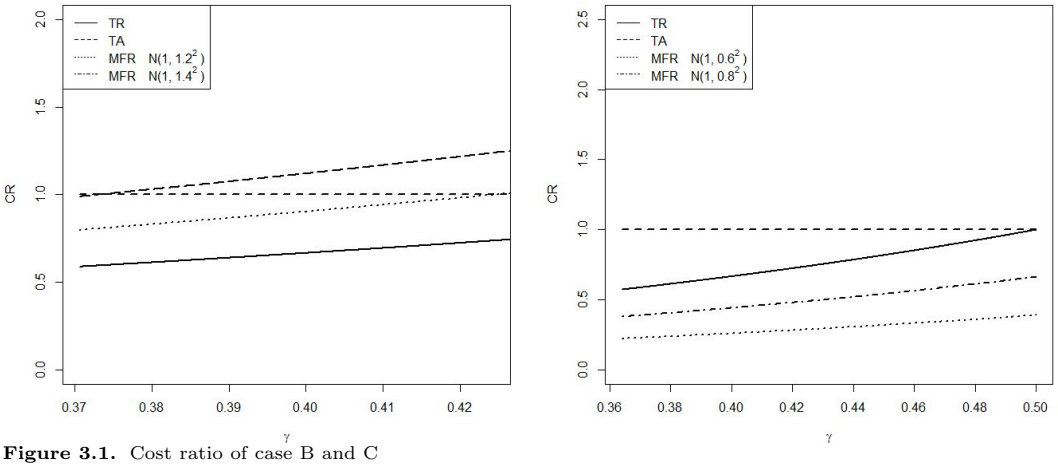

Figure 3.1. Cost ratio of case B and C

경우 B1, B2, C를 만족하는 다양한 정규분포를 고려하면서 각각의 비용비율을 구한다. $F_d(\cdot)$를 표준정규분포 $N(0,1)$ 그리고 $F_n(\cdot)$을 평균이 1인 정규분포 $N(1,\sigma_n^2)$로 간주한다. 세 가지 경우에 따라 분산 σ_n^2을 다양하게 변화시켜 TA, TR, MFR를 각각 최대화했을 때의 $\mathrm{CR_{TA}}$, $\mathrm{CR_{TR}}$, $\mathrm{CR_{MFR}}$ 비용비율과 정리 3.2에서 $\mathrm{NEC_{MFR}}$을 최소화했을 때의 경우 B1, B2, C의 비용비율의 범위를 각각 $\mathrm{CR_{B1}}$, $\mathrm{CR_{B2}}$, $\mathrm{CR_C}$을 구하고 분석한다. 여기서 $\mathrm{CR_{TA}}$는 항상 1이기 때문에 Table에서 제외했다.

경우 A는 Hong과 Yoo (2010)와 동일하기 때문에 제외했다. 경우 B1인 $\sigma_d^2 < \sigma_n^2$의 경우에 σ_n^2을 1보다 큰 1.2^2, 1.4^2로 설정하고 다양한 γ에 대하여 비용비율 $\mathrm{CR_{TR}}$, $\mathrm{CR_{MFR}}$과 $\mathrm{CR_{B1}}$를 Table 3.2에 요약하였다. Table 3.2를 통해 살펴보면 σ_n^2의 분산이 각각 증가할수록 세 종류의 비용비율은 증가한다. 특히 γ가 증가할수록 비용비율은 모두 급격히 상승하며, $\mathrm{CR_{MFR}}$이 1보다 작다. 경우 B2는 $\sigma_d^2 < \sigma_n^2$의 경우이지만 경우 B1보다 γ가 높을 경우에 발생한다. Table 3.3을 살펴보면 $\mathrm{CR_{MFR}}$이 1보다 크고 $\mathrm{CR_{B2}}$도 1을 초과한다. Figure 3.1의 왼쪽을 통해 살펴보면, 경우 B1에서는 $\mathrm{CR_{TR}} < \mathrm{CR_{MFR}} < \mathrm{CR_{TA}} = 1$이

Table 3.4. Cost ratio of case C

$F_n(\cdot)$	γ	$\mathrm{CR_{TR}}$	$\mathrm{CR_{MFR}}$	$\mathrm{CR_C}$
$N(1, 0.6^2)$	0.25	0.333	0.1314	0.1980
	0.35	0.5385	0.2122	0.3198
	0.45	0.8182	0.3225	0.4858
$N(1, 0.8^2)$	0.38	0.6129	0.4058	0.4957
	0.42	0.7141	0.4795	0.5858
	0.45	0.8518	0.5418	0.6618

Table 4.1. Confusion matrix

Rate		State of Nature		
		e_1	e_2	e_3
Decision	d_1	$F_1(x_1)$	$F_2(x_1)$	$F_3(x_1)$
	d_2	$F_1(x_2) - F_1(x_1)$	$F_2(x_2) - F_2(x_1)$	$F_3(x_2) - F_3(x_1)$
	d_3	$1 - F_1(x_2)$	$1 - F_2(x_2)$	$1 - F_3(x_2)$

고 $\mathrm{CR_{B1}}$의 구간은 $\mathrm{CR_{MFR}}$보다 크고 $\mathrm{CR_{TA}}$보다 작다. 그리고 경우 B2에서는 $\mathrm{CR_{TR}} < \mathrm{CR_{TA}} = 1 < \mathrm{CR_{B2}} < \mathrm{CR_{MFR}}$임을 파악할 수 있다. 일반적으로 손실비용 $(\mathrm{C_{FN}} - \mathrm{C_{TP}})$이 기회비용 $(\mathrm{C_{FP}} - \mathrm{C_{TN}})$보다 큰 $\mathrm{C_{FN}} - \mathrm{C_{TP}} > \mathrm{C_{FP}} - \mathrm{C_{TN}}$이므로 비용비율은 1보다 작다. 따라서 $\mathrm{CR_{MFR}} > 1$이므로 현실적이지 않다. 그리고 $\mathrm{CR_{MFR}} > \mathrm{CR_{TR}}$이므로 TR에 비해서 MFR의 손실비용 $(\mathrm{C_{FN}} - \mathrm{C_{TP}})$이 기회비용 $(\mathrm{C_{FP}} - \mathrm{C_{TN}})$보다 크다고 판단할 수 있다.

경우 C인 $\sigma_d^2 > \sigma_n^2$인 경우에 σ_n^2을 1보다 작은 0.6^2, 0.8^2으로 설정하고 다양한 γ에 대하여 $\mathrm{CR_{TR}}$, $\mathrm{CR_{MFR}}$과 $\mathrm{CR_C}$를 Table 3.4에 요약하고 이를 Figure 3.1의 오른쪽에 표현하였다. Figure 3.1를 살펴보면 γ가 증가할수록 세 종류의 비용비율은 모두 급격히 상승하는 것은 경우 B와 동일하다. 그러나 경우 B와 다른 점으로 $\mathrm{CR_{MFR}} < \mathrm{CR_{TR}} < \mathrm{CR_{TA}} = 1$이며 분산이 1보다 작아질수록 $\mathrm{CR_{MFR}}$이 $\mathrm{CR_{TR}}$보다 작아지는 것을 파악할 수 있다. 따라서 현실적인 $\mathrm{CR_{MFR}} < 1$을 만족하고 특히 $\mathrm{CR_{MFR}} < \mathrm{CR_{TR}}$이므로 TR에 비해서 MFR의 비용비율이 작으며 따라서 정규화된 기대비용(NEC)가 작다. 그러므로 경우 C는 $F_d(\cdot)$의 분산이 $F_n(\cdot)$의 분산보다 큰 실제 상황의 경우이며, TA와 TR 보다는 MFR을 사용하여 구한 최적분류점에 대한 비용비율이 현실적이며 기대비용이 작다는 것을 확인할 수 있다.

4. 다차원 ROC 분석에서의 MFR

본 논문의 앞부분에서는 두 분포함수 $F_d(x)$, $F_n(x)$에 대한 최적분류기준을 특히 오분류율곱(MFR)에 대하여 연구하였다. 본 절에서는 분류모형을 세 분포함수 이상에 대하여 일반적인 k범주로 판별하는 문제에서 MFR을 확장 연구한다. 우선 삼차원에 대하여 살펴본다. 확률변수 X에 분포함수를 $F_i(x)$, $i = 1, 2, 3$로 정의하고 모든 x에서 $F_1(x) \geq F_2(x) \geq F_3(x)$로 가정한다. 그리고 임의의 절단점 x_1, $x_2(x_1 \leq x_2)$에서 삼차원 혼동행렬은 Table 4.1과 같이 나타난다.

ROC 곡면(surface)은 $(F_1(x_1), F_2(x_2) - F_2(x_1), 1 - F_3(x_2))$를 좌표로 작성한 그림으로(Hong 등, 2013), 삼차원에서의 MFR은 삼차원 그림에서 (1, 1, 1) 좌표로부터 ROC곡면까지의 직선을 대각선으로 하는 사각형의 부피를 최대화하는 기준이며, 최대화될 때의 ROC 곡면상의 점에 대응하는 스코어를 분류점으로 설정한다. 식 (2.1)의 MFR을 삼차원으로 확장하면서 정의 4.1에 설명하고, ROC 곡면 윗부분에 나타난 육면체의 부피로 표현할 수 있다.

정의 4.1 ROC 곡선에서 MFR은 다음과 같이 정의한다.

$$\mathrm{MFR}^3 = \max\{(1 - F_1(x_1)) \cdot (1 - (F_2(x_2) - F_2(x_1))) \cdot F_3(x_2)\}.$$

4차원 이상인 k차원까지 확장하여 MFR에 대하여 살펴보자. 분포함수를 $F_i(x),\ i = 1, \ldots, k$로 정의하고 모든 x에서 $F_1(x) \cdots F_k(x)$로 가정한다. 그리고 순서있는 절단점 $x_1, \ldots, x_{k-1}(x_1 \leq \cdots \leq x_{k-1})$에 대하여 k차원으로 확장하여 혼동행렬을 고려할 수 있고 이를 바탕으로 MFR은 다음과 같이 정의할 수 있다.

정의 4.2 k차원 ROC 분석의 MFR 기준은 다음과 같이 정의한다.

$$\mathrm{MFR}^k = \max\{(1 - F_1(x_1)) \cdots (1 - (F_i(x_i) - F_i(x_{i-1}))) \cdots F_k(x_{k-1})\}.$$

Zhou 등 (2002)은 k차원 혼동행렬에서의 대각원소 빈도들의 합을 전체 빈도수로 나눈 값으로 정분류율(correct classification rates; CCR)을 정의하였고 각 분포에서의 정분류율(CCR_i)의 가중평균으로 제안하였다. 본 연구에서 제안한 k차원의 MFR과 CCR과의 관계를 정리하면 다음과 같다.

$$\mathrm{MFR}^k = \prod_{i=1}^{k} (1 - \mathrm{CCR}_i),$$

여기서 각 분포의 정분류율은 $\mathrm{CCR}_1 = F_1(x_1)$, $\mathrm{CCR}_i = F_i(x_i) - F_i(x_{i-1})$, i= $2, 3, \ldots, k-1$, $\mathrm{CCR}_k = 1 - F_k(x_{k-1})$이다.

5. 결론

본 연구는 ROC 곡선에서 면적 형태로 표현되는 최적분류기준인 오분류율곱(MFR) 기준을 제안하였다. MFR 분류기준은 (0, 1) 좌표로부터 ROC 곡선까지의 직선을 대각선으로 하는 ROC 곡선의 윗부분의 직각사각형의 면적을 최대화하는 기준이며 이 면적값을 최대화될 때의 ROC 곡선상의 점에 대응하는 스코어를 분류점으로 설정한다.

MFR 기준으로부터 얻는 최적분류점과 다른 분류기준들과 분류성과를 비교하였다. 분류성과를 분석하기 위해서 다양한 분포함수에 대하여 TA, TR과 MFR 기준에 기반하는 최적분류점을 구하고 이에 대응하는 제1종 오류와 제2종 오류 그리고 두 오류의 합을 비교하면서 MFR의 장점을 토론하였다. 대부분의 실제 상황에서는 $F_d(\cdot)$의 분산이 $F_n(\cdot)$의 분산보다 큰 경우가 많다. 이런 상황에서는 MFR 기준으로 구한 최적분류점에 대한 제1종 오류인 FNR이 TR이나 TA 기준으로 얻은 최적분류점의 제1종 오류보다 작은 값을 갖는다. 따라서 중요하고 손실함수가 큰 제1종 오류를 최소화하는 측면에서 살펴보면, MFR 기준으로 구한 최적분류점이 본 연구에서 논의한 최적분류점 기준 중에서 현실적이라고 파악할 수 있다.

일반적인 비용곡선을 이용하여 본 연구에서 제안한 MFR을 최대화하는 비용비율을 정의하였다. 대표적인 분류기준인 TA와 TR 그리고 MFR로 구한 각각의 최적분류점에 대응하는 NEC와 PCF에 대한 비용비율을 살펴보면서, 비용곡선으로부터 TA, TR, MFR 기준으로 구한 분류점들에 대한 비용비율의 관계를 정리하고, 정규화된 기대비용 측면에서 MFR이 선호되는 비용비율을 살펴 보았다. 많은 실증 예제에서와 같이 $F_d(\cdot)$의 분산이 $F_n(\cdot)$의 분산보다 클수록 MFR에 대응하는 비용비율이 TR에 대응하는 비용비율보다 작은 값을 갖는 것을 발견하면서 비용비율의 측면에서의 MFR 기준은 현실적이라는 장점을 유도하였다. 마지막으로 이차원의 ROC 곡선을 확장한 다차원의 ROC 분석에서 MFR 기준을 정의하고 MFR과 다른 기준과의 관계를 정리하면서 다차원으로 확대 가능함을 설명하였다.

References

Adams, N. M. and Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain, *Pattern Recognition*, **30**, 1139–1147.

Antonie, M. L., Zaiane, O. R. and Holte, R. C. (2006). Learning to use a learned model: A two-stage approach to classification, *Proceedings of the* 6^{th} *IEEE International Conference on Data Mining(ICDM'06)*, 33–42.

Brasil, P. (2010). Diagnostic test accuracy evaluation for medical professionals, Package DiagnosisMed in R.

Briggs, W. M. and Zaretzki, R. (2007). The skill plot: A graphical technique for the evaluating the predictive usefulness of continuous diagnostic tests, *Biometrics*, **63**, 250–261.

Cantor, S. B., Sun, C. C., Tortolero-Luna, G., Richards-Korturn, R. and Follen, M. (1999). A comparison of CB ratios from studies using receiver operating characteristic curve analysis, *Journal of Clinical Epidemiology*, **52**, 885–892.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves, *Proceedings of the* 23^{rd} *International Conference on Machine Learning(ICML'06)*, 233–240.

Drummond, C. and Holte, R. (2000). Explicitly representing expected cost: An alternative to ROC representation, Technical Report, School of Information Technology and Engineering, University of Ottawa.

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*, Frankfurt.

Fawcett, T. (2003). ROC Graphs: Notes and practical considerations for data mining researchers, *Technical Report HPL-2003-4*, HP Laboratories Palo Alto, 1–28, Palo Alto.

Freeman, E. A. and Moisen, G. (2008). PresenceAbsence: An R package for presence absence analysis, *Journal of Statistical Software*, **23** 1–31.

Greiner, M. and Gardner, I. A. (2000). Epidemiologic issues in the validation of veterinary diagnostic tests, *Preventive Veterinary Medicine*, **45**, 3–22.

Hand, D. J. (2009). Mismatched models, wrong results, and dreadful decisions: On choosing appropriate data mining tools, *Proceedings of the* 15^{th} *ACM SIGKDD international conference on Knowledge discovery and data mining.*

Hand, D. J. and Zhou, F. (2009). Evaluating models for classifying customers in retail banking collections, *Journal of the Operational Society*, DOI: 10.1057/jors.2009.129, London.

Hilden, J. and Glasziou, P. (1996). Regret graphs, diagnostic uncertainty and Youden's index, *Statistics in Medicine*, **15**, 969–986.

Holte, R. C. and Drummond, C. (2008). Cost-sensitive classifier evaluation using cost curves, *Advances in Knowledge discovery and Data Mining*, **5012**, 26–29

Hong, C. S. (2009). Optimal threshold from ROC and CAP curves, *Communications in Statistics-Simulation and Computation*, **38**, 2060–2072.

Hong, C. S. and Lee W. Y. (2011). ROC curve fitting with normal mixture, *The Korean Journal of Applied Statistics*, **24**, 269–278.

Hong, C. S. and Yoo, H. S. (2010). Cost ratios for cost and ROC curves, *Communications of The Korean Statistical Society*, **17**, 755–765.

Hong, C. S. and Joo, J. S. (2010). Optimal thresholds from non-normal mixture, *The Korean Journal of Applied Statistics*, **23**, 943–953.

Hong, C. S., Joo, J. S. and Choi, J. S. (2010). Optimal thresholds from mixture distributions, *The Korean Journal of Applied Statistics*, **23**, 13–28.

Hong, C. S., Lin, M. H. and Hong, S.W. (2011). ROC function estimation, *The Korean Journal of Applied Statistics*, **24**, 987–994.

Hong, C. S., Jung, E. S. and Jung, D. G. (2013). Standard criterion of VUS for ROC surface, *The Korean Journal of Applied Statistics*, **26**, 977–985.

Hoshino, R., Coughtrey, D., Sivaraja, S., Volnyansky, I., Auer, S. and Trishtchenko, A. (2009). Applications and extensions of cost curves to marine container inspection, *Annals of Operations Research*, DOI:

10.1007/s10479-009-0669-2.

Jund, J., Rabillous, M., Wallon, M. and Ecochard, R. (2005). Methods to estimate the optimal threshold for normally or log-normally distributed biological tests, *Medical Decision Making*, **25**, 406–415.

Kim, J. H. (2004). Roc and cost graphs for general cost matrix where correct classification incur non-zero costs, *Communications of the Korean Statistical Society*, **11**, 21–30.

Liu, Y. and Shriberg, E. (2007). Comparing evaluation metrics for sentence boundary detection, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP'07)*, **4**, 185–188.

Liu, Z., Tan, M. and Jiang, F. (2009). Regularized F-measure maximization for feature selection and classification, *Journal of Biomedicine and Biotechnology*, 617946.

Lambert, J. and Lipkovich, I. (2008). A macro for getting more out of your ROC curve, SAS Global forum, paper 231, Indianapolis.

Metz, C. E. (1978). Basic principles of ROC analysis, *Seminars in Nuclear Medicine*, **8**, 283–298.

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

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

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

Turney, P. D. (1995). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm, *Journal of Artificial Intelligence Research*, **2**, 369–409.

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

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

Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*, Wiley, New York.

대안적인 분류기준: 오분류율곱

홍종선[a,1] · 김효민[a] · 김동규[a]

[a]성균관대학교 통계학과

(2014년 8월 12일 접수, 2014년 9월 29일 수정, 2014년 9월 30일 채택)

요 약

본 연구는 ROC 곡선에서 형성되는 면적 형태로 나타나는 분류정확도기준인 오분류율곱(multiplication of false rates; MFR)를 제안한다. MFR 기준과 다른 기준로부터 구한 최적분류점의 분류성과에 대하여 비교 분석한다. 다양한 분포함수에 대하여 최적분류점을 구하고 이에 대응하는 FNR과 FPR을 비교하면서 MFR의 특징과 장점을 유도한다. 일반적인 비용함수를 바탕으로 분류점에 대한 비용비율을 다양한 분류기준을 이용하여 구한다. 비용곡선에 대한 비용비율의 관계를 정리하여 MFR 기준의 장점을 탐색한다. MFR 기준의 정의를 다차원 ROC 분석으로 확장하고 다차원의 다른 분류기준과의 관계를 설명하면서 토론한다.

주요용어: 부도, 분류점, 분류성과, 비용비율, 혼동행렬.

[1]교신저자: (110-745) 서울시 종로구 명륜동 3가 53, 성균관대학교 경제대학 통계학전공.
E-mail: cshong@skku.edu