한국통계학회논문집
2010, 17권, 6호, 755–765

DOI: 10.5351/CKSS.2010.17.6.755

비용곡선과 ROC곡선에서의 비용비율

홍종선[1,a], 유현상[a]

[a]성균관대학교 통계학과

요 약

혼합분포의 분류문제에서 비용함수를 고려한 분류점은 최소 기대비용이라는 측면에서 최적이다. 비용에 관한 어떠한 정보가 주어지지 않은 경우에 ROC곡선을 이용하여 분류정확도 측도인 전체정확도와 진실율이 최대일 때의 분류점에 대응하는 기대비용에서의 비용비율을 제안하고, 최소 기대비용의 비용비율과의 관계를 설명한다. 그리고 비용곡선을 이용하여 분류정확도 측도들에 기반하는 최소 기대비용에서의 비용비율을 제안하였고 이 비용비율은 대표적인 두 종류의 분류정확도가 최대일 때의 기대비용에 대한 비용비율들 사이에 존재하며, 최소 기대비용에서의 비용비율에 수렴하는 것을 발견하였다. 본 연구는 기대비용과 정규화된 기대비용을 최소화할 때의 비용비율과 분류정확도가 최대일 때의 비용비율들의 관계를 토론한다.

주요용어: 기대비용, 부도, 분류점, 분류정확도, 신용평가, 판별력.

1. 서론

두 분포함수의 혼합분포로부터 분류모형의 성과를 평가하는 문제를 고려하자. 분류문제는 경영학, 공학 그리고 의학 등의 여러 분야에서 사용되지만 본 연구는 신용평가모형을 기반으로 설명하고자 한다. 차주(borrower)의 미래상태인 부도상태(default) 혹은 정상상태(non-default)에 대한 예측력을 최대화하는 신용등급시스템(credit rating system)에서 모형의 판별력(discriminative power)은 차주(borrower, debtor)에게 대출여부를 결정할 때 부도(default, positive)인지 또는 정상(non-default, negative)인지의 미래 상태를 미리 식별해야 하므로 중요한 문제로 인식된다 (Servigny와 Renault, 2004). 왜냐하면 두 종류의 상태에 관한 예측문제에서 부도를 정상으로 또는 정상을 부도로의 예측은 큰 손실을 초래할 수 있기 때문이다.

확률변수 X는 차주에게 부여되는 연속형 등급스코어(rating score)이고, 확률밀도함수 $f(x)$는 부도 또는 정상상태의 모수로 이루어진 전체 모수공간 $\Theta = \{\theta_d, \theta_n\}$에서 정의되는 확률밀도함수들의 혼합함수로 다음과 같이 가정한다.

$$\begin{aligned} f(x\,|\,\theta) &= \gamma f(x\,|\,\theta_d) + (1-\gamma) f(x\,|\,\theta_n) \\ &\equiv \gamma f_d(x) + (1-\gamma) f_n(x), \quad x \in (-\infty,\ \infty), \end{aligned} \tag{1.1}$$

여기서 $f_d(x)$와 $f_n(x)$는 각각 부도와 정상의 조건부 확률밀도함수를 나타내고, γ는 모집단의 전체부도율(total probability of default)로서 혼합모수(mixing parameter)이며 일반적으로 전체 차주 중에서 부도차주는 정상차주보다 많지 않으므로 0.5보다 작다고 가정한다.

차주에 대한 미래 상태의 예측은 위의 혼합분포에서 X의 특정한 스코어 x_0를 통해 분류된다. 따라서 모형의 분류성과(classification performance)는 2×2 혼동행렬(confusion matrix)로 나타나는데 각

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

표 1: 비용 행렬

		실제 그룹(Actual Class)	
		Default/Positive	Non-default/Negative
예측된 그룹 (Predicted Class)	Default/Positive	C_{TP}	C_{FP}
	Non-default/Negative	C_{FN}	C_{TN}

칸은 TP(true positive), FP(false positive), TN(true negative), FN(false negative)으로 구성된다. TP와 TN은 정분류된 경우이고, FP와 FN은 오분류된 경우이다. 그리고 혼동행렬의 각 칸에 대응하는 비용(손실)을 나타낸 비용행렬(cost matrix)은 표 1과 같다. C_{FP}와 C_{FN}은 오분류 비용(misclassification costs), C_{TP}와 C_{TN}은 정분류 비용이며, 모두 비음수(non-negative)로 발생함을 가정한다.

위와 같은 비용행렬이 주어졌을 때 비용을 최소화하는 분류점은 다음과 같이 정의되는 기대비용(expected cost; EC) 함수를 통해 얻어질 수 있다 (Metz, 1978; Zhou 등, 2002; Pepe, 2003; 김지현, 2004).

$$\mathrm{EC} = C_0 + C_{TP}\gamma F_d(x) + C_{FN}\gamma(1 - F_d(x)) + C_{FP}(1-\gamma)F_n(x) + C_{TN}(1-\gamma)(1 - F_n(x)), \tag{1.2}$$

여기서 C_0는 고정비용이다. 기대비용 EC는 스코어 x의 함수로서 표 1에서 각 칸이 발생할 확률(열비율로 정의되고, 가정된 누적분포함수를 이용하여 얻어진다)에 기대비용(각 칸의 비용에 발생확률을 곱한다)을 곱해서 모두 더한 형태이다.

기대비용함수를 최소화하는 스코어를 찾는 방법은 최적분류점을 추정하는 문제로 간주되기 때문에 Jund 등 (2005), Hand (2009) 그리고 Hand와 Zhou (2009) 등은 기대비용함수를 최소화하는 최적분류점 x_0는 다음을 만족하는 스코어로 제안하였다.

$$\frac{f_d(x_0)}{f_n(x_0)} = \left(\frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}\right)\frac{1-\gamma}{\gamma}, \tag{1.3}$$

여기서 $(C_{FP} - C_{TN})/(C_{FN} - C_{TP})$는 비용비율(cost ratio; CR)로 정의하며, Hand (2009)는 위 식으로부터 기대비용함수를 최소화할 때의 비용비율은 최적분류점 x_0에 대응하는 가능도비 $f_d(x_0)/f_n(x_0)$와 부도오즈(the odds of default) $\gamma/(1-\gamma)$의 곱으로 식 (1.4)를 유도하였다.

$$\mathrm{CR} = \left(\frac{f_d(x_0)}{f_n(x_0)}\right)\frac{\gamma}{1-\gamma}. \tag{1.4}$$

실제로 비용함수를 이용한 접근은 특히 신용평가부분에서는 미래에 발생할 각각의 차주들이 서로 다른 비용위험을 갖고 있고, 예측의 불확실성이 존재하기 때문에 이론적으로 접근하기 어렵다 (Adams와 Hand, 1999; Liu, 2002). 최소 손실하의 분류점은 결국 그룹분포, 정확한 비용비율에 관한 정보 그리고 전체부도율 추정값이 분류점의 성과를 결정하게 된다. 만일 스코어의 확률밀도함수가 알려져 있고 특정한 분류점이 주어져서 식 (1.3)의 우측항의 값을 알고 있는 경우에는 좌측의 두 항들은 상호교환 가능한 관계를 갖게 된다. 따라서 최근의 연구들은 적절한 분류점을 찾기 위해서 전체부도율이나 혹은 비용비율에 대한 결정문제로 접근하게 된다. Cantor 등 (1999)는 다양한 진단시험에서의 비용비율에 대한 추정값을 제시하였으며, Adams와 Hand (1999)는 비용비율의 최대가능도 추정값과 구간문제를 언급하였다.

분류정확도 측도들의 결과를 그래픽적으로 살펴볼 수 있는 방법으로 performance graphs (Turney, 1995), regret graphs (Hilden과 Glasziou, 1996), ROC곡선, 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 (Briggs와 Zaretzki, 2007) 그리고 비용곡선(cost curve) 등이 있다. 이러한 그래픽적인 방법들 중에 몇몇 방법들은 분류문제에서 특히 오분류 비용이나 분포에 관해서 불확실성을 갖을 때 유용하며, 본 연구에서는 대표적인 ROC곡선과 비용곡선을 바탕으로 비용비율에 대하여 토론한다. ROC곡선은 두 분포함수의 혼합분포로부터 분류모형의 성과를 평가하는 그래픽적인 방법이며 (Provost와 Fawcett, 1997; Zhou 등, 2002; Drummond와 Holte, 2006; Tasche, 2006; Vuk와 Curk, 2006), 비용곡선은 수평축에 확률비용을 그리고 수직축은 혼합분포함수의 함수로서의 분류성과를 의미하는 정규화된 기대비용으로 표현하고 두 축의 좌표 모두를 0과 1사이의 값을 갖도록 표준화한 그래픽적 방법이다 (Drummond와 Holte, 2006; Holte와 Drummond, 2008; Hoshino 등, 2009).

분류성과를 결정하는 두 종류의 학습방법에는 최소 손실하의 비용-의존학습(cost-sensitive learning)을 통한 접근에서 정확한 비용비율에 관한 연구와 분류정확도에 기초하는 비용-비의존학습(cost-insensitive learning)에서는 높은 정확도를 나타내는 측도에 관한 연구로 구분된다. 먼저 비용-의존학습 측면에서 기대비용 EC 함수를 ROC곡선의 좌표 $F_n(x)$과 $F_d(x)$의 선형함수로 표현한 등비용선(iso-cost line)으로 변환하여 등비용선과 ROC곡선과의 접점에 대응하는 분류점에서 기대비용을 최소화하고 이때의 비용비율을 구할 수 있으나, 비용과 비용비율에 대한 정보의 부족으로 기대비용을 최소화하는 비용비율을 유도하는데 한계가 있다. 따라서 비용-비의존학습 측면에서 대표적인 두 종류의 분류정확도(classification accuracy) 측도를 등성과선(iso-performance line)으로 변환하여, ROC곡선과의 접점에서의 분류정확도가 최대이므로 이 접선에서의 기대비용을 최소로 간주하여 비용비율을 유도하고자 한다. 그리고 두 종류의 분류정확도 측도를 최대화하는 분류점에 대응하는 비용직선(cost line)들의 최소값으로 비용곡선을 나타내어, 특정한 비용에서 기대비용이 더 적은 분류정확도 측도를 선호하고, 선호된 분류정확도 측도에 대응하는 비용비율을 설정한다. 또한 본 연구에서 유도한 비용비율과 Hand (2009)가 제안한 비용비율의 관계를 토론하고자 한다.

본 연구의 구성은 다음과 같다. 2절에서는 두 종류의 분류정확도 측도들을 설명하고, ROC곡선에서 두 분류정확도를 최대화하는 분류점에 대응하는 비용비율을 제시하고, 3절에서는 비용곡선에 대한 설명과 두 분류정확도 측도 중에서 기대비용이 적은 측도에 기초하는 비용비율의 새로운 기준에 대해서 제안한다. 그리고 이러한 비용비율들의 관계를 2절과 3절에서 토론한다. 4절에서는 다양한 정규혼합분포를 가정하여 본 연구에서 제안한 비용비율을 구하고 분포함수의 평균과 분산 그리고 전체부도율의 변화에 따라 비용비율과의 관계를 설명한다. 마지막 5절에서는 결론을 유도한다.

2. ROC곡선과 비용비율

혼합분포로부터 분류모형의 성과를 측정하는 그래픽적 방법으로 많이 사용하는 ROC곡선이 있다. ROC곡선에서 EC를 최소화하는 분류점은 식 (1.2)의 EC 함수를 ROC곡선의 좌표인 $F_n(x)$과 $F_d(x)$의 선형함수로 식 (2.1)과 같이 표현한 등비용선과 ROC곡선과의 접점에서 구해질 수 있다 (Zweig와 Campbell, 1993; Cantor 등, 1999; Zhou 등, 2002; Liu, 2002; Fawcett, 2006; Kaivanto, 2008; Krzanowski와 Hand, 2009).

$$F_d(x) = \left(\frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}\right)\left(\frac{1-\gamma}{\gamma}\right)F_n(x) + \frac{C_0' - \mathrm{EC}}{C_{FN} - C_{TP}}\frac{1}{\gamma}, \tag{2.1}$$

여기서 $C_0' = C_0 + C_{FN}\gamma + C_{TN}(1-\gamma)$이다. 식 (2.1)의 기울기가 식 (1.3)과 동일함을 확인할 수 있다. 신용평가모형에서 제 I종 오류에 대응하는 비용(C_{FN})은 주요 손실금이 포함되고, 제 II종 오류(C_{FP})에 대응하는 비용은 좋은 고객이 갖는 이자소득의 손실금 및 대체 고객들에 대한 손실 등의 기회비용(opportunity cost)으로 구성되어 있으며 $C_{FN} \geq C_{FP}$이 성립된다 (Fielding과 Bell, 1997; Liu,

2002). ROC곡선과의 접선인 등비용선이 양의 기울기를 갖기 위해서는 $C_{FN} - C_{TP} \geq C_{FP} - C_{\mathrm{TN}}$이 성립되어야 하는 제약을 필요로 한다 (Kaivanto, 2008). 따라서 일반적인 조건은 $\mathrm{CR} \leq 1$이다.

이러한 최소 손실하의 비용-의존학습을 통한 접근에서 정확한 비용비율에 관한 정보가 분류성과를 결정하는데 비용과 비용비율에 대한 정확한 정보를 파악하기 어렵기 때문에 기대비용을 최소화하는 분류점을 발견하는데 한계가 있다. 따라서 분류정확도에 기초하는 비용-비의존학습을 이용하여 분류정확도를 최대화하는 분류점을 선택하고 이 분류점에서의 접선의 기울기를 식 (2.1)의 기울기와 동일하게 설정한다. 이때의 기대비용을 최소로 간주하여 비용비율을 구하여 본다. 비용-비의존 학습을 통한 분류점을 찾고자 할 때 가장 대표적으로 사용되는 분류정확도 통계량으로는 전체정확도(total accuracy; TA)와 진실율(true rate; TR)이 있다. 전체정확도 TA는 아래의 식 (2.2)에 정의되며 가장 일반적으로 사용되는 분류정확도 측도이다. 균형된 정확도(balanced accuracy) (Velez 등, 2007)이라 불리기도 하는 진실율 TR은 홍종선과 최진수 (2009)에 의해 제안되고 홍종선 등 (2010)이 이론적으로 발전시켰으며 TA와 비교하여 정확도와 비용 측면에서 장점을 가진 통계량으로 식 (2.3)과 같이 정의한다.

$$\mathrm{TA} = \gamma F_d(x) + (1-\gamma)(1 - F_n(x)), \tag{2.2}$$

$$\mathrm{TR} = \frac{1}{2}[F_d(x) + (1 - F_n(x))]. \tag{2.3}$$

다음의 정리 1과 2는 전체정확도 TA와 진실율 TR을 각각 최대화하는 분류점에 대응하는 비용비율을 제안하였다.

정리 1. 전체정확도 TA를 최대화하는 비용비율 CR은 다음과 같이 결정된다.

$$\mathrm{CR} = 1.$$

증명: Vuk과 Curk (2006)는 전체정확도 TA를 최대화하는 최적분류점을 다음과 같은 등성과선과 ROC곡선과의 접점으로부터 추정하였다.

$$F_d(x) = \frac{1-\gamma}{\gamma} F_n(x) + \frac{1}{\gamma}(\mathrm{TA} + \gamma - 1).$$

EC 함수를 등비용선으로 표현한 식 (2.1)의 기울기를 ROC곡선에서 TA를 최대화하는 등성과선의 기울기와 동일하게 다음과 같이 설정하면, 비용비율 CR은 1이 된다.

$$\left(\frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}\right) \frac{1-\gamma}{\gamma} = \frac{1-\gamma}{\gamma}. \tag{2.4}$$

□

정리 2. 진실율 TR을 최대화하는 비용비율 CR은 다음과 같다.

$$\mathrm{CR} = \frac{\gamma}{1-\gamma}.$$

증명: 홍종선 등 (2010)은 ROC곡선과 진실율 TR를 최대화하는 다음의 등성과선과의 접점으로부터 분류점을 추정하였다.

$$F_d(x) = F_n(x) + (2\,\mathrm{TR} - 1). \tag{2.5}$$

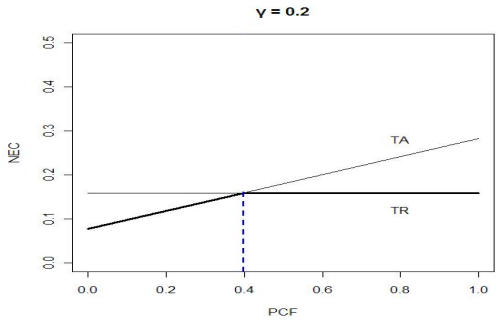

그림 1: *두 비용직선과 비용곡선*

EC 함수를 등비용선으로 표현한 식 (2.1)의 기울기를 ROC곡선에서 TR을 최대화하는 등성과선의 기울기와 동일하게 다음과 같이 설정하면, 비용비율 CR은 $\gamma/(1-\gamma)$이 된다.

$$\left(\frac{C_{FP}-C_{TN}}{C_{FN}-C_{TP}}\right)\frac{1-\gamma}{\gamma}=1.$$

□

Hand (2009)가 제안한 비용비율 식 (1.4)와 정리 1과 2에서 제안한 비용비율과의 관계를 보조정리 1에서 설명한다.

보조정리 1. Hand (2009)가 제안한 식 (1.4)의 비용비율이 TA와 TR을 최대화하는 조건과 결합하면 정리 1과 2의 비용비율과 동일하다.

증명: 홍종선 등 (2010)은 TA와 TR을 최대화하는 조건은 $f_d(x)/f_n(x)$이 $(1-\gamma)/\gamma$과 1로 각각 제안하였다. 이 조건을 식 (1.4)에 각각 대입하면, 정리 1과 2에서 유도한 CR = 1 그리고 CR = $\gamma/(1-\gamma)$이 된다. □

정리 1에서는 CR이 1이며 정리 2에서는 γ가 0.5미만일 때는 CR이 1보다 작기 때문에 일반적인 비용비율의 조건 CR ≤ 1을 만족한다.

3. 비용곡선과 비용비율

비용비율을 구하기 위하여 2절에서는 기대비용을 최소화하는 방법을 ROC곡선을 이용하였고, 3절에서는 정규화된 기대비용을 최소화하기 위하여 비용곡선을 이용하여 비용비율을 유도하고자 한다.

비용곡선은 분류결과의 성과에 대한 비용분석이 요구될 때의 확률비용과 특정한 분류점에서의 기대비용을 이차원 평면에 나타내어 분류성과를 그래픽적으로 표현한 방법으로, 비용곡선의 수평축은 비용비율 CR과 γ로 이루어진 확률비용함수(probability cost function; PCF)이며, 비용곡선의 수직축은 분류성과를 나타내는 정규화된 기대비용(normalized expected cost; NEC)으로 다음과 같이 정의된다.

$$\text{PCF}=\frac{(C_{FN}-C_{TP})\gamma}{(C_{FN}-C_{TP})\gamma+(C_{FP}-C_{TN})(1-\gamma)}=\frac{1}{1+\text{CR}(1-\gamma)/(\gamma)}, \tag{3.1}$$

$$\text{NEC}=F_n(x)+(1-F_d(x)-F_n(x))\text{PCF}. \tag{3.2}$$

비용곡선의 축을 나타내는 PCF와 NEC는 모두 0과 1사이의 값을 갖도록 표준화하였다. ROC곡선에서의 모든 점 $(F_n(x),\ F_d(x))$들에 대응하는 비용직선(cost line)으로 나타낼 수 있고, 모든 비용직선들의 최소덮개(minimum envelope 또는 lower envelope)를 연결하여 비용곡선을 얻는다. 만약 두 종류의 분류정확도 측도를 최대화하는 분류점에 대응하는 비용직선에서 NEC가 가장 적은 선을 연결하면 간단한 비용곡선이 되며 두 종류의 분류정확도 측도 중에서 어느 구간의 PCF에서 선호되는 측도가 무엇인지를 파악할 수 있다.

전체부도율 $\gamma = 0.2$이고 $f_d(x)$가 표준정규분포 $\phi(x|0,\ 1)$, $f_n(x)$는 평균이 2이고 분산이 1인 정규분포 $\phi(x|2,\ 1)$인 경우에, TA와 TR을 최대화하는 분류점은 각각 0.5763과 1이며, ROC곡선에서의 좌표는 각각 $(0.0773,\ 0.7178)$과 $(0.1587,\ 0.8413)$이다. 식 (3.1)과 (3.2)를 이용하여 구한 비용직선을 그림 1에 나타내었다. TA에 기초한 비용직선은 PCF에 비례하며, 반면에 TR에 기초한 비용직선은 PCF에 일정하다. 두 비용직선의 교차점 보다 큰 PCF > 0.3793 (CR < 0.3973)일때 TR의 NEC가 더 낮고, 그 외에는 TA의 NEC가 더 낮게 나타난다. 다음의 정리는 TA와 TR의 비용직선들을 살펴보고, 그들 중에 기대비용이 적은 확률비용함수의 구간에서 비용비율을 결정할 수 있는 기준을 제시한다.

대표적인 분류정확도 측도인 TA와 TR을 바탕으로 PCF와 NEC로 표현되는 비용직선을 구하고, 두 직선에서의 최소덮개를 설정하여 비용비율을 살펴보자.

정리 3. 전체정확도 TA와 진실율 TR을 최대화하는 분류점에 대응하는 ROC곡선에서의 좌표를 각각 $(F_n(x)_{TA},\ F_d(x)_{TA})$와 $(F_n(x)_{TR},\ F_d(x)_{TR})$이고, 두 분류점들에 대응하는 정규화된 기대비용함수를 각각 NEC_{TA}와 NEC_{TR}라고 할 때, NEC_{TR}이 NEC_{TA} 보다 작거나 같을 경우에 CR은 다음과 같은 상한값 아래에서 결정한다.

$$\left(\frac{F_d(x)_{TR} - F_d(x)_{TA}}{F_n(x)_{TR} - F_n(x)_{TA}}\right)\frac{\gamma}{1-\gamma}. \tag{3.3}$$

증명: 식 (2.4)의 기울기 $(1-\gamma)/\gamma$는 $\gamma < 0.5$인 경우에 식 (2.5)의 기울기 1보다 크므로 TA와 TR을 최대화하는 분류점의 각 좌표는 $F_d(x)_{TR} > F_d(x)_{TA}$, $F_n(x)_{TR} > F_n(x)_{TA}$이므로 $\mathrm{NEC}_{TA} \geq \mathrm{NEC}_{TR}$인 점에서의 PCF은 다음을 만족한다.

$$\mathrm{PCF} \geq \frac{1}{1 + \dfrac{F_d(x)_{TR} - F_d(x)_{TA}}{F_n(x)_{TR} - F_n(x)_{TA}}}.$$

확률비용함수의 식 (3.1)로부터 $\mathrm{NEC}_{TA} \geq \mathrm{NEC}_{TR}$을 만족하는 CR은 다음과 같이 구한다.

$$\mathrm{CR} \leq \left(\frac{F_d(x)_{TR} - F_d(x)_{TA}}{F_n(x)_{TR} - F_n(x)_{TA}}\right)\frac{\gamma}{1-\gamma}.$$

□

정리 3에서 제안한 비용비율의 상한값과 Hand (2009)의 비용비율과의 관계는 보조정리 2에서 설명한다.

보조정리 2. TR을 최대화하는 분류점에서의 정규화된 기대비용이 TA에 대응하는 정규화된 기대비용보다 작거나 같을 경우에 CR의 상한값은 전체부도율 γ가 0.5로 접근할 때 Hand (2009)가 제안한 비용비율 식 (1.4)로 수렴한다.

표 2: 정상상태 분포의 평균 변화에 따른 비용비율

$\phi(x;\mu_n,1)$	γ	CR_2	CR_3
$\phi(2,1)$	0.3	0.4286	0.6506
	0.35	0.5385	0.7320
	0.4	0.6667	0.8159
	0.45	0.8182	0.9045
$\phi(3,1)$	0.3	0.4286	0.6528
	0.35	0.5385	0.7330
	0.4	0.6667	0.8162
	0.45	0.8182	0.9045
$\phi(4,1)$	0.3	0.4286	0.6536
	0.35	0.5385	0.7333
	0.4	0.6667	0.8164
	0.45	0.8182	0.9045
$\phi(5,1)$	0.3	0.4286	0.6540
	0.35	0.5385	0.7335
	0.4	0.6667	0.8164
	0.45	0.8182	0.9045

표 3: 정상상태 분포의 분산 변화에 따른 비용비율

$\phi(x;2,\sigma_n^2)$	γ	CR_2	CR_3
$\phi(2,0.5^2)$	0.3	0.4286	0.5990
	0.35	0.5385	0.6991
	0.4	0.6667	0.7994
	0.45	0.8182	0.8998
$\phi(2,1^2)$	0.3	0.4286	0.6506
	0.35	0.5385	0.7320
	0.4	0.6667	0.8159
	0.45	0.8182	0.9045
$\phi(2,1.5^2)$	0.3	0.4286	0.6706
	0.35	0.5385	0.7439
	0.4	0.6667	0.8215
	0.45	0.8182	0.9059
$\phi(2,2^2)$	0.3	0.4286	0.6874
	0.35	0.5385	0.7523
	0.4	0.6667	0.8250
	0.45	0.8182	0.9068

증명: γ가 0.5에 가까운 값을 가지면, TA를 최대화하는 분류점은 TR을 최대화하는 분류점으로 수렴하므로 Δx를 0으로 접근시키는 것과 동일하다. 즉 정리 3의 CR에 Δx를 0으로 수렴시키면, 다음과 같이 식 (1.4)의 CR이 된다.

$$\begin{aligned}\lim_{\Delta x\to 0} \mathrm{CR} &= \lim_{\Delta x\to 0}\left(\frac{(F_d(x)_{TR}-F_d(x)_{TA})/\Delta x}{(F_n(x)_{TR}-F_n(x)_{TA})/\Delta x}\right)\frac{\gamma}{1-\gamma}\\ &\equiv \lim_{\Delta x\to 0}\left(\frac{\Delta F_d(x)/\Delta x}{\Delta F_n(x)/\Delta x}\right)\frac{\gamma}{1-\gamma}\\ &= \frac{f_d(x)}{f_n(x)}\frac{\gamma}{1-\gamma}.\end{aligned}$$

□

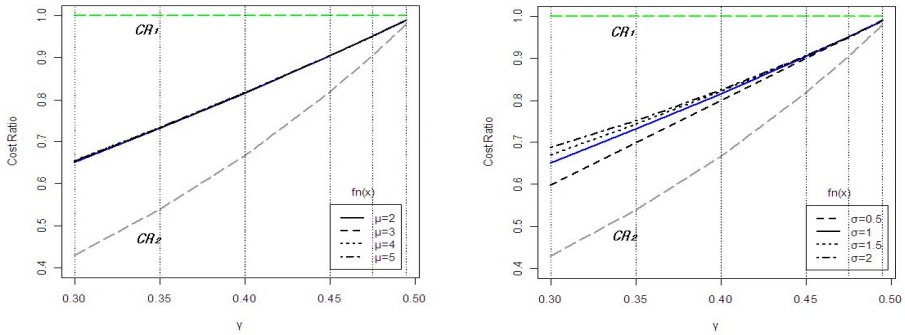

그림 2: *정상상태 분포의 평균과 분산 변화에 따른 비용비율*

정리 3에서 얻은 비용비율의 상한값은 Hand (2009)가 제안한 비용비율로 수렴하며, 이 상한값은 정리 1과 정리 2에서 제안한 비용비율 사이인 $(\gamma/(1-\gamma), 1)$에 존재한다. 분류정확도 측도인 TR에 대응하는 정규화된 기대비용이 TA에 대응하는 정규화된 기대비용 보다 작다면 비용비율이 $\gamma/(1-\gamma)$와 식 (3.3) 사이에 존재하고, 그 이외에는 비용비율이 식 (3.3)과 1 사이에 존재한다.

4. 정규혼합분포 예제

본 연구에서 제안하는 비용비율에 관한 이론을 통계적 확률밀도함수를 이용하여 비교 토론하고자 한다. 식 (1.1)에서의 부도와 정상상태의 조건부 확률밀도함수 $f_d(x)$과 $f_n(x)$를 각각 표준정규분포 $\phi(x|0,\ 1)$와 정규분포 $\phi(x|\mu_n, \sigma_n^2)$으로 가정한다.

우선 정상의 확률밀도함수의 분산을 1로 부도상태의 분산과 일치시키고, 정상상태의 평균을 2부터 5까지 그리고 전체부도율 γ를 0.3부터 0.45까지 변화시키면서 비용비율을 구하여 표 2에 정리하고, 다음으로는 정상의 확률밀도함수의 평균을 2로 고정시키고, 표준편차를 0.5부터 2까지 0.5만큼 증가시키면서 그리고 전체부도율 γ는 표 2에서와 같이 0.3부터 0.45까지 변화시키면서 구한 비용비율을 표 3에 나타내었다.

본 논문 2절의 정리 1과 2에서 유도한 두 종류의 분류정확도 통계량인 전체정확도와 진실율이 최대일 때의 분류점에 대응하는 기대비용 EC에서의 비용비율을 각각 CR_1과 CR_2으로, 3절의 정리 3에서 제안한 정규화된 기대비용 NEC에서의 비용비율을 CR_3으로 표기하자. 여기에서 CR_1은 정상상태의 분포와 전체부도율의 크기와 독립이며 항상 1이고, CR_2는 정상상태의 분포와는 독립이지만 전체부도율의 크기에만 의존하여 $\gamma/(1-\gamma)$값을 갖는다. 그리고 TR를 최대화하는 분류점에 대응하는 정규화된 기대비용함수가 TA에 대응하는 정규화된 기대비용함수보다 작거나 같을 경우에 비용비율의 상한값은 정상상태의 분포와 전체부도율의 크기에 의존한다. 그러므로 표 2와 3에 CR_1을 제외한 CR_2와 CR_3만을 나타내고 그림 2에서는 모든 값을 표현하였다.

우선 표 2와 그림 2의 왼쪽 그래프를 바탕으로 정상상태 확률밀도함수의 평균 변화에 따른 비용비율을 살펴보면, CR_3 값은 CR_1과 CR_2 사이에 위치하며 전체부도율 γ가 0.5로 증가할 때 모든 값은 1에 수렴한다. 그리고 값은 정상상태의 평균 μ_n에 의존하지만 그 값의 변화는 매우 적어 소숫점 3자리 이하에서만의 변화만큼 증가한다.

다음으로 표 3과 그림 2의 오른쪽 그래프를 살펴보면, 전체부도율 γ가 0.5로 증가할 때 역시 CR_3 값은 CR_1과 CR_2 사이에 위치하며 모든 값은 1에 수렴한다. 그러나 정상상태의 분산 $\sigma^2{}_n$이 증가할수록 CR_3 값은 증가한다. 그리고 정상상태 분포의 평균이 증가함에 따라 CR_3의 증가는 매우 적으나 분산이 증가하면 CR_3는 큰 값을 갖는다. 그러나 전체부도율이 클수록 그 차이는 작고 모든 CR_3는 1로 수렴하는 것을 파악할 수 있다.

5. 결론

부도와 정상상태의 확률밀도함수들의 혼합분포에서 최소 기대비용을 고려한 분류점은 비용-의존학습 방법인 기대비용함수 측면에서 최적이다. 과거자료를 통해 미래에 발생할 손실을 최소화하기 위한 분류점을 찾을 때 비용비율을 추정하는 문제는 특히 경기변동에 따른 부도의 심각도가 변하므로 이론적으로 접근하기 어려운 부분이 많다. 본 연구에서는 신용평가모형에서 비용에 관한 어떠한 정보가 주어지지 않은 경우에 비용-비의존학습 방법을 사용하여 분류정확도 측도의 성과를 최대화할 때의 비용비율을 유도하여 새로운 비용비율을 추정하는 방법을 제안하였다.

두 종류의 분류정확도 측도인 전체정확도와 진실율이 최대일 때의 분류점에 대응하는 기대비용에서의 비용비율을 제안하였으며, 분류정확도를 최대화하는 이론과 결합하여 기대비용을 최소화하는 비용비율과의 관계를 유도하였다. 전체정확도인 분류정확도가 최대일 때의 분류점에 대응하는 기대비용에서의 비용비율은 1이며, 진실율이 최대일 때의 비용비율은 전체부도율 γ에 의존하며 그 값은 $\gamma/(1-\gamma)$이다.

그리고 두 종류의 분류정확도가 최대일 때의 분류점에 대응하는 정규화된 기대비용을 비교하여 기대비용이 적은 확률비용함수의 구간에서 용비율을 결정할 수 있는 기준을 제안하였다. 정규화된 기대비용을 기준으로 비용비율의 상한값을 유도하였고 이 비용비율의 상한값은 분류정확도에 대응하는 기대비용에 대한 비용비율 사이 $(\gamma/(1-\gamma), 1)$에 존재하며 전체부도율이 0.5로 증가할수록 모든 비율이 수렴한다는 것을 증명하였다. 또한 정규화된 기대비용에 대한 비용비율은 비용-의존학습 방법에 의한 최소 기대비용에 대응하는 비용비율에 수렴하는 것을 보였다.

본 연구는 비용-의존학습 방법을 사용한 기대비용을 최소화할 때의 비용비율과 비용-비의존학습 방법을 사용하여 분류정확도 통계량이 최대일 때의 기대비용에서의 비용비율을 제안하고 그 비용비율들의 관계를 유도하였다. 그러므로 비용-의존학습 방법과 비용-비의존학습 방법의 관계를 설명함으로서로 다른 학습방법으로의 두가지 목적에 부합할 수 있음을 보였다.

참고 문헌

김지현 (2004). ROC and cost graphs for general cost matrix where correct classifications incur non-zero costs, <한국통계학회논문집>, **11**, 21–30.

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

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

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 classication, *In Proceedings of the 6th IEEE International Conference on Data Mining(ICDM'06)*, 33–42.

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-Kortum, R. and Follen, M. (1999). A comparison of C/B 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, *In Proceedings of the 23rd International Conference on Machine Learning(ICML'06)*, 233–240.

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

Fawcett, T. (2006). ROC graphs with instance-varying costs, *Pattern Recognition Letters archive*, **27**, 882–891.

Fielding, A. H. and Bell, J. F. (1997). A review of methods for the measurement of prediction errors in conservation presence/absence models, *Environmental Conservation*, **24**, 38–49.

Hand, D. J. (2009). Mismatched models, wrong results, and dreadful decisions: on choosing appropriate data mining tools, *In Proceedings of the 15th 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.

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.

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

Kaivanto, K. (2008). Maximization of the sum of sensitivity and specificity as a diagnostic cutpoint criterion, *Journal of Clinical Epidemiology*, **61**, 516–518.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*, Champman & Hall/CRC, Boca Raton, Florida.

Liu, Y. (2002). The evaluation of classification models for credit scoring, *Arbeitsbericht*, **2**, 1–65.

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

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*, 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*, 43–48.

Servigny, A. D. and Renault, O. (2004). *Measuring and Managing Credit Risk*, McGraw-Hill, New York.

Tasche, D. (2006). Validation of Internal Rating Systems and PD Estimates, *arXiv.org, eprint arXiv: physics/0606071*.

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., Ritchie, 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, *Metodoloski Zvezki*, **3**, 89–108.

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

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic(ROC) plots: A fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, **39**, 561–577.

2010년 7월 접수; 2010년 10월 채택

Cost Ratios for Cost and ROC Curves

Chong Sun Hong[1,a], Hyun Sang Yoo[a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

For classification problems on mixture distribution, a threshold based on cost functions is optimal from the viewpoint of a minimum expected cost. Assuming that there is no cost information, we propose cost ratios in the expected cost corresponding to thresholds where the total accuracy and the true rate are maximized to explain the relation of these cost ratios minimizing the expected cost. Other cost ratios are also proposed by comparing the normalized expected costs when classification accuracy is maximized. The values of these cost ratios are located between two cost ratios for the expected costs based on classification accuracies, and converge to that of the minimum expected cost. This work suggests two cost ratios: one is minimized by the expected cost and the normalized expected cost, and the other in the expected cost and the normalized expected cost functions that are maximized classification accuracies. We discuss their compatibility based on the relation of these cost ratios.

Keywords: Classification accuracy, credit evaluation, default, expected cost, discriminant power, threshold.

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