The Korean Communications
in Statistics Vol. 11 No. 1, 2004
pp. 21-30

ROC and Cost Graphs for General Cost Matrix Where Correct Classifications Incur Non-zero Costs[1)]

Ji-Hyun Kim[2)]

Abstract

Often the accuracy is not adequate as a performance measure of classifiers when costs are different for different prediction errors. ROC and cost graphs can be used in such case to compare and identify cost-sensitive classifiers. We extend ROC and cost graphs so that they can be used when more general cost matrix is given, where not only misclassifications but correct classifications also incur penalties.

Keywords : cost-sensitive learning, cost matrix, iso-performance line, CART

1. 서론

분류문제(classification problem)에 지도학습 알고리즘(supervised learning algorithm)을 적용할 때 예측의 정확도(accuracy)에만 관심을 두는 경우가 많다. 정확도를 분류기(classifier)의 성능평가기준으로 쓸 때에는, 모든 잘못된 예측은 같은 비용을 초래하며 정확한 예측에는 비용이 발생하지 않는다는 것을 가정하고 있다. 하지만 이런 가정이 만족되지 않는 예가 많이 있다. 은행대출 신청자에게 대출여부를 결정할 때 불량고객에게 대출을 허용했을 때와 우량고객에게 대출을 허용하지 않았을 때 발생하는 비용은 같지 않으며, 질병여부를 진단하는 문제에서 환자를 정상인으로 진단할 때와 정상인을 환자로 진단할 때 발생하는 비용 또한 같지 않을 것이다. 비용민감학습(cost-sensitive learning)은 예측의 정확도에만 의존하지 않고 비용까지 고려하는 학습으로서, 주어진 관측개체가 어떤 계급에 속할 확률이 비록 낮더라도 비용을 고려하여 그 계급으로 예측할 수 있는 분류기(classifier)를 생성한다. 따라서 비용민감학습에서는 비용행렬(cost matrix)의 정의가 중요하다.

두 계급 분류문제에서 일반적인 비용행렬의 형태는 [표 1]과 같다. 흔히, 제대로 분류하면 비용이 발생하지 않는 것으로 보고 $c_{00} = c_{11} = 0$ 으로 둔다. Elkan (2001)은 비용행렬에서 각 칸의 비용을 정할 때 같은 기준

1) This work was supported by the Soongsil University Research Fund.

2) Professor, Department of Statistics, Soongsil University, Dongjak-Ku Sangdo-Dong, Seoul 156-743, Korea.
E-mail: jhkim@stat.ssu.ac.kr

선(baseline)에서 생각하지 않는 오류를 흔히 범한다고 지적하며, 은행대출 신청고객을 분류하는 문제의 예를 들어, $c_{00}=c_{11}=0$ 이면 논리적 모순이 생긴다는 것을 설득력 있게 논하였다. 그의 설명대로가 아니더라도, 대출금을 상환하는 실제 우량고객을 우량고객으로 제대로 예측할 때 발생하는 비용과, 대출금을 갚지 않는 실제 불량고객을 불량고객으로 제대로 예측할 때 발생하는 비용은, 은행에 생기는 이익이라는 관점에서 보면 같지 않다. 우량고객에게 대출해주면 대출이자를 받게 되어 이익이 발생하지만 불량고객에게 대출을 기각했다고 해서 이익이 발생하지는 않기 때문이다. 우량고객을 음의 계급, 불량고객을 양의 계급이라고 하고 이익을 비용의 음수값이라고 하면, 이럴 때 $c_{00}<c_{11}=0$ 이어야 더 합리적이다. (비용행렬의 각 칸의 값에 같은 상수를 곱하거나 더해도 최적의사결정(optimal decision)에는 영향을 미치지 않으므로(Elkan, 2001), $c_{00}=0<c_{11}$ 으로 바꾸어 생각할 수 있다. 그리고 Elkan (2001)은 비용행렬에서 당연히 만족되어야 할 조건(reasonableness conditions)으로서 $c_{10}>c_{00}$ 와 $c_{01}>c_{11}$ 을 제시하였는데, 본 논문에서도 이 조건은 만족되는 것으로 가정한다.)

분류기의 성능을, 비용을 고려하여 비교하기 위해 Provost & Fawcett (1997)과 Drummond & Holte (2000)는 ROC 그래프와 비용 그래프를 각각 제안하였다. 이들 연구에서는 $c_{00}=c_{11}=0$ 를 가정하고 있는데, 본 논문에서는 $c_{00}=c_{11}=0$ 이라는 제약을 두지 않을 때 이들 그래프를 어떻게 확장하여 적용할 수 있는지에 대해 연구하였다. ROC 그래프와 비용 그래프, 두 방법은 분류기의 성능을 효과적으로 비교할 수 있는 방법으로서 많이 쓰이고 있다. 따라서 보다 일반적인 비용행렬이 주어졌을 때 이들 그래프를 어떻게 확장할 수 있는가 하는 것도 해결되어야 할 중요한 주제라고 생각한다.

제2절에서 ROC 그래프와 비용 그래프에 대해 간단히 소개하고, 제3절에서 일반적인 비용행렬이 주어졌을 때 이들 그래프를 어떻게 확장할 수 있는지 그 방법을 제시하였다. 제4절에서 예제를 통해 구체적으로 보였다.

[표 1] 일반적인 비용행렬

	실제 −	실제 +
− 로 예측	c_{00}	c_{01}
+ 로 예측	c_{10}	c_{11}

2. ROC 그래프와 비용 그래프

분류기의 성능을 평가하는 기준으로 흔히 오분류율(misclassification rate) 또는 정확도(accuracy)를 사용하는데, 이는 비용을 고려하지 않는 기준으로서 $c_{00}=c_{11}=0$ 이고 $c_{10}=c_{01}$ 임을 암묵적으로 가정하고 있다. 비용을 고려하여 분류기의 성능을 비교하기 위해 ROC 그래프(Receiver Operating Characteristic graphs)를 많이 쓴다(Egan, 1975). ROC 그래프를 간략히 설명하기 위해 아래 기호를 먼저 정의하자.

$$p_1 = P(Y=1),\ p_0 = 1-p_1 = P(Y=0)$$
$$TP = P(\hat{Y}=1|Y=1),\ FP = P(\hat{Y}=1|Y=0)$$

즉 관측개체가 양의 계급에 속할 확률이 p_1이며, 양의 계급에 속하는 관측개체를 제대로 예측할 확률을 TP(True Positive rate)라고 정의한다. ROC 그래프는 가로축이 FP(False Positive rate)이고 세로축이 TP인 ROC 공간(ROC space)에, 주어진 분류기의 (FP, TP)의 값을 점으로 나타낸 그래프이다. 여러 분류기의 성능을 비교하고 싶을 때 '등성능선'(iso-performance line)의 기울기를 이용하는데, Provost & Fawcett (1997)은 $c_{00}=c_{11}=0$ 일 때 등성능선의 기울기가 다음과 같음을 보였다.

$$\frac{p_0 c_{10}}{p_1 c_{01}} \tag{2.1}$$

등성능선의 기울기를 이용하여 여러 분류기의 성능을 비교하는 방법과 'ROC 최소볼록집합'(ROC convex hull)을 이용하여 효과적으로 잠재적 최적분류기(potentially optimal classifier)를 찾아내는 방법은 Provost & Fawcett (1997)에 잘 설명되어 있다. 본 논문의 4절에서 ROC 그래프의 예를 보였다.

Drummond & Holte (2000)는 ROC 그래프의 단점을 지적하며 새로운 방법을 제안하였다. ROC 그래프로는 한 분류기가 다른 분류기보다 얼마나 더 좋은 성능을 갖는지를 판단하기 어렵고, 주어진 분류기의 기대비용도 쉽게 읽어낼 수 없다는 점을 지적하며, 이러한 점을 보완하는 비용 그래프(cost graph)를 제안하였다. Drummond & Holte (2000)는 $c_{00}=c_{11}=0$ 을 가정하였는데, 그리는 방법을 간략히 정리해 본다. 비용 그래프의 세로축은 정규화기대비용(Normalized Expected Cost)으로서

$$NE(C) = \frac{p_1 c_{01}(1-TP) + p_0 c_{10} FP}{p_0 c_{10} + p_1 c_{01}} \tag{2.2}$$

인데, 기대비용(expected cost)을 0과 1사이의 값을 갖도록 표준화시킨 값이다. 가로축은 확률비용함수(probability cost function)라고 부르는 값으로서 다음과 같이 정의하였다.

$$PCF = \frac{p_1 c_{01}}{p_0 c_{10} + p_1 c_{01}}$$

만약 $c_{01}=c_{10}$ 이면 $PCF=p_1$ 이 되며, $p_0=p_1=1/2$ 이면 $PCF=c_{01}/(c_{01}+c_{10})$ 로서 두 오분류비용의 합에서 양의 계급을 오분류하는 비용 c_{01}이 차지하는 비율을 나타낸다. PCF를 이용하여 식 (2.2)를

$$NE(C) = (1-TP-FP)PCF + FP$$

와 같이 표현할 수 있다. ROC 공간의 한 점 (FP, TP)은 비용공간(cost space)에서 점 $(0, FP)$와 점 $(1, 1-TP)$를 잇는 선에 대응되며, 반대로 비용공간의 한 점은 ROC 공간의 한 선에 대응되어 두 그래프는 동등한 표현방법(dual representation)임을 보일 수 있다. 그리고 ROC 최소볼록집합에 대응하는 곡선을 비용공간에 그릴 수 있는데 이를 '아래덮개'(lower envelope)라고 부른다(Drummond & Holte, 2000). 비용 그래

프를 이용하면 두 분류기의 기대비용을 직접 비교할 수 있어 해석하기 쉽다는 장점이 있다. 본 논문의 4절에서 비용 그래프의 예를 보였다.

기존의 ROC 그래프와 비용 그래프의 연구에서는 $c_{00}=c_{11}=0$ 을 가정하였는데, c_{00}와 c_{11}이 반드시 0일 필요가 없는 일반적인 비용행렬이 주어졌을 때, ROC 그래프와 비용 그래프를 어떻게 확장할 수 있는지를 다음 절에서 알아보자.

3. 일반 비용행렬에 대한 ROC 그래프와 비용 그래프

먼저 c_{00}와 c_{11}이 반드시 0일 필요가 없는 일반적인 비용행렬이 주어졌을 때 비용 그래프를 그리는 방법을 유도하고자 한다. TP와 FP 값이 정해진 분류기의 기대비용은 아래와 같다.

$$\begin{aligned} E(C) &= p_1(1-TP)c_{01} + p_0c_{10}FP + p_1c_{11}TP + p_0(1-FP)c_{00} \\ &= p_1(1-TP)(c_{01}-c_{11}) + p_0FP(c_{10}-c_{00}) + p_1c_{11} + p_0c_{00} \end{aligned} \tag{3.1}$$

위 식으로부터 $TP=0$ 이고 $FP=1$ 인 최대기대비용을 갖는 분류기(the worst possible classifier)의 기대비용은

$$p_0c_{10} + p_1c_{01} \tag{3.2}$$

이며, $TP=1$ 이고 $FP=0$ 인 최소기대비용을 갖는 분류기(the best possible classifier)의 기대비용은

$$p_0c_{00} + p_1c_{11} \tag{3.3}$$

임을 알 수 있다. 위 극단적인 두 분류기의 기대비용의 차는

$$p_0(c_{10}-c_{00}) + p_1(c_{01}-c_{11}) \tag{3.4}$$

이 된다. 여기서

$$v_0 = p_0(c_{10}-c_{00}), \quad v_1 = p_1(c_{01}-c_{11})$$

$$PCF* = \frac{v_1}{v_0+v_1} \tag{3.5}$$

이라고 각각 정의하자. 만약, $c_{10}-c_{00}=c_{01}-c_{11}$ 이면 $PCF*=p_1$ 으로서 양의 계급일 확률이 되며, $p_0=p_1=1/2$ 이면 $PCF*=(c_{01}-c_{11})/(c_{10}-c_{00}+c_{01}-c_{11})$ 으로서 $c_{10}-c_{00}$ 과 $c_{01}-c_{11}$, 두 추가비용의 합에서 $c_{01}-c_{11}$, 즉 양의 계급을 제대로 분류할 때에 비해 잘못 분류할 때 드는 추가적인 비용이 차지하는 비율을 나타낸다. 한편 최소기대비용을 갖는 분류기를 기준으로 했을 때, 주어진 분류기에 의해 추가적으로 드는 비용의 기대값을 $E(AC)$라고 나타내기로 하면, 식 (3.3)에 의해

$$E(AC) = E(C) - (p_0c_{00} + p_1c_{11}) \tag{3.6}$$

이다. 식 (3.4)에 의해 $E(AC)$를 0과 1 사이의 값을 갖도록 표준화한

$$E(AC)/(v_0+v_1) \tag{3.7}$$

를 표준화기대추가비용(Normalized Expected Additional Cost)이라 부르기로 하자. 이 때 식 (3.1)과 (3.5), (3.6)을 이용하면

$$\begin{aligned} NE(AC) &= (1-TP)PCF*+FP(1-PCF*) \\ &= (1-TP-FP)PCF*+FP \end{aligned} \tag{3.8}$$

임을 쉽게 보일 수 있다. 따라서 분류기의 (FP, TP)가 주어지면, 즉 ROC 공간상의 점 하나가 주어지면, $PCF*$을 가로축으로 하고 $NE(AC)$를 세로축으로 하는 비용공간(cost space)에, 대응하는 선을 그을 수 있다. (정확히 얘기하면 비용공간이 아니라 추가비용공간(additional cost space)이라고 불러야 하지만 추가비용도 비용이므로 편의상 비용공간으로 부르기로 한다.) 이상의 결과를 Drummnod & Holte (2000)의 연구결과와 비교했을 때, 기대비용($E(C)$) 대신에 기대추가비용($E(AC)$)을 고려했다는 점과 확률비용함수(PCF)의 정의가 조금 달라진 점에 차이가 있다.

다음으로 c_{00}와 c_{11}이 반드시 0일 필요가 없는 일반적인 비용행렬이 주어졌을 때 ROC 그래프가 어떻게 달라지는지 알아보자. ROC 그래프의 가로축과 세로축은 각각 FP와 TP이므로 변화가 없고, 단지 분류기의 성능을 비교할 때 쓰는 등성능선의 기울기의 정의만 달라진다. ROC 공간의 두 점 (FP_1, TP_1), (FP_2, TP_2)의 기대비용이 같다면

$$p_1(1-TP_1)(c_{01}-c_{11})+p_0FP_1(c_{10}-c_{00})=p_1(1-TP_2)(c_{01}-c_{11})+p_0FP_2(c_{10}-c_{00})$$

이고, 따라서 등성능선의 기울기는 식 (2.1)과 달리

$$\frac{TP_2-TP_1}{FP_2-FP_1}=\frac{p_0(c_{10}-c_{00})}{p_1(c_{01}-c_{11})} \tag{3.9}$$

로 새롭게 정의된다.

지금까지의 논의에서와 같이 $c_{00}=c_{11}=0$ 의 조건이 반드시 만족될 필요가 없는 경우 ROC 그래프와 비용그래프는 지적한 대로 달라지지만, ROC공간과 비용공간의 동등성은 그대로 유지된다. 예를 들어 ROC 공간의 점이 비용공간의 선에 대응되며, ROC 최소볼록집합에 대응해서 비용공간에 아래덮개를 그릴 수 있는 등 (Drummond & Holte, 2000), 두 표현방법의 동등성은 그대로 유지된다.

4. 실제 자료의 예

캘리포니아 주립대학 자료저장소(UCI Repository, Blake and Merz (1998))에 있는 독일 신용자료를 이용해 앞 절에서 설명한 ROC 그래프와 비용 그래프를 그려 봄으로써, 이 두 그래프를 기존의 그래프와 비교했을 때

어떤 차이점이 있는가를 구체적으로 알아본다. 또한 이 두 그래프를 이용하여 최적 분류기를 찾는 방법에 대해서도 알아본다.

4.1 비용행렬

이 자료는 은행대출 신청자 1,000명에 대한 자료인데, 양의 계급, 즉 대출금을 상환하지 않는 불량고객에 대한 자료가 전체의 30%인 300개이다. 17개의 문자형 설명변수와 3개의 연속형 설명변수가 있다. 자료출처에 보면 자료와 함께 다음과 같은 비용행렬이 주어진다.

	실제 양호	실제 불량
양호로 예측	0	5
불량으로 예측	1	0

Elkan (2001)은 이 비용행렬이 같은 기준선에서 정해지지 않아 논리적 모순이 생긴다고 주장하였는데, 본 논문에서는 [표 2]와 같은 비용행렬을 고려하기로 한다. 여기서 비용행렬의 타당성에 대한 논의는 중요하지 않으며, 단지 $c_{00}=c_{11}=0$ 이 만족되지 않을 때 ROC 그래프와 비용 그래프를 어떻게 그리는가를 보이고자 한다. 참고로 $c_{10}-c_{00}$ 와 $c_{01}-c_{11}$ 의 값은 두 비용행렬에서 모두 같다.

[표 2] 독일 신용자료의 비용행렬

	실제 양호	실제 불량
양호로 예측	0	6
불량으로 예측	1	1

4.2 세 가지 분류기

학습 알고리즘으로 나무모형(Breiman et al. 1984)을 이용하였는데, 공개된 통계분석용 프로그래밍 언어인 R의 rpart package(Therneau & Atkinson, 1997)를 이용하였다. 분리기준(splitting criteria)과 정지규칙(stopping rule) 등은 자동설정값을 이용하였고 가지치기(pruning)는 하지 않았다.

두 계급의 분류문제에서 한 계급의 관측개체수가 다른 계급에 비해 현저하게 적을 때가 흔히 있다(관측개체수가 적은 계급을 양의 계급으로 두고 소수계급이라고 부르기로 한다). 이러한 자료를 분석할 때, 흔히 현장에서는 과대표집(over-sampling)이나 과소표집(under-sampling)을 통해 인위적으로 두 계급의 관측개체수가 균형을 이루도록 한 다음 분석하기도 한다. Elkan (2001)은, 비용행렬을 알고 있다면 무조건 균형을 맞추어 주는 대신에 비용을 고려하여 다수계급(음의 계급)의 크기를 $(c_{10}-c_{00})/(c_{01}-c_{11})$배 만큼 과소표집할 것을 제안하였다(Elkan (2001), 정리 1). 본 논문에서는 자료의 크기를 변화시키는 과소표집 대신에 다수계급에 속하는 관측개체에 대한 가중값(weight)을 바꾸어주는 방법을 택하였다.

세 가지 방법으로 훈련자료(train data)를 구성하여 나무모형을 각각 적용하였는데, 먼저 자료를 그대로 두는 방법(A)과 다수계급에 속하는 관측개체의 가중값을 1에서 3/7으로 낮추어 다수계급의 가중값의 합과 소수계급의 가중값의 합이 같도록 하는 방법(B), 그리고 Elkan (2001)의 제안대로 다수계급의 가중값을 1에서 $(c_{10}-c_{00})/(c_{01}-c_{11})=1/5$ 로 낮추어 주는 방법(C)을 각각 적용하였다. 분류기의 성능을 평가하기 위해 10중 교차타당성(10-fold cross-validation)을 이용하였으며, 교차타당성의 안정성을 보기 위하여 랜덤순열(random permutation)로 자료의 순서를 바꾸어가며 100번 반복하였다.

세 방법에 의해 생성되는 분류기의 성능은 [표 3]과 같다. 100번 반복한 결과의 평균을 구하였으며 평균의 표준오차(standard error)를 괄호 안에 표시하였다. [표 3]을 보면, 두 계급의 크기가 불균형을 이루는 다른 자료에서와 같이, 다수계급의 가중값을 낮춰주면 전체 오분류율은 높아지지만 소수계급오분류율(전체 자료 중에서 소수계급을 다수계급으로 잘못 분류한 비율)은 낮아지며, 소수계급으로 예측할 조건부확률인 TP와 FP는 모두 높아진다는 것을 알 수 있다. 식 (3.1)을 이용하여 구한 기대비용의 기준에서는, 예상대로, 비용을 고려한 방법 C가 우월했다. 만약 주어진 비용행렬이 정확하지 않을 때에나 소수계급의 확률 p_1이 정확한 값이 아닐 때 분류기의 성능을 효과적으로 비교하기 위해 ROC 그래프와 비용 그래프를 이용하는데, 이 두 그래프를 그려본다.

4.3 ROC 그래프와 비용 그래프

[그림 1]은 ROC 그래프를 나타낸다. 세 분류기를 나타내는 점 A, B, C 이외에 점 $(0,0)$은 무조건 다수계급으로 예측하는 분류기를 나타내고 점 $(1,1)$은 이와 반대로 무조건 소수계급으로 예측하는 분류기를 나타내는데, 이 두 분류기를 '자명한 분류기'(trivial classifiers)라고 한다. [그림 1]에서 이 다섯 개의 점을 잇는 곡선이 최소볼록집합(convex hull)이 되므로 다섯 개의 분류기가 모두 잠재적 최적분류기이다(Provost & Fawcett, 1997). 즉 각 분류기의 기대비용이 최소가 되는 사전확률(p_1, p_0)과 비용($c_{00}, c_{01}, c_{10}, c_{11}$)의 조합이 존재한다. 주어진 사전확률과 비용의 값으로 식 (3.9)에 의한 등성능선의 기울기를 구해보면 약 0.467이다. 이 기울기를 갖는 등성능선을 그려 판단해보면, A, B, C 세 개의 분류기 중에서는 C가 최적이며, 두 개의 자명한 분류기를 포함한 다섯 개의 분류기 중에서는 점 $(1,1)$에 대응하는 분류기가 최적임을 알 수 있다.

ROC 그래프를 이용하면 두 분류기의 우열을 가릴 수 있지만 성능에 얼마나 차이가 나는지를 쉽게 알 수 없다. 이를 보완한 것이 비용 그래프이다([그림 2]). 식 (3.8)에 의해 $PCF*=0$ 이면 $NE(AC)=FP$ 이고, $PCF*=1$ 이면 $NE(AC)=1-TP$ 이다. 이를 이용하여 ROC 그래프 위의 다섯 점에 대응하는 선을 비용 그래프에 각각 그릴 수 있다. 주어진 사전확률과 비용의 값에서 $PCF*$의 값은 약 0.682이다. 이 값에서 $NE(AC)$가 최소가 되는 분류기는, ROC 그래프에서와 마찬가지로 A, B, C 세 개의 분류기 중에서는 C가, 두 개의 자명한 분류기를 포함한 다섯 개의 분류기 중에서는 점 $(1,1)$에 대응하는 분류기임을 [그림 2]에서 알 수 있다. ROC 그래프의 결과와 비용 그래프의 결과가 일치하는 것은 일반적으로 성립하는 사실인데, 그 이유는 두 분류기의 $NE(AC)$의 차는 식 (3.6)과 (3.7)에 의해 두 분류기의 $E(C)$의 차의 상수곱과 같기 때문이다. 비용그래프에서는 주어진 확률비용함수($PCF*$)의 값에서 세로축의 높이 차가 바로 두 분류기의 비용

(엄격하게는 정규화기대추가비용) 차를 나타내므로 ROC 그래프에서보다 더 쉽게 비교할 수 있다는 장점이 있다.

[그림 2]에서 보면 PCF^*가 약 0.5보다 작은 영역에서 분류기 A의 성능이 더 좋을 수 있다. 하지만 소수계급의 오분류비용이 다수계급의 오분류비용보다 높을 때, 보다 엄밀하게는

$$c_{01} - c_{11} > (p_0/p_1)(c_{10} - c_{00})$$

일 때 $PCF^* > 0.5$ 이므로, 기대비용의 관점에서 볼 때 분류기 A는 실제로 좋은 분류기가 되기 힘들다. 한편 $PCF^* = 0.682$ 에서 분류기 C보다, 자명한 분류기 $(1,1)$의 성능이 더 우수하게 나오는 이유는 $p_1 = 0.3$ 이 실제 분포를 반영하지 못하기 때문일 것으로 판단된다. 독일 신용자료를 만들 때 인위적으로 불량고객의 비율을 높여서 구성한 것으로 짐작되며, 실제 불량고객의 비율은 30%보다 작은 값일 것이다. 따라서 실제 적용할 때 p_1의 값이 더 작아져 PCF^*의 값은 0.5에 더 가까워지고 이 때 최적분류기는 분류기 C가 될 가능성이 높다. ('가능성이 높다'라는 유보적인 표현을 쓴 이유는 검증자료(test data)에서 p_1이 작아지면 PCF^*뿐만 아니라 분류기의 FP와 TP도 달라지기 때문이다.) 이 사실을 모의실험을 통해 확인해볼 수 있겠으나, 본 절의 목적이 최적분류기를 찾는 데에 있지 않고, 확장된 ROC 그래프와 비용 그래프의 예를 보이는 데에 있으므로 생략하기로 한다.

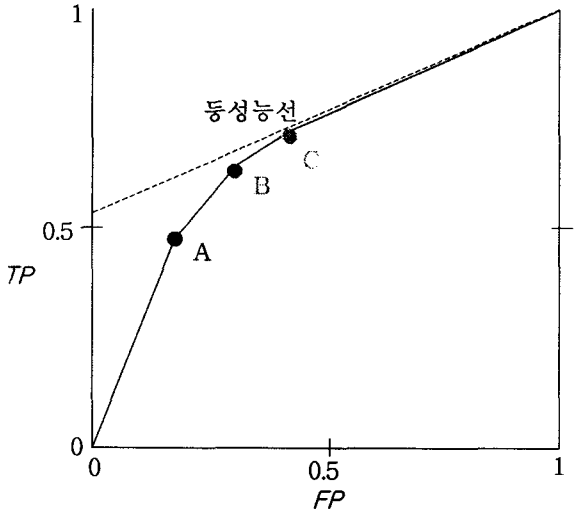
[표 3] 세 가지 훈련자료 구성방법에 따른 세 분류기의 성능

	오분류율	소수계급오분류율	TP	FP	$E(C)$
방법 A	.285 (.0012)	.160 (.0008)	.468 (.0025)	.180 (.0013)	1.225 (.0033)
방법 B	.340 (.0011)	.122 (.0008)	.592 (.0027)	.311 (.0015)	1.128 (.0039)
방법 C	.390 (.0013)	.092 (.0007)	.695 (.0024)	.426 (.0020)	1.055 (.0037)

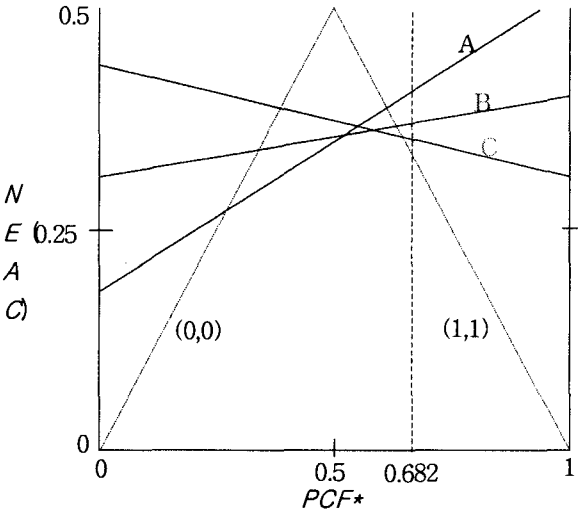
5. 결론

은행대출 신청고객의 분류 예에서 보듯이 제대로 예측했을 때 발생하는 비용인 c_{00}와 c_{11}이 0이 아닌 경우가 있다. 따라서 이러한 일반적인 비용행렬이 주어졌을 때 분류기의 성능을 비교하는 방법이 필요하다. 본 연구에서는 기존의 ROC 그래프와 비용 그래프를 확장하여 일반적인 비용행렬이 주어졌을 때에도 적용할 수 있도록 하였다. ROC 그래프의 경우에는 축의 변경은 필요하지 않지만, 최적분류기를 찾기 위해 필요한 등성능선의 기울기를 다시 정의해야 한다는 것을 보였다. 비용 그래프의 경우에는 가로축과 세로축을 다시 정의하여야 하며 해석도 그에 따라 달라짐을 보였다.

확장된 ROC 그래프와 비용 그래프를 적용한 예에서 나무모형에 의한 분류기를 이용하였는데, 어떤 분류기든 상관없이 그 성능을 비교할 때 이 두 그래프를 적용할 수 있다는 것을 밝혀둔다.



[그림 1] ROC 그래프



[그림 2] 비용 그래프

참고문헌

[1] Blake, C. L. and Merz, C. J. (1998). UCI Repository of machine learning databases, http://www.ics.uci.edu/~mlearn/ML-Repository.html, University of California in Irvine, Department of Information and Computer Science.

[2] Breiman, L., Friedman, J. H., Olshen, J. A., and Stone, C. J. (1984). *Classification and Regression Trees*, Belmont, CA, Wadsworth.

[3] 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.

[4] Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*, Series in Cognition and Perception, New York: Academic Press.

[5] Elkan, C. (2001). The foundations of cost-sensitive learning, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence.*

[6] 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.

[7] Therneau, T. M. and Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines, Technical Report, Mayo Foundation.

[2003년 9월 접수, 2004년 1월 채택]