아시아교육연구 12권 4호
Asian Journal of Education
2011, Vol. 12, No. 4, pp. 193-216.

진단, 선발, 분류에 대한 의사결정에 있어서 Bayes' Theorem의 적용 가능성 및 교육적 함의

박태학(朴泰學)*

논문 요약

본 연구는 임상적 진단이나 선발과 분류와 같은 의사결정에 있어서 '베이즈 정리'의 적용 방안을 모색하고 구체적인 적용 사례를 통해 그 교육적 함의를 논의하였다. 임상적 진단에 있어서 '잘못된 양성반응의 패러독스'는 매우 중요한 시사점을 제시한다. 모집단의 유병율이 낮은 조건에서 상당한 정확성을 가진 진단도구라고 할지라도 대부분의 양성결과는 잘못 판정된 것일 수 있다는 점이다. 대부분의 임상적 장애가 낮은 유병율을 보이기 때문에 임상적 평가에서 총평관적 접근의 중요성을 강조한다. 본 연구는 선발 혹은 분류에 있어서 분할점수 추정과 그에 따른 준거타당도 평가에 대해 베이즈 정리를 어떻게 적용할 수 있는지를 구체적인 적용 사례와 함께 제시하였다. 기존 방식과 달리 베이즈 정리에 의한 방식은 분할점수 추정에 있어서 통계적 오차를 반영한 의사결정을 가능하게 해준다. 또한 연속변수의 경우 기존방식은 무수히 많은 분할표를 필요로 하지만, 베이즈 정리에 의해 이러한 문제를 쉽게 해결할 수 있다. 특히 베이즈 방식은 ROC 곡선의 분석을 병행함으로써 최적분할모델을 손쉽게 식별할 수 있다. 결과적으로 임상적 진단, 선발과 분류에 관련한 준거설정과 준거타당도는 주기적(periodically)으로 평가되고 수정·보완되어야 한다. 베이즈 정리는 이러한 준거설정과 준거타당도 연구에 있어서 논리적으로 적합한 모델을 제시한다.

주요어: 베이즈 정리, 임상적 진단, 선발, 분류, 분할표, 준거설정, 준거타당도

Ⅰ. 서 론

우리는 일상의 교육현장에서 학생들에 대한 임상적 진단이나, 검사결과에 의거하여 특정한 집단을 선발하거나, 학생들을 여러 범주나 수준으로 분류 혹은 배치하는 상황을 흔히 접하게 된

* 신라대학교 교육학과 부교수

다. 여기서 여러 가지 검사자료나 준거에 입각하여 학생들을 진단, 선발, 분류하는 모든 의사결정의 경우 공통적으로 발생하는 문제점은 두 가지의 오류, 즉 긍정적 오류와 부정적 오류가 발생할 수 있다는 사실이다. 따라서 이러한 의사결정의 정확성을 높이기 위해서는 그 오류의 원인을 잘 파악하고, 그것을 최소화할 수 있는 방법을 모색해야 한다.

본 연구의 목적은 이러한 진단, 선발 및 분류에 관련한 준거설정과 그 준거에 따른 의사결정의 타당성 평가에 있어서 '베이즈 정리(Bayes' Theorem)'의 적용 가능성을 모색하는 것이다. 기존 방식은 이러한 의사결정에 있어서 여러 가지 본질적인 문제점과 한계를 지니고 있다. 첫째, 준거점수에 대한 정보만 제공하지 근처 점수들에 대한 오차에 대한 정보를 제공하지 못하며, 따라서 오차를 반영한 의사결정이나 오차를 고려한 분할점수 조정이 불가능하다. 둘째, 진단, 선발 혹은 분류에 대한 준거설정에 필요한 정보만 제공하지, 이러한 단순한 차원을 넘어 개별적 학생 상담에 필요한 부과적인 정보를 제공하지 못한다. 또한 기존방식은 검사점수가 연속변수인 경우 분할점수 추정을 위해 무한히 많은 분할표의 분석을 필요로 하며, 그 준거에 따른 의사결정의 정확성을 평가하는데 한계가 있다. 베이즈 정리에 의한 방식은 기존 방식의 이러한 문제점과 한계를 극복할 수 있는 부과적인 정보와 해결방법을 제공한다. 본 연구는 학생들에 대한 진단, 선발, 분류 상황에서 기존방식에 의해 발생할 수 있는 이러한 문제점과 한계를 극복할 수 있는 방안으로 베이즈 정리의 적용을 통해 그 구체적 적용 사례와 해결방법 등을 종합적으로 제시하였다.

Ⅱ. Bayes' Theorem의 이해

베이즈 정리는 영국의 수학자이자 목사인 Thomas Bayes(1701-1761)가 "우연이라는 원칙으로 문제를 해결하는 방법에 관한 수필(Essay towards solving a problem in the doctrine of chances)"이라는 제목으로 발표한 이론이다. 베이즈 정리는 두 확률사건(random events)에 있어 조건확률(conditional probability)과 주변확률(marginal probability) 사이의 관계성을 기술한 것이다. 일반적으로 기초통계학에서는 베이즈 정리를 확률사건에 한정하여 설명하지만 확률변수의 경우에도 쉽게 적용이 가능하다.

1. 확률사건(Random Events)

두 임의의 확률사건 A와 B가 있다고 가정하자. 또한 A는 n개의 서로 배반인 범주 $A_1, A_2, \cdots, A_N$으로 분할되고, 그 중 하나의 범주는 반드시 일어난다고 가정하자. 사건 B와 관련하여 A의 조건부 확률에 대한 베이즈 정리는 다음과 같이 정의한다.

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{P(B)} \quad (1)$$

$P(A|B)$는 관측된 B에 대한 A의 조건부 확률이며, 이것은 B의 관찰 결과에 의해 결정되기 때문에 사후확률(posterior probability)이라고 불린다. $P(A)$는 A의 주변확률로서 사전확률(prior probability)이라 불리며, '사전'이라는 것은 아직 사건 B에 관한 어떤 정보도 고려하지 않음을 의미한다. $P(B|A)$는 사전정보 A가 주어졌을 때 B가 일어날 조건부 확률이며, 우도(likelihood) 혹은 우도함수(likelihood function)로 불린다. $P(B)$는 정규화 상수(normalizing constant)의 역할을 하는 것으로 아래와 같이 구할 수 있다.

$$P(B)=\sum_{j=1}^{N}P(B|A_j)P(A_j) \quad (2)$$

사건 A가 단순히 두 범주 A와 그 여집합 $\overline{A}$로 양분되는 경우에 기초통계학에서 흔히 제시되는 베이즈 정리의 가장 단순하고도 기본적인 공식이 정리된다.

$$P(A\mid B)=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\overline{A})P(\overline{A})} \quad (3)$$

이와 같이 두 배반 범주의 경우 베이즈 정리를 승산비와 우도비(odds-likelihood ratio)의 형태로 정의하기도 한다.

$$R=\frac{P(A|B)}{P(\overline{A}|B)}=\frac{P(B|A)}{P(B|\overline{A})}\times\frac{P(A)}{P(\overline{A})} \quad (4)$$

사후승산비(혹은 승산비) R은 A와 $\overline{A}$에 대한 사후확률의 비율로서, A가 $\overline{A}$에 비해 선호되는 비율이며, 베이즈 인자(Bayes factor)인 우도의 비율에 사전승산비(prior odds)를 곱한 것과 같다. 이러한 형태의 베이즈 정리는 두 배반 가설이나 이론 중에서 하나의 가설이나 이론을 선택하는 의사결정에 유용하며, 특히 이것은 베이지안 가설검정의 근간을 이루는 개념이다.

베이즈 정리의 요점은 새로운 정보 B에 의해 이를 근거로 어떤 사건 A가 발생할 조건부 확률을 개선하는 방법을 제공한 것이다. 이는 특히 사전확률 $P(A)$가 밝혀지지 않았을 때 이를 확인할 수 있다는 점에서 중요성이 있으나, 사전확률이 밝혀져 있다고 하더라도 이를 새로운 정보

B에 의해 조정할 수 있다는 의미가 더 크다. 다시 말해 베이즈 정리는 A에 대한 사전 지식이나 개인적인 확률적 신념이 B에 대한 새로운 정보에 근거하여 개선되어지는 과정을 묘사한 것이다.

2. 확률변수(Random Variables)

베이즈 정리는 확률사건을 대신하여 확률변수를 사용하더라도 거의 변화가 없다. 확률변수에는 두 가지의 종류가 있다. 하나는 모든 값을 셀 수 있는 이산확률변수(discrete random variable)이고, 다른 하나는 어떤 구간 내에서 모든 값을 취할 수 있는 연속확률변수(continuous random variable)이다. 임의의 두 변수 X와 Y가 연속변수인 경우, 베이즈 정리는 확률밀도함수(probability density function: pdf)의 조건부 확률에 의해 아래와 같이 정의될 수 있다.

$$f_X(x|Y=y)=\frac{f_Y(y|X=x)f_X(x)}{f_Y(y)} \tag{5}$$

여기서 $f_X(x)$와 $f_Y(y)$는 각각 X와 Y의 확률밀도함수이며, 음이 아닌 함수이다. X 값의 특정한 구간 $(a \le X \le b)$에 대한 확률을 확률밀도함수로 표현하면 아래와 같고,

$$f_X(a \le X \le b)=\int_a^b f_X(x)dx \tag{6}$$

X 값의 범위가 $-\infty$ 에서 ∞ 이면, $\int_{-\infty}^{\infty} f_X(x)dx=1$이다. X 값의 범위가 $-\infty$ 에서 ∞ 이면, $\int_{-\infty}^{\infty} f_X(x)dx=1$이다. $f_Y(y)$는 정규화 상수의 역할을 하는 것으로 X의 특정한 값 x에 대한 Y의 조건부 확률밀도함수로 나타낼 수 있으며, 만약 Y 값의 범위가 $+\infty$ 에서 $-\infty$ 이면 아래와 같다.

$$f_Y(y)=\int_{-\infty}^{+\infty} f_Y(y|X=x)f_X(x)dx \tag{7}$$

베이즈 정리를 적용함에 있어서 때때로 한 변수가 연속적 변수이고 다른 변수가 이산형 변수인 경우를 흔히 접하게 된다. 만약에 X가 이산형 변수이고, N개의 서로 배반인 범주

$x_1, x_2, \cdots, x_N$으로 분할되며, $P(X=x_i)$의 사전확률을 갖는다고 가정하자. 또한 임의의 Y변수가 $f_Y(y)$의 확률밀도함수를 갖는다면, 두 변수에 대한 베이즈 정리는 아래와 같이 정의할 수 있다(Johnsonbaugh & Jost, 1996).

$$P(X=x_i|\,Y=y)=\frac{f_Y(y|X=x_i)P(X=x_i)}{f_Y(y)} \tag{8}$$

여기서 모든 x_i에 대해 $0 \leqq P(X=x_i) \leqq 1$이고, $\sum_{i=1}^{N} P(X=x_i)=1$이다. 분자에 있는 $f_Y(y|X=x_i)$는 X가 x_i인 경우 Y의 조건부 확률밀도이며, 분모 $f_Y(y)$는 정규화 상수로서 아래와 같다.

$$f_Y(y)=\sum_{j=1}^{N} f_Y(y|X=x_i)P(X=x_i) \tag{9}$$

Ⅲ. 임상적 진단과 거짓 양성반응 패러독스

1. 임상적 진단과 그 타당성

초·중등학교 현장에서는 아동과 청소년의 발달과 관련하여 정신장애, 학습장애 등 임상적 진단을 필요로 하는 문제가 많이 발생한다. 하지만 어떤 임상적 진단이든 완벽한 판정을 기대할 수 없다. 모든 임상적 진단이 공통적으로 지니는 문제는 두 가지의 오류가 항상 수반된다는 것이다. 임상적 진단법에 있어서 정확도를 높이기 위해서는 판단의 오류가 어떻게 일어나는지와 그것을 최소화할 수 있는 방법에 대한 이해가 필요하다. 예를 들어 어떤 정신질환 여부에 대한 적격성 판정을 진단검사의 결과에 의존하여 결정할 경우, 두 가지의 올바른 판정과 두 가지의 판정 오류가 동시에 발생할 수 있다(참조: <표 1>). 첫째, 정신질환자를 올바르게 양성으로 판정하는 경우(진양성)이며; 둘째, 비정신질환자를 음성으로 올바르게 판정하는 경우(진음성)이고; 셋째, 비정신질환자를 양성으로 잘못 판정하는 경우(위양성)로서 제1종 오류의 개념이고; 마지막으로 정신질환자를 음성으로 잘못 판정하는 경우(위음성)로서 제2종 오류의 개념이 그것이다.

<표 1> 진단법에서의 올바른 판정과 판정오류의 유형

		실제		
		정신질환자	비질환자	계
검사 결과	양성	진양성 a	위양성 b	a+b
	음성	위음성 c	진음성 d	c+d
	계	a+c	b+d	a+b+c+d

베이즈 정리를 적용하여 진단법의 타당성 여부를 평가하기 위해 필요한 정보를 아래와 같이 요약할 수 있다.

$$\text{민감도(sensitivity)} = \frac{a}{a+c} \qquad \text{위음성율(FNR)} = \frac{c}{a+c}$$

$$\text{특이도(specificity)} = \frac{d}{b+d} \qquad \text{위양성율(FPR)} = \frac{b}{b+d}$$

$$\text{양성예측도(PPV)} = \frac{a}{a+b} \qquad \text{음성예측도(NPV)} = \frac{d}{c+d}$$

$$\text{양성승산비(R+)} = \frac{\text{민감도}}{1-\text{특이도}} \qquad \text{음성승산비(R-)} = \frac{1-\text{민감도}}{\text{특이도}}$$

[그림 1] 베이즈 정리에 의한 진단법의 타당성 평가에 관련된 정보

이러한 정보를 근거로 하여 진단법의 타당성을 평가하는데 있어서 최소한 세 가지 측면을 고려해야만 한다. 첫째, 바람직한 진단법이라면, 민감도와 특이도가 높아야 할 것이다. 민감도(sensitivity)는 실제 정신질환이 있는 학생들이 검사를 통해 양성이라고 판단될 확률로서, 그 검사가 정신질환자를 얼마나 잘 선별해 내는가를 알려주는 지표이다. 또한 민감도는 정신질환이 있는 데도 검사를 통해 음성이라고 판단될 확률인 위음성율(Flase Negative Rate: FNR)이 얼마나 적은지를 보여준다(참조: 위음성율 = 1-민감도). 특이도는 실제 정신질환이 없는 학생들이 검사를 통해 음성이라고 판단될 확률로서 이 검사가 비정신질환자를 얼마나 잘 배제시키는가를 나타내는 지표이다. 동시에 특이도는 정신질환이 없는 데도 검사를 통해 양성이라고 판단될 확률인 위양성율(False Negative Rate: FNR)이 얼마나 적은지를 보여준다(참조: 위양성율 = 1 - 특이도). 민감도가 매우 높은 경우, 검사 결과가 음성이라면 정신질환 판정으로부터 배제(rule-out)를 고려해 볼 수 있다. 반면 특이도가 매우 높은 경우, 검사 결과가 양성이라면 정신질환 확정

(rule-in)을 고려볼 수 있다. 그러나 진단법의 타당성에 있어서 민감도와 특이도의 정확성은 필요조건이지 충분조건은 되지 못한다.

둘째, 진단법의 타당성 평가에 있어서 가장 중요한 판단 준거는 양성예측도이다. 양성예측도(Positive predictive value: PPV)는 양성반응이 나온 학생이 실제 정신질환이 있을 확률을 나타내는 지표이기 때문이다. 사전확률이 유병율(prevalence)이라면, 양성예측도는 베이즈 정리의 사후확률과 같다. 진단법이 최소한의 유용성을 갖기 위해서는 양성예측도가 무작위 판정의 수준인 0.5 보다 높아야 한다. 같은 맥락에서 양성승산비($R+$)는 최소한 1.0 보다 커야 한다. 양성승산비는 비정신질환자가 검사결과 양성이 나올 가능성에 비해 정신질환자가 검사결과 양성으로 나올 가능성이 얼마나 더 큰지를 나타내는 승산 지표이다. 무작위 판정의 경우 양성예측도의 기대치는 .5이고, 양성우도비의 기대치는 1.0일 것이다. 부과적인 정보로 음성예측도(Negative Predictive Rate: NPR)는 정신질환자 대신 비정신질환자를 선별하는 경우의 사후확률로서 높을수록 좋을 것이다. 또한 음성승산비($R-$)는 양성승산비의 역비율이며, 비정신질환자가 검사결과 음성으로 나올 가능성에 비해 정신질환자가 검사결과 음성으로 나올 가능성이 얼마나 더 큰지 알려주는 승산 지표이다. 따라서 음성승산비는 적을수록 좋을 것이다.

마지막으로 판정오류를 일으키는 주요한 원인의 하나인 기초비율(base rate), 즉 유병율을 고려해야 한다. 일반적으로 양성예측도는 모집단의 유병율에 의해 크게 달라지며, 특히 유병율이 낮은 경우 민감도와 특이도가 높아도 양성예측도는 현저히 낮아지게 된다. 따라서 모집단의 유병율이 아주 낮은 조건(rare conditions)에서는 정확도가 상당히 높은 검사도구라 할지라도 바람직한 양성예측도를 기대할 수 없다.

2. 가상적 자료에 의한 적용 사례

초등학교 학생들을 대상으로 정신질환 여부에 대한 적격성 판정을 내리기 위해 간편한 진단검사를 개발한다고 가정하자. 진단검사의 타당성 여부를 평가하기 위해 다음과 같은 소정의 절차가 필요하다. 먼저 과거의 통계치를 통해 전체 초등학교 학생들 중에 정신질환자로 적격성 판정된 학생들의 비율(사전확률)이 얼마인지를 조사해 볼 필요가 있고, 적절한 정신질환자 표본과 비정신질환자 표본을 대상으로 사전검사(pilot testing) 시행을 통해 최소한 두 가지 정보에 대한 조사가 필요하다. 하나는 이 검사가 정신질환자 표본으로부터 올바르게 양성반응을 판정해 내는 비율(민감도)과 비정신질환자 표본으로부터 잘못된 양성판정을 하는 비율(위양성율)에 대한 정보가 그것이다. 이들 정보를 근거로 베이즈 정리는 어떤 학생이 이 진단검사를 통해 양성으로 판정되었을 때 그 학생이 실제 정신질환자일 확률은 얼마인지에 대한 양성예측도를 제공해 준다. 정신질환자와 비정신질환자를 각각 A와 $\overline{A}$로 표기하고, 양성반응과 음성반응을 각각 B,

$\overline{B}$로 표기하여 아래와 같은 결과를 얻었다고 가정하자.

$P(A) = .001$	유병율(기초비율)
$P(B \mid A) = .99$	민감도(정신질환자에 대한 양성반응의 확률)
$P(B \mid \overline{A}) = .01$	위양성율(비정신질환자에 대한 양성반응의 확률)

이들 자료를 베이즈 정리의 공식(3)에 대입하면 아래와 같은 사후확률을 얻게 된다.

$$P(A|B) = \frac{.99 \times .001}{.99 \times .001 + .01 \times .999} = .09$$

양성예측도가 .05보다 적기 때문에 이 검사에 의한 진단법은 적합하지 않다. 위의 자료를 공식(4)에 대입하면 양성승산비는 .09가 되며, 이것은 1.0 보다 적으므로 같은 결론을 얻게 된다.

$$R = \frac{.99}{.01} \times \frac{.001}{.999} = .09$$

이러한 통계적 결과는 일반인의 직관에 매우 반하는 것이기 때문에 많은 사람에게 의외의 사실로 인식될 것이다. 이 검사는 정신질환자를 대상으로 민감도 99% 수준의 정확한 양성반응을 보였고, 비질환자를 대상으로 특이도 99% 수준의 올바른 음성을 보였다. 즉, 정신질환자 표본을 대상으로 단지 1% 수준의 위양성율을 보였다. 이런 수준의 민감도와 특이도를 갖는 검사는 일반인에게 직관적으로 매우 정확한 진단도구로 인지될 것이다. 그러나 이 검사에 의해 양성반응을 나타낸 학생이 실제 정신질환자일 가능성은 10% 미만이며, 양성으로 판정된 사람의 90% 이상은 실제 비정신질환자일 것이다. 이렇게 직관에 반대되는 모순된 결과를 두고 '거짓 양성반응 패러독스(false positive paradox)'라 한다. 이 패러독스는 양성반응 결과가 정확히 정신질환자일 확률은 진단도구의 정확성뿐만 아니라 표집하는 모집단의 특성에 의존한다는 것을 보여준다.

3. 거짓 양성반응 패러독스(False Positive Paradox)

거짓 양성반응(false positive)은 불행히도 모든 진단법이 안고 있는 문제점이다. 우리는 때때로 잘못된 양성반응에 의한 심각한 피해를 겪었던 사례를 언론을 통해 접하게 된다. 예를 들어 다소 희귀한 질병에 대해 양성으로 판정된 환자가 오랜 세월 동안 엄청난 정신적 피해와 육체적 고통을

겪고 난 이후 그 판정이 잘못된 것으로 밝혀지는 사례이다. 베이즈 정리가 우리에게 주는 중요한 시사점은 유병율이 매우 낮은 조건에서는 검사도구의 정확성, 즉 민감도와 특이도가 상당히 높은 경우라도 대부분의 양성결과가 잘못된 양성이라는 것이다. 이러한 문제점의 심각성은 임상적 진단을 필요로 하는 경우의 대부분이 유병율이 낮은 조건에서 이루어진다는 것이다.

다양한 모집단의 기초비율에 따라 양성예측도와 양성승산비가 얼마나 심각하게 달라지는지 보여주기 위해 아래 <표 2>의 자료를 살펴보자.

<표 2> 진단검사의 유병율에 따른 가상적 양성예측도와 양성승산비

유병율	민감도	위양성율	양성예측도	양성승산비
.001	.99	.01	.090	.10
.002	.99	.01	.166	.20
.003	.99	.01	.230	.30
.004	.99	.01	.284	.40
.005	.99	.01	.332	.50
.006	.99	.01	.374	.60
.007	.99	.01	.411	.70
.008	.99	.01	.444	.80
.009	.99	.01	.473	.90
.010	.99	.01	.500	1.00

위의 자료는 민감도 .99와 위양성율 .01(즉, 특이도 .99)의 정확성을 가진 검사도구의 양성예측도와 양성승산비가 유병율이 증가함에 따라 얼마나 민감하게 변화하는지 보여준다. 유병율이 .001일 때 양성예측도는 .09이고 양성승산비는 .10에 불과하다. 유병율이 .001에서 .002으로 증가함으로써 양성승산비가 2배로 증가함을 볼 수 있다. 유병율이 위양성율의 수준인 .01에 이르러서야 양성예측도와 양성승산비가 무작위 판정의 수준인 .5과 1.0이 된다. 위양성율이 모집단의 유병율 보다 높음에도 불구하고, 이러한 정보를 무시한 채, 단순히 양성반응 결과에 따라 정신질환자로 판정하는 것을 '기초비율오류(base rate fallacy)'라고도 한다. 이러한 결과는 모집단의 기초비율이 낮은 조건에서는 인간의 특성을 하나의 검사나 절차로 평가할 것이 아니라 여러 가지 다양한 검사와 가능한 모든 준거들을 동원하여 종합적으로 평가해야 한다는 사실 보여주는 것이다. 다시 말해, 거짓 양성반응 패러독스는 임상적 진단에 있어서 총평관적 접근의 중요성을 시사한다.

Ⅳ. 분할점수 설정과 준거타당도

1. 분할점수 설정

교육현장에서 검사결과에 의거하여 특정한 집단을 선발하거나, 학업성취의 목표도달-목표미달 여부를 결정하거나, 학생들을 여러 범주 혹은 수준으로 분류 혹은 배치하는 문제를 흔히 접하게 된다. 이러한 선발, 분류, 배치의 목적으로 개발한 검사를 준거참조검사라 하며, 준거참조검사를 개발하는 과정에는 분할점수(cut-off score)를 설정하는 절차와 그 준거의 타당성에 대한 연구가 필요하다.

어떤 검사결과를 근거로 하여 학생들을 긍정적 범주와 부정적 범주(즉, 성공-실패, 합격-불합격, 추천-비추천 등)로 분류하는 경우를 고려해 보자. 이러한 절차는 특정한 분할점수를 기준으로 하여 그 점수 이상을 긍정적 집단으로, 그 점수 미만을 부정적 집단으로 분류하는 과정을 포함한다. 이러한 분할점수에 근거한 의사결정은 임상적 진단에서의 경우와 마찬가지로 항상 두 가지의 올바른 판정과 두 가지의 판정오류를 수반하게 된다(참조: <표 3>). 의사결정의 정확성(decision-making accuracy)은 이러한 판정오류를 최소한으로 줄이고 올바른 판정의 비율을 최대로 늘이는 분할점수 선정에 달려 있다(백순근, 2007; Cizek & Bunch, 2007).

<표 3> 진단법에서의 올바른 판정과 판정오류의 유형

		실제 결과		
		성공	실패	계
검사에 의한 예측 결과	성공	긍정적 적중 (+적중) a	긍정적 오류 (爲긍정) b	(선발) a+b
	실패	부정적 오류 (爲부정) c	부정적 적중 (−적중) d	(배제) c+d
	계	a+c	b+d	a+b+c+d

베이즈 정리를 적용하여 적합한 분할점수를 추정하거나, 이미 설정된 분할점수에 의한 의사결정의 정확도를 평가하는데 도움이 되는 정보를 아래와 같이 요약할 수 있다. 아래 [그림 2]에 제시된 정보는 임상적 진단의 경우와 동일한 것으로 추정하는 방법도 같다. 단지 분할점수에 의한 의사결정 상황에 적합한 용어로 재정의했을 뿐이다.

$$+\text{적중률} = \frac{a}{a+c} \qquad \text{爲부정율} = \frac{c}{a+c}$$

$$-\text{적중률} = \frac{d}{b+d} \qquad \text{爲긍정율} = \frac{b}{b+d}$$

$$\text{성공예측도} = \frac{a}{a+b} \qquad \text{실패예측도} = \frac{d}{c+d}$$

$$\text{성공승산비}(R+) = \frac{\text{진선발율}}{1-\text{진배제율}} \qquad \text{실패승산비}(R-) = \frac{1-\text{진선발율}}{\text{진배제율}}$$

[그림 2] 베이즈 정리에 의한 분할점수 추정에 관련된 정보

기존의 측정이론이나 교육평가이론에서 주로 논의해 온 방식은 적중률(hit rate), 긍정적(+) 적중률, 선발비율(selection rate)에 의한 분할점수 선정방법이다. 적중률은 의사결정 전체에 대한 정확한 결정의 비율(a+d/a+b+c+d)로서 전체 학생들 중에서 성공할 것으로 올바르게 예측된 학생과 올바르게 실패할 것으로 예측된 학생의 비율이다. 적중률은 선발에 있어서 긍정적 오류(爲긍정율)과 부정적(-) 오류(爲부정율)의 상대적 심각성을 고려할 필요가 없는 경우 일반적으로 사용되는 지수이다. 그러나 만약 어떤 이유로 특정한 오류가 상대적으로 더 심각하다면 두 유형의 오류 혹은 두 유형의 예측적중에 대해 서로 다른 가중치를 부과해야 한다. +적중율은 성공할 것으로 예측하여 선발한 학생 중에서 실제로 성공한 학생의 비율이고, Brown(1970)에 의해 처음으로 제시된 지수이다. 만약 배제된 학생에 대해 전혀 관심이 없고, 선발된 학생 중 얼마나 많은 사람이 성공적인가에만 관심이 있다면, +적중률에 기초한 분할점수 선정이 더 적합할 것이다. 이렇게 기존 방식에 입각한 분할점수 선정방법의 문제점은 특정한 분할점수에 따른 효과에만 관심을 갖는다는 것이다. 다시 말해 검사점수의 연속선 전반에 관련된 정보가 아니라 실패-성공 등 몇 개로 나누어진 유목으로 요약된 정보만을 사용한다는 것이다(백순근, 2007). 또한 분할점수는 모집단을 대표하는 표본을 대상으로 추정한 값이기 때문에 분할점수 근처에 있는 점수에 대한 오차를 의사결정에 고려해야 한다. 오차에 대한 정보가 없는 의사결정은 부정확하기 쉬우며, 통계적으로 유의할 수 없다. 분할점수 설정은 정책적·정치적·경제적 배경에 의해 조정이 필요한 경우가 많으며, 오차는 흔히 분할점수 조정의 기본 자료로 사용된다(Angoff, 1971; Ebel, 1972; Cizek & Bunch, 2007) 기존방식에 입각한 분할점수 선정은 오차를 반영한 분할점수의 조정이 불가능하다. 베이즈 정리는 이러한 문제점에 대해 보다 적절하게 대처할 수 있는 방법을 제시해 주고 있다.

2. 가상적 자료에 의한 적용 사례

어떤 프로그램에 성공한 집단과 실패한 집단에 새로 개발한 준거참조검사를 시행하여 <표 4>과 같은 결과를 얻었다고 가정하자.

<표 4> 성공한 집단과 실패한 집단의 가상적 검사점수 빈도분포

점수	성공집단	실패집단
20	3	0
19	5	0
18	12	2
17	8	1
16	10	2
15	3	5
14	2	8
13	1	10
12	2	7
11	2	5
10	2	5
9	0	3
8	0	2
합계	50	50

위 자료를 근거로 베이즈 정리를 적용하여 적절한 분할점수를 선정하는데 필요한 정보를 <표 5>과 같이 요약·정리할 수 있다. 적중률을 최대한 높이는 것이 목적이라면, +적중률과 -적중률을 가중치 없이 동시에 높이는, 즉 爲부정률과 爲긍정률을 동시에 낮추는 방법으로 사후확률인 성공예측도와 실패예측도가 같거나 가능한 일치되는 점수를 분할점수로 선정하면 된다. <표 5>에서 보면 15점을 분할점수로 할 때 두 사후확률(성공예측도 80.4%, 실패예측도는 81.6%)이 거의 같아지며, 적중률은 81%이다. 기존의 방식에 의하면 적중률(83%)이 가장 높은 16점을 분할점수로 선정할 것이나, 이 경우 성공예측도(88.4%)와 실패예측도(78.9%)가 다소 불균형적이다. 또한 판정오류에 있어서도 기존 방식에 의한 결과(위부정률 24%, 위긍정률 10%)가 베이즈 정리에 의한 결과(위부정률 18%, 위긍정률 20%) 보다 더 불균형적이다. 또한 두 방식 간의 적중률에 있어 2%의 미세한 차이는 사례수가 다소 적고, 단지 한 표본을 대상으로 한 추정치로서 통계적으로 유의미한 차이라고 볼 수 없으며, 따라서 베이즈 정리에 의한 방식이 기존 방식보다 더 안정적인 추정치를 제시한다고 단정할 수 있다.

<표 5> 베이즈 정리에 의한 분할점수 추정에 필요한 정보

						사후확률	
점수	적중률	+적중률	위부정률	−적중률	위긍정률	성공예측도	실패예측도
20	.53	.06	.94	1.0	.00	1.000	.515
19	.58	.16	.84	1.0	.00	1.000	.517
18	.68	.40	.60	.96	.04	.909	.615
17	.75	.56	.44	.94	.06	.903	.681
16	.83	.76	.24	.90	.10	.884	.789
15	.81	.82	.18	.80	.20	.804	.816
14	.75	.86	.14	.64	.36	.705	.821
13	.66	.88	.12	.44	.56	.611	.786
12	.61	.92	.08	.30	.70	.568	.789
11	.58	.96	.04	.20	.80	.545	.833
10	.55	1.0	0.0	.10	.90	.526	1.000
9	.52	1.0	0.0	.04	.96	.510	1.000
8	.50	1.0	0.0	.00	1.0	.500	−−−

의사결정의 정확성에 대한 평가에 있어서 적중률 81%는 적합한 것인가? 이에 대한 판단은 단순히 적중률에만 기초하여 결정할 수 없으며, 기초비율과 비교하여 상대적으로 판단할 문제이다. 이 검사의 대안으로 사용할 수 있는 다른 기존의 검사도구나 절차가 있다면, 그 중에서 제시한 최고의 적중률을 기초비율로 사용해야 할 것이다. 만약에 대안으로 사용할 검사나 절차가 없을 경우, 실제 성공한 학생 혹은 실제 실패한 학생의 비율 중 높은 것을 기초비율로 사용하면 된다. 이 경우 <표 4>에서 보듯이 기초비율은 50%이며, 적중률이 이 보다 높기 때문에 이 검사가 이러한 분류의 목적으로 사용하기에 유용한 것으로 판단할 수 있다.

베이즈 정리에 의한 방식은 기존 방식이 제시하지 못하는 부과적인 정보를 제시해 준다. 성공예측도는 성공할 것으로 예측한 학생이 실제로 성공할 확률에 대한 지표이다. 어떤 프로그램에 참여할 학생들을 선발하는 과정에서 위와 같은 결과에 따라 15점을 분할점수로 선정할 경우, 선발된 학생 중에서 80%는 그 프로그램에 성공할 것으로 예측된다. 만약 그 프로그램의 예산 관계로 선발된 학생 중 90% 이상의 성공률을 요구한다면, 분할점수로 17점 혹은 그 이상 점수를 고려해 볼 수 있다. 실패할 수 있는 학생을 선발하는 爲긍정률을 최소화하기 위해서 19점을 분할점수로 선정하면 성공예측도는 100%가 된다. 반면 그 프로그램의 혜택을 더 많은 학생들에게 부여하기 위하여 60%의 성공률에 만족한다면 13점을 분할 점수로 선정할 수 있다. 성공할 수 있는 학생을 잘못하여 배제하는 위부정률을 최소화하려면 10점을 분할 점수로 선정할 수 있으며, 이 경우 성공예측도는 52.6%이고 실패예측도는 100%가 된다. 또한 이러한 의사결정 과정에서 추정치의 오차를 분할점수 조정의 기본 자료로 사용함으로써 보다 유연한 의사결정을 내릴 수 있다. 분할점수 설정을 반복적으로 시행한다고 가정할 때, 성공예측도는 상호 배반인 Bernoulli 시행

(trial)이라 볼 수 있고, 이항분포(Binominal distribution)의 확률분포를 갖는다(김달호, 2005). 따라서 성공예측도에 대한 측정의 표준오차(standard error of measurement: SE)는 아래와 같이 구할 수 있다.

$$\text{성공예측도 SE} = \sqrt{\text{성공예측도} \times (1 - \text{성공예측도}) / (a + b)}$$

분할점수가 15점일 때 성공예측도의 SE는 .0556이며, 95% 수준의 신뢰구간은 69.5%와 91.3% 사이이다. 이 신뢰구간은 14점에서 19점까지의 성공예측도를 포함하며, 사례수가 적은 관계로 SE가 상당히 크다고 볼 수 있다. 따라서 지속적인 분할점수의 타당성에 대한 검정이 요구된다.

이상의 논의와 달리 특정한 학생의 프로그램 참여 여부에 대한 상담의 경우를 생각해 보자. 이 검사에서 14점을 획득한 학생이 이 프로그램에 참여하기를 적극적으로 희망한다고 가정하자. 이 학생의 성공예측도는 70.5%이고, 95% 수준의 신뢰구간은 58%와 83% 사이다. 또한 이 학생의 성공승산비는 1.0보다 크다.

$$R+ = \frac{.86}{1 - .64} = 2.39$$

따라서 이 프로그램이 허락할 수 있다면, 학생에게 참여할 것을 권유할 수 있다. 참고로 승산비의 신뢰구간은 분할표에 대한 로그승산비의 점근적 표준오차(asymptotic standard error: ASE)를 구하는 방식을 적용하여 구할 수 있다(박광배, 2006).

Ⅴ. 결정지역 설정과 준거타당도

1. 결정지역(Decision Regions)과 결정경계 설정

검사점수에 근거하여 두 개 이상의 집단으로 분류하는 의사결정의 경우, 우리는 그 점수가 연속적 확률변수인 경우를 실제로 더 많이 접하게 된다. 이런 경우 분할표에 근거하여 적중률이 높은 분할점수를 선정하는 기존 방식을 사용하기엔 한계가 있다. 베이즈 정리를 적용하면 이러한 문제를 쉽게 해결할 수 있다.

어떤 연속변수의 점수를 준거로 하여 대학 응시자를 합격과 불합격으로 분류하는 경우를 가정해 보자. 두 가지 경우는 상호 배반(mutually exclusive)이며, 편리상 각각의 범주를 G와 $\overline{G}$로 표기하고 연속변수인 Y의 점수를 y로 표기하면, 베이즈 정리에 의해 사후확률은 아래와 같이

정의할 수 있다(참조: 공식 8, 공식 9).

$$P(G|y) = \frac{f_Y(y|G)P(G)}{f_Y(y|G)P(G)+f_Y(y|\overline{G})P(\overline{G})} \tag{10}$$

그리고

$$P(\overline{G}|y) = \frac{f_Y(y|\overline{G})P(\overline{G})}{f_Y(y|\overline{G})P(\overline{G})+f_Y(y|G)P(G)} \tag{11}$$

검사점수에 근거해 두 집단으로 분류하는 방법은 사후확률을 비교하여 각 집단에 대해 최적의 y 결정지역(decision region)을 정하는 것이다. 각 결정지역 사이를 구분하는 경계선을 결정경계(decision boundary)라 한다. 베이지안 통계에서는 연속변수의 경우 분할점수를 대신하여 결정경계라는 용어를 사용한다(Johnsonbaugh & Jost, 1996). 두 집단 G와 $\overline{G}$ 사이의 적절한 결정경계는 두 사후확률이 같아지는 것을 만족시키는 y 선상의 경계를 찾는 것이다. 즉,

$$P(G|y) = P(\overline{G}|y) \tag{12}$$

여기서 두 집단의 사후확률은 공식 (10)과 공식(11)에서 볼 수 있듯이 같은 분모를 공유하기 때문에 위 등식은 아래와 같음을 알 수 있다.

$$f_Y(y|G)P(G) = f_Y(y|\overline{G})P(\overline{G}) \tag{13}$$

따라서 학생들을 y 값에 근거해 G와 $\overline{G}$ 집단으로 분류하는데 따른 최적의 결정경계는 위의 등식을 만족시키는 것이다.

만약에 두 범주 G와 $\overline{G}$의 점수 분포는 각각 정규분포를 이루다고 가정할 수 있다면, 위 방정식(13)을 아래와 같은 정규분포 밀도함수로 표현할 수 있다.

$$P(G)\frac{1}{\sigma_G\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-\mu_G}{\sigma_G}\right)} = P(\overline{G})\frac{1}{\sigma_{\overline{G}}\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-\mu_{\overline{G}}}{\sigma_{\overline{G}}}\right)} \tag{14}$$

위 방정식에서 양변의 $\sqrt{2\pi}$를 제거하고, 양변에 natural log를 취하고 -2를 곱하면

$$-2\ln\left(\frac{P(G)}{\sigma_G}\right)+\left(\frac{y-\mu_G}{\sigma_G}\right)^2=-2\ln\left(\frac{P(\overline{G})}{\sigma_{\overline{G}}}\right)+\left(\frac{y-\mu_{\overline{G}}}{\sigma_{\overline{G}}}\right)^2 \quad (15)$$

이다. 최적의 결정경계는 이 방정식을 만족하는 y값에 위치한다. 참고로 위의 방정식은 다음과 같은 판별함수(discriminant function)로 변환 할 수 있기 때문에

$$D=-2\ln\left(\frac{P(G)}{\sigma_G}\right)+\left(\frac{y-\mu_G}{\sigma_G}\right)^2+2\ln\left(\frac{P(\overline{G})}{\sigma_{\overline{G}}}\right)-\left(\frac{y-\mu_{\overline{G}}}{\sigma_{\overline{G}}}\right)^2 \quad (16)$$

D가 0인 지점이 결정지역이다. D가 음수이면 G지역에 속할 확률이 높고, D가 양수면 $\overline{G}$지역에 속할 확률이 높다.

2. 가상적 자료에 의한 적용 사례

모 대학의 기록에 의하면 신입생이 4년 이내에 성공적으로 졸업할 확률이 0.8이고, 이 범주(G)에 속하는 학생의 점수는 평균 26이고, 표준편차 2를 가지며, 정규분포를 이룬다고 가정하자. 반면 4년 내에 졸업하지 못하는 범주($\overline{G}$)에 속하는 학생들의 점수는 평균 22이고, 표준편차 3을 가지며, 정규분포를 이룬다. 위 방정식(15)에 각각의 사전확률 $P(G)=0.8$, $P(\overline{G})=0.2$와 각각의 모수치 $\mu_G=26$, $\sigma_G=2$ 그리고 $\mu_{\overline{G}}=22$, $\sigma_{\overline{G}}=3$을 대입하면 다음의 값이 구해진다.

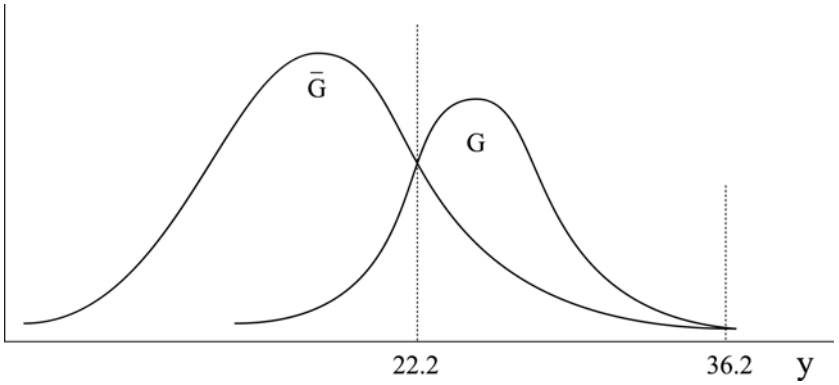
$$-2\ln\left(\frac{0.8}{2}\right)+\left(\frac{y-26}{2}\right)^2=-2\ln\left(\frac{0.2}{3}\right)+\left(\frac{y-22}{3}\right)^2$$

$$5y^2-292y+4018.99=0$$

$$y=22.2 \text{ 혹은 } y=36.2$$

이러한 두 개의 결정경계는 [그림 3]에서 볼 수 있듯이 y값 선상에 3개의 결정지역으로 나누어진다. 즉 $22.2\le y\le 36.2$인 G지역, $y<22.2$인 $\overline{G}$지역 그리고 $y>36.2$인 $\overline{G}$지역이다. 그

러나 마지막 $\overline{G}$지역 $y > 36.2$에 속하는 학생은 실제로 졸업 가능한 G로 분류되어야 한다. 따라서 한 개의 결정경계 $y = 22.2$에 의한 두 개의 결정지역만이 의미가 있을 것이다. 즉, $y \leq 22.2$인 G지역과 $y > 22.2$인 $\overline{G}$지역이다.



[그림 3] 구체적 사례에 대한 가상적 결정지역과 결정경계

이상의 논의와 달리 특정한 학생이 위 대학에 진학을 앞두고 자기가 진학하기에 적합한 대학인지에 대해 상담을 필요로 하는 경우를 가정해 보자. 만일 이 학생의 점수가 22이라면, 진학상담자는 이 학생이 4년 뒤에 졸업할 확률에 관심을 갖게 될 것이다. 이 경우 아래의 조건부 확률밀도함수에 필요한 값을 대입하여, 두 사후확률의 값들을 구한 후,

$$f_Y(y = 22|G) = \frac{1}{\sigma_G\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{22-\mu_G}{\sigma_G}\right)} = 0.027$$

$$f_Y(y = 22|\overline{G}) = \frac{1}{\sigma_{\overline{G}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{22-\mu_G}{\sigma_G}\right)} = 0.133$$

이들 값을 공식(10)에 대입하면 다음과 같은 값을 얻는다.

$$P(G|y = 22) = \frac{f_Y(y = 22|G)P(G)}{f_Y(y = 22|G)P(G) + f_Y(y = 22|\overline{G})P(\overline{G})} = 0.448$$

따라서 해당 학생의 졸업예측도는 0.5 보다 낮은 것을 알 수 있다. 이 검사에의 표준오차에 대한 정보가 있다면 해당 학생 점수의 95% 수준 신뢰구간은 22-1.96SE와 22+1.96SE 사이이다.

이 신뢰구간의 가장 낮은 점수와 가장 높은 점수에 대한 사후확률을 위와 같은 절차로 구함으로써 이 학생의 졸업예측도에 대한 95% 수준의 신뢰구간을 제시할 수 있다. 또한 졸업할 것인지 못할 것인지에 대한 졸업승산비에 관심이 있다면, 아래의 사후승산비의 공식에 대입하여 다음의 값을 갖게 된다.

$$R = \frac{P(G|y)}{P(\overline{G}y)} = \frac{f_Y(y|G)P(G)}{f_Y(y|\overline{G})P(\overline{G})}$$

$$R = \frac{0.027 \times 0.8}{0.133 \times 0.2} = 0.812$$

$R < 1$이기 때문에 졸업을 못할 확률이 높다는 것을 알 수 있다.

3. ROC 곡선에 의한 최적분할모델의 선정

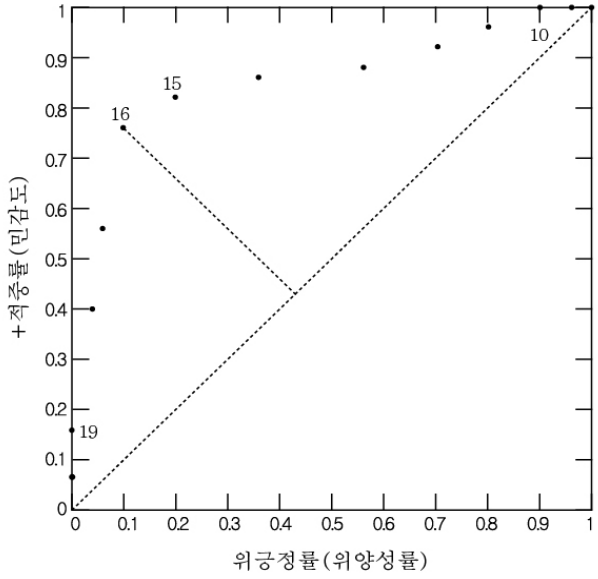
ROC(Receiver Operating Characteristic) 곡선은 진단, 선발, 분류 상황에서 사용되는 검사의 분할점수 설정이나 그에 따른 의사결정의 정확도에 대해 시각적으로 쉽게 평가할 수 있는 방법을 제공한다. 특히 ROC 분석은 연속확률분포의 '최적분할모델(optimal classification model)'을 선정하는데 유용한 도구로 사용할 수 있다(박광배, 2006; Fawcett, 2006; Zou, O'Malley, & Mauri, 2007).

ROC 도표는 각 분할점수에 따른 긍정적 적중률(혹은 민감도)과 위긍정률(혹은 위양성율)을 각각 y축과 x축으로 하여 연결한 도표로서 각 점에서의 기울기는 사후승산비(R+)이다. <표 5>에 제시된 자료에 대한 ROC 도표는 [그림 4]와 같다. 그래프의 왼쪽 하부로부터 오른쪽 상부로 45도의 점선으로 표시된 대각선은 R+가 1인 지점들로서 무작위 선정 결과에 해당한다고 볼 수 있다. 대각선의 위쪽은 R+가 1 이상인 지점들이고, 그 아래쪽은 R+가 1 이하인 지점들이다. 따라서 해당 분할점이 유용하기 위해서는 ROC 도표의 45도 대각선 위쪽에 존재해야 한다. 예를 들어, [그림 4] 도면의 왼쪽 하단 모서리에 맞닿아 있는 19점의 경우 위긍정율이 0.0(따라서 부정적 적중률은 1.0)이고 긍정적 적중률이 .16이다. 반면 오른쪽 상단의 모서리에 맞닿아 있는 10점의 경우 긍정적 적중률이 1.0(따라서 위부정률은 0.0)이고 위긍정률이 .90이다. 이 그림에서 볼 수 있듯이 적중률(즉, 긍정적 적중률+부정적 적중률)을 최대로 높이는 분할점수는 16점임을 쉽게 알 수 있다. 즉, 각 점수로부터 수직으로 대각선을 향해 선을 끄으면 16점의 경우에 그 선이 가장 길다는 것을 쉽게 분간할 수 있다. 결과적으로 16점이 분할점수로서 가장 높은 분할력

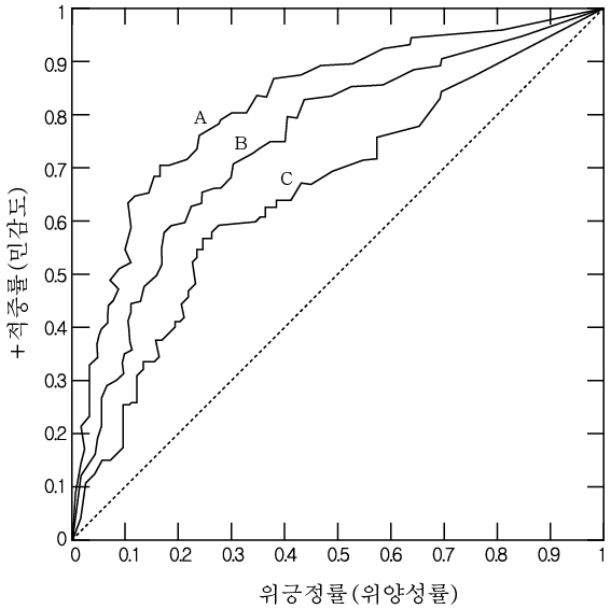
(seperation power) 혹은 준거 타당성을 갖는다는 사실을 알 수 있다.

[그림 5]는 연속확률분포를 이루는 세 유형의 대안적 검사결과에 대한 가상적 ROC 곡선을 도표를 제시한 것이다. '곡선 아랫부분의 면적(Area Under Curve: AUC)'라고 하며, 곡선 아랫부분의 면적(AUC)이 넓을수록 검사의 변별력 혹은 준거 타당도가 우수하다는 것을 의미한다. 즉 대각선으로부터 더 상부로 볼록할수록 더 우수한 분할모델(classification model)을 제공하는 검사이다. 진단, 선발 혹은 분류의 목적으로 사용 가능한 여러 대안검사가 있는 경우, ROC 분석은 최적분할모델을 제공하는 검사를 손쉽게 식별할 수 있게 해준다. 예를 들어, [그림 5]의 경우 A, B, C 검사 중에서 A검사가 가장 넓은 곡선 아랫부분의 면적(AUC)을 갖고 있기 때문에 A검사가 적합한 분할모델임을 알 수 있다.

실제 ROC 곡선을 적용함에 있어, [그림 5] 경우와 달리 쉽게 분할모델을 식별하기 다소 어려운 경우가 있을 수 있다. ROC 곡선들이 서로 중첩되거나 특정 지역에서 더 뽈록하다가도 다른 지역에서는 더 홀쭉해 질 수도 있다. 이런 경우 관심의 분할점수 지역 근처에서 더 뽈록한 ROC 곡선을 제공하는 검사가 최적분할모델이 된다.



[그림 4] <표 5>의 점수들에 대한 ROC 도면



[그림 5] 세 유형 검사의 가상적 ROC 곡선

Ⅵ. 요약 및 논의

본 연구의 목적은 교육현장에서 흔히 필요로 하는 학생들에 대한 임상적 진단이나 선발과 분류와 같은 의사결정에 있어서 베이즈 정리의 적용 가능성을 모색하고, 구체적인 적용 사례와 함께 그 교육적 함의를 논의하는데 있다. 이러한 진단, 선발, 분류와 관련한 모든 의사결정에 있어서 공통적으로 지니는 문제점은 항상 두 가지의 오류, 즉 긍정적 오류와 부정적 오류가 항상 수반될 수 있다는 사실이다. 따라서 의사결정의 정확성을 높이기 위해서는 그 오류의 원인을 잘 파악하고, 그것을 최소화할 수 있는 방법을 모색해야 한다.

임상적 진단에서 있어서 베이즈 정리는 매우 중요한 시사점을 제시한다. 어떤 검사 결과에 의한 임상적 장애에 대한 의사결정의 타당성은 그 진단도구의 정확성뿐만 아니라 모집단의 특성에 의존한다는 것이다. 이에 따라 본 연구는 모집단의 다양한 기초비율에 따라 양성예측도와 양성승산비가 얼마나 심각하게 달라지는지를 구체적인 가상 자료를 통해 살펴보았다. 즉, 진단도구의 정확성을 나타내는 민감도와 특이도가 각각 99% 수준의 정확성을 가졌더라도 유병율이 낮은 조건에서는 대부분의 양성반응 결과가 잘못된 양성이라는 것이다. 이러한 모순을 두고 '거짓 양성반응의 패러독스(false positive paradox)'라고 한다. 이러한 문제점의 심각성은 초·중등학교에서 아동과 청소년의 발달과 관련하여 정신장애, 학습장애 등의 임상적 진단을 필요로 하는 경우 대부분은 낮은 모집단의 유병율을 보인다는 점이다. 이러한 패러독스가 임상적 진단에서

시사하는 바는 모집단의 기초비율이 낮은 조건에서 인간의 특성을 하나의 검사나 절차에 의해 평가할 것이 아니라, 여러 가지 다양한 검사와 진단에 도움이 될 수 있는 가능한 모든 준거들을 동원하여 종합적으로 평가해야 한다는 것이다. 다시 말해 임상적 진단에 있어서 총평관적 접근의 중요성을 시사한다.

이와 더불어 본 연구는 교육현장에서 학생들을 선발하거나 여러 범주 혹은 수준으로 분류·배치하기 위하여 분할점수를 추정하거나, 이미 설정된 분할점수에 의한 의사결정의 정확도를 평가하는데 있어서 베이즈 정리의 적용 가능성을 구체적 사례와 함께 제시하였다. 기존의 측정이론이나 교육평가이론에서 주로 논의해 온 방식은 적중률, 긍정적 적중률과 선발비율에 의한 분할점수 선정방법이다. 이러한 기존 방식의 본질적 문제점은 특정한 기준점수에 따른 효과에만 의존하여 의사결정을 할 수 밖에 없다는 것이다. 다시 말해 검사점수의 연속선 전반에 관련된 정보가 아니라 성공-실패 등 몇 개로 나누어진 유목으로 요약된 정보만을 사용한다는 것이다. 분할점수는 모집단을 대표하는 표본을 대상으로 추정하기 때문에, 그 기준점수 근처에 있는 점수에 대한 통계적 오차를 고려하여 의사결정을 내려야 한다. 불행히도 기존방식으로는 이러한 오차를 의사결정에 반영할 방법이 없으며, 오차를 고려한 분할점수의 조정이 불가능하다. 그러나 현실적으로는 정책적·정치적·경제적 배경에 의해 분할점수의 조정이 필요한 경우가 많으며, 오차를 분할점수 조정의 기본 자료로 사용함이 바람직하다. 본 연구는 베이즈 정리를 적용함으로써 이러한 문제점에 적절하게 대처할 수 있는 방법을 표준오차 추정방법과 함께 제시하였다. 기존방식에 의한 의사결정에 있어서 또 다른 문제점은 기존방식은 학생들을 단지 서로 다른 범주로 분류하는데 필요한 정보만 제공할 뿐이며, 학생 개개인의 입장에서 상담에 필요한 정보를 제시해 주지 못한다는 점이다. 그러나 베이즈 정리에 의한 방식은 학생 개개인의 능력에 따른 특정 프로그램의 참여 여부에 대한 적합성을 결정하는데 필요한 정보를 제공해 줄 수 있다. 이에 따라, 본 연구는 구체적인 적용 사례를 통해 단순한 분류의 차원을 넘어 개별 학생에 대한 상담에 필요한 정보와 그 정보의 오차를 반영하여 의사결정을 할 수 있는 방법을 제시하였다.

한편 선발, 분류, 배치와 관련된 의사결정에 있어서 우리는 검사점수가 이산변수보다 연속적 변수인 경우를 실제로 더 많이 접하게 된다. 이런 경우 기존 방식에 따라 무수히 많은 분할표에 근거해 분할점수를 추정하거나 의사결정의 정확도를 평가하는데 한계가 있다. 이에 따라 본 연구는 베이즈 정리를 적용함으로써 이러한 문제를 쉽게 해결할 수 있음을 구체적 사례와 함께 제시하였다. 지면상의 문제로 구체적 사례를 검사점수가 정규분포를 이루는 경우에 한해 설명하였지만, 베이즈 정리에 의한 방식은 포아송(Poisson distribution)이나 균일분포(Uniform distribution) 등 어떤 확률분포 가정하에서도 적용 가능하다는데 그 장점이 있다. 더군다나 베이즈 정리에 의한 방식은 분할점수가 세 개 이상인 경우나, 구인이 두 개 이상인 경우에도 적용 가능하다. 또한 연속적 변수에 대한 베이즈 정리의 또 다른 장점은 ROC 곡선의 분석을 병행함

으로써 진단 혹은 분류 목적으로 사용 가능한 여러 가지 대안검사가 있는 경우 최적분할모델을 제시하는 검사를 손쉽게 식별할 수 있다는 것이다.

임상적 진단, 선발과 분류에 관련한 준거설정과 그에 따른 의사결정의 타당성은 단순한 일회성의 연구가 아니라 주기적(periodically)으로 평가하고 수정, 보완 및 개선 작업을 필요로 한다. 베이즈 정리는 사전 지식이 새로운 정보나 관찰에 의해 개선되어지는 과정의 수학적 체계화(mathematical formulation)라고 볼 수 있다. 베이즈 정리의 관점에서 보면, 준거타당도에 대한 사전지식(prior)이 새로운 타당도 연구를 통해 개선된 새로운 사후지식(posterior)이 되며, 추후에 또 다른 타당도 검증에서 이전의 사후지식은 사전지식이 되고 새로운 사후지식을 얻게 되면서 준거타당도는 개선될 수 있다. 그리고 이러한 과정을 반복함으로써 준거타당도에 대한 지식이 계속 개선되어 나가는 것이 '계층적 베이지안 모델(Hierarchical Bayesian model)'이 지향하는 바이다. 결론적으로 베이지안적 접근은 준거설정과 준거타당도 연구에 있어서 논리적으로 적합한 모델을 제시한다.

참고문헌

김달호 (2005). **R과 WinBUGS를 이용한 베이지안 통계학**. 경기도 파주시 : 자유아카데미.

박광배 (2006). **범주변인분석.** 서울 : 학지사.

백순근 외 (2011). **교육측정의 이론과 실제.** 서울 : 교육과학사.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. I. Thorndike (Ed.), *Educational measurement* (pp.508-600). Washington, DC: American Council on Education.

Cizek, J., & Bunch, M. B. (2011). *Standard Setting.* Bevery Hills, CA: Sage.

Ebel, R. L. (1972). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice Hall.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27,* 861-874.

Johnsonbaugh, R., & Jost, S. (1996). *Pattern recognition and image analysis*. Englewood Cliffs, NJ: Prentice Hall.

Zou, K. H., O'Malley, A. J., Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 6, 115*(5), 654-657.

* 논문접수 2011년 10월 31일 / 1차 심사 2011년 11월 30일 / 게재승인 12월 16일

* 박태학(朴泰學, Park, Tae Hak): 고려대학교 교육학과를 졸업하고, Michigan State Univ. 교육심리학과에서 석사학위를 취득하였으며, Univ. of Wisconsin-Madison 교육심리학과에서 통계 및 연구방법을 전공으로 하여 박사학위를 취득하였다. 현재 신라대학교 교육학과 부교수로 재직 중이다.
* e-mail : thpark@silla.ac.kr

Abstract

The Application of Bayes' Theorem in Decision Making such as Diagnosis, Selection, Classification and Its Implications

Park, Tae Hak*

This study seeks to recognize the applicable methods of the Bayes' Theorem on decision making such as clinical diagnosis, selection, and classification. Also this study discusses, through specific cases, the educational implications of such applicable methods. The 'False Positive Paradox' brings up an extremely important point when it comes to clinical diagnosis. Although it has accuracy when applied on a population with such rare conditions of low prevalence, a majority of the positive results could be judged as false. Because most clinical disorders are characterized by a low prevalence rate, the clinical evaluations stress the need for a comprehensive approach during clinical evaluations. This study demonstrates, through specific cases, how the Bayes' Theorem could be applied to cut-off score estimations and their validity evaluations regarding selection and classification. Opposed to previously existing methods, through Bayes' Theorem, one can make decisions regarding cut-off score setting with statistical errors. In the case of continuous variables, though classic methods require a myriad of contingency tables, Bayes' Theorem can resolve such problems. Bayes' Theorem, accompanied by the analysis of ROC curves, provides a way to easily discriminate the optimal classification model from other alternatives. Consequently, criterion settings and their validity for clinical diagnosis, selection, and classification need to be evaluated and remedied periodically. Bayes' Theorem provides an logically adequate model in respect to such criterion settings and criterion-related validity studies.

Key words : Bayes' theorem, clinical diagnostic, classification, contingency table, criterion setting, criterion validity

* Associate professor, Department of Education, Silla University