The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.2.291
32(2), 291–300

Comparison of evaluation measures for classification models on binary data

Byungsoo Kim[a,1] · Soyoung Kwon[b]

[a]Department of Statistics, Inje University;
[b]Medical Device Policy Division, Ministry of Food and Drug Safety

(Received January 18, 2019; Revised February 15, 2019; Accepted February 15, 2019)

Abstract

This study investigates the characteristics of evaluation measures for classification models on a binary response variable in order to evaluate their suitability for use. Six measures are considered: Accuracy, Sensitivity, Specificity, Precision, F-measure, and the Heidke's skill score (HSS). Evaluation measures are reformulated using x(ratio of actually 1), y(ratio predicted by 1), z(ratio of both actual and predicted by 1) from the confusion matrix. We suggest two necessary conditions to assess the suitability of the evaluation measures. The first condition is that the measure function is constant for x and y in the case of a random model. The second condition is that the measure function is increasing for z and decreasing for x and y. Since only HSS satisfies the two conditions, that is always appropriate as an evaluation measure for the classification model on the binary response variable, and the other measures should be used within a limited range.

Keywords: binary variable, measure, classification, random, Heidke's skill score

1. 서론

데이터마이닝이나 빅데이터와 같은 통계분야에서 관심 있는 목표변수에 대한 분류나 예측을 위해 모형을 구축하는 경우가 많다. 모형을 구축한 후에 그 모형이 얼마나 우수한지에 대한 모형 평가와 여러 모형들 중에서 최적 또는 최상의 모형을 선택하기 위한 모형 비교에서 다양한 측도들을 사용하게 된다. 모형 평가와 비교를 위해 사용되는 측도들은 목표변수의 형태에 따라 달라지며, 그 형태는 크게 연속형, 범주형, 순위형으로 나눌 수 있다. 범주형 중에서 목표변수가 가질 수 있는 값이 두 개인 이진형인 경우는 다양한 분야에서 접할 수 있으며 이러한 변수에 대한 예측모형이 중요하게 다루어지고 있다. 이진변수의 값은 보통 질병, 사망, 불량, 양성과 같이 사건이 일어난 경우를 1로 그렇지 않은 반대의 경우를 0으로 표기한다.

이진변수에 대한 모형이 구축되면 1에 대한 확률인 스코어가 계산되며 임계치(cutoff) 이상이면 1로 그렇지 않으면 0으로 예측하게 된다. 모형 평가와 비교는 크게 세 가지 방법으로 첫 번째 방법은 목표변

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A3B03034032).
[1]Corresponding author: Department of Statistics/Institute of Statistical Information, Inje University, 197 Inje-ro, Gimhae-si, Gyeongnam-do 50834, Korea. E-mail: statkbs@inje.ac.kr

Table 1.1. Accuracy score in each random model

Actual	Predicted								
	Case 1			Case 2			Case 3		
	0	1	Total	0	1	Total	0	1	Total
0	81	9	90	72	18	90	25	25	50
1	9	1	10	8	2	10	25	25	50
Total	90	10	100	80	20	100	50	50	100
Accuracy	82%			74%			50%		

수가 연속형인 경우와 유사하게 목표값과 스코어의 차이를 이용하는 것으로 root mean squared error (RMSE) 등이 있고, 두 번째 방법은 목표값과 0 또는 1의 예측값로 구성되는 정오분류표(confusion matrix)에 기반한 방법으로 정분류율 등이 있고, 세 번째 방법은 다양한 임계치에 따라 만들어지는 그래프를 이용하는 방법으로 receiver operation characteristics (ROC) 등이 있다 (Kim 등, 2018a).

모형 평가와 비교에서 측도들을 사용한 문헌들을 살펴보면, Sung (2013)은 기술금융 여신 선정 적절성 분석에서 정분류율을 사용하였고, Park 등 (2013)은 결핵환자 분류 모형비교에서 root average squared error (RASE)와 오분류율을 사용하였다. Kim 등 (2015)은 문제음주 예측모형의인 의사결정나무의 적합성을 평가하고자 정분류율과 이득도표를 사용하였고, Leem과 Ryu (2016)는 산업재해 예측모형을 비교하기 위해 정분류율, 오분류율, 민감도, 특이도를 사용하였으며, Kim (2016)은 병원성 위험인자의 분석모형을 비교하기 위해 민감도, 특이도, 정밀도를 이용하였다. Sohn 등 (2005)은 집중호우 예측을 위한 모형선택 기준으로 하이드게 점수(Heidke's skill score; HSS)를 사용하였고, Sakong (2012)은 주가 등락예측 모형비교에서 특이도, 민감도, 정분류율을 사용하였으며 임계치를 선택하기 위해 HSS를 사용하였다. Bekkar 등 (2013)은 1의 자료가 0에 비해 훨씬 적은 imbalanced dataset에서 G-mean, F-measure와 ROC의 area under curve (AUC)에 대해서 다루었고, Kim 등 (2018b)은 잡음제거 오토인코더를 응용한 새로운 협업 필터링 방법의 평가를 위해 F-measure를 사용하였다.

다양한 분야에서 이진자료 분류모형을 평가하기 위해서 여러 측도들이 사용되고 있지만, 특정 측도를 사용하는 근거 제시가 없거나 여러 측도들을 나열하여 종합적으로 판단하는 경우가 많다. Table 1.1은 세 모형의 예측 결과와 모형 평가에서 많이 쓰이는 정분류율을 보여주는 정오분류표이다. 세 모형에서 각 셀의 빈도는 예측이 실제와 독립적으로 이루어지는 랜덤모형일 때의 기대빈도와 같으며, 세 모형은 랜덤모형과 같이 예측력이 전혀 없는 모형이다. Case 1과 Case 2는 같은 자료에 대한 것으로 임계치에 따라 다르게 예측되는 것을 가정한 모형이고, Case 1과 Case 3은 전혀 다른 자료에 대한 예측결과를 가정한 것이다. 세 모형은 모두 랜덤모형과 같이 예측력이 없음에도 정분류율은 다른 값을 가지며, 정분류율이 모형을 평가하고 비교하는 측도로서 불완전하다는 것을 보여주고 있다.

측도들에 대한 올바른 이해와 선정이 모형 개발과 선택에 중요한 문제이며 본 연구에서는 모형 평가와 비교에서 사용되는 측도들의 특성을 파악하고 합리적인 측도를 선정하기 위한 기준을 마련할 것이다. 연구의 범위는 Kim 등 (2018a)에서 언급한 세 가지 방법 중에서 두 번째 방법인 정오분류표을 이용한 방법에 한정하였다. 본 논문의 구성은 2절에서 비교를 위한 선택한 6가지 측도들에 대해 설명하고, 3절에서 랜덤모형인 경우에 측도들의 특성을 파악하고 사용하기 적합한 측도인가를 살펴보았다. 4절에서 일반적인 모형에서의 측도들의 특성을 살펴보고, 5절에서 결론 및 본 연구의 성과에 대해 살펴보았다.

2. 분류모형 평가 측도

이진자료에 대한 분류모형이 구축되면 Table 2.1의 이원분할표가 만들어지며 이 표를 기반으로 하여 분

Table 2.1. General components of confusion matrix

Actual	Predicted		
	0	1	Total
0	a	b	$t_0 = a + b$
1	c	d	$t_1 = c + d$
Total	$f_0 = a + c$	$f_1 = b + d$	$n = a + b + c + d$

Table 2.2. Measures for confusion matrix

Measure	Formula	Measure	Formula
Accuracy	$\frac{a+d}{n}$	Precision	$\frac{d}{f_1}$
Sensitivity	$\frac{d}{t_1}$	F-measure	$\frac{2d}{t_1+f_1}$
Specificity	$\frac{a}{t_0}$	HSS	$\frac{\text{PCM}-\text{PCR}}{1-\text{PCR}}$

PCM $= (a+d)/n$; PCR $= (t_0 f_0 + t_1 f_1)/n^2$. HSS = Heidke's skill score.

Table 2.3. Confusion matrix reformulated by ratios (x, y, z)

Actual	Predicted		
	0	1	Total
0	$1 - x - y + z$	$y - z$	$1 - x = t_0/n$
1	$x - z$	$z = d/n$	$x = t_1/n$
Total	$1 - y = f_0/n$	$y = f_1/n$	1

류모형에 대한 평가와 비교가 이루어진다.

본 연구에서는 이원분할표를 기반으로 하고 모형평가와 비교를 위해 많이 사용되는 측도들인 정분류율(accuracy), 민감도(sensitivity), 특이도(specificity), 정밀도(precision), F-measure, HSS의 6가지 측도를 선정하였으며 각 측도들의 식은 Table 2.2와 같다.

정분류율은 전체 자료 중에서 맞게 분류되는 비율로서 가장 많이 사용되는 측도이다. 오분류율은 '1 − 정분류율'이고 정분류율의 성질로부터 오분류율의 성질을 쉽게 유추할 수 있으므로 논의에서 제외하였다. 민감도는 실제 값이 1인 자료 중에서 1로 맞게 예측되는 비율이고, 특이도는 실제 값이 0인 자료 중에서 0으로 맞게 예측되는 비율이며, 정밀도는 1로 예측되는 자료 중에서 실제 값이 1인 자료의 비율이다. F-measure는 정밀도와 민감도의 조화평균이며 balanced F-score 또는 F_1 score라고도 불린다. HSS의 식에서 PCM $= (a+d)/n$은 고려하는 모형의 정분류율이고 PCR $= (t_0 \times f_0 + t_1 \times f_1)/n^2$은 랜덤모형의 정분류율이다. HSS는 고려하는 모형이 랜덤모형보다 나은 정도를 나타내는 측도로서 $-\infty <$ HSS ≤ 1 범위의 값을 가지며 고려하는 모형이 랜덤모형과 같으면 HSS $= 0$이 된다.

6가지 측도들의 성질들을 쉽게 파악하고 비교하기 위해 Table 2.1의 빈도들을 Table 2.3과 같이 비율로 표현하였다. 여기서 x는 전체에서 실제로 1인 비율이고 y는 1로 예측되는 비율이며 z는 실제와 예측이 모두 1인 비율이다.

Table 2.3에서 (x, y, z) 이외의 값들은 (x, y, z)에 따라 정해지므로 이원분할표에 기반한 측도들은 모두 (x, y, z)의 함수가 되며 Table 2.4는 6개 측도를 (x, y, z)에 따라 표현한 것이다. 6개의 모든 측도들은 x와 y가 고정되었을 때 z가 클수록 커지며, z가 고정되었을 때 x와 y가 증가할수록 감소한다. 민감도는 y의 영향을 받지 않고, 정밀도는 x의 영향을 받지 않는다.

Table 2.4. Measures reformulated by ratios (x, y, z)

Measure	Formula	Measure	Formula
Accuracy	$1 - x - y + 2z$	Precision	$\frac{z}{y}$
Sensitivity	$\frac{z}{x}$	F-measure	$\frac{2z}{x+y}$
Specificity	$\frac{1-x-y+z}{1-x}$	HSS	$\frac{2z-2xy}{x+y-2xy}$

HSS = Heidke's skill score.

Table 3.1. Measures in case of random model $(z = xy)$

Measure	Formula	Measure	Formula
Accuracy	$1 - x - y + 2z$	Precision	x
Sensitivity	y	F-measure	$\frac{2xy}{x+y}$
Specificity	$1 - y$	HSS	0

HSS = Heidke's skill score.

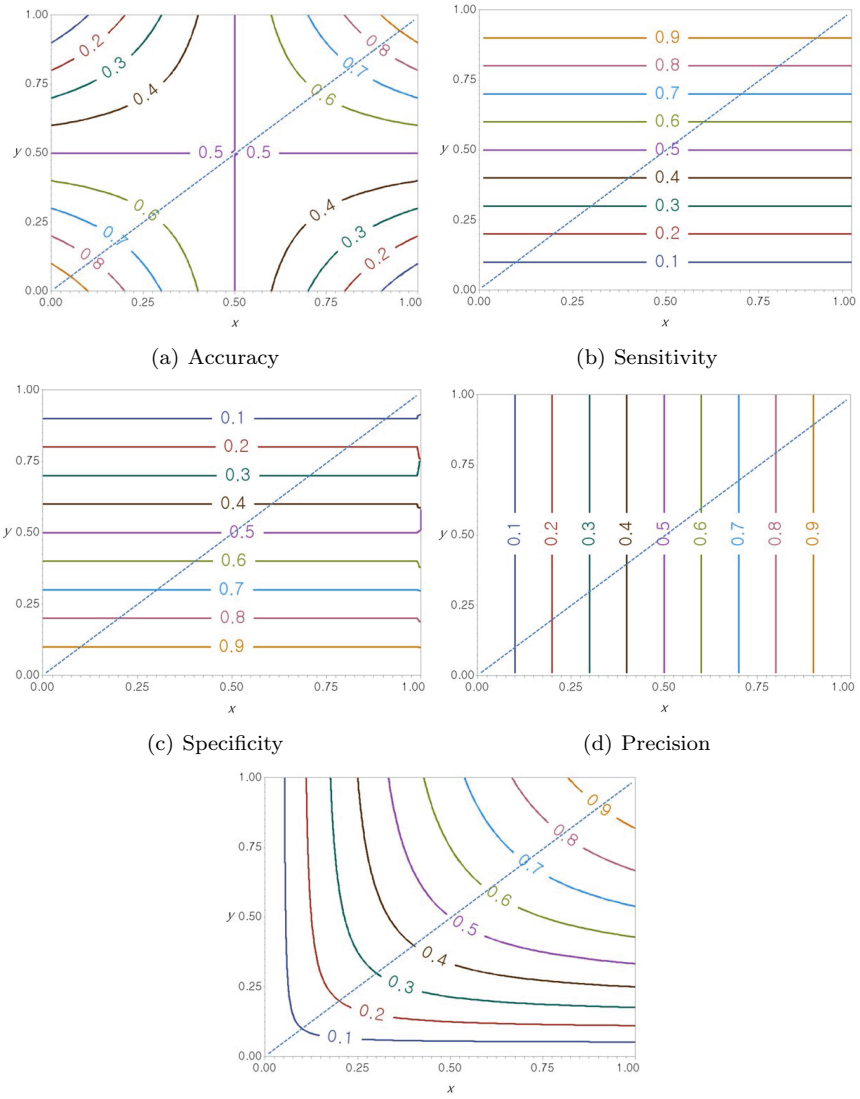
3. 랜덤모형인 경우

랜덤모형은 예측이 실제와 관계없이 이루어지는 모형으로 모형의 성능이 전혀 없다. 본 논문에서는 랜덤모형일 때 각 측도들의 성질을 쉽게 파악하고 비교하기 위해 랜덤모형을 Table 2.2에서 각 셀의 비율이 실제와 예측이 서로 독립일 때 나타나는 기대비율과 정확히 같아지는 것으로 한정하였다. 이러한 랜덤모형에서는 z가 xy와 같으며 Table 2.3에서 z를 xy로 바꾸면 Table 3.1과 같은 식을 얻을 수 있다. HSS는 x와 y의 값과 관계없이 0의 값을 갖으며, 정분류율과 F-measure는 x와 y의 함수로 주어지며 x와 y의 값이 바뀌어도 같은 측도의 값을 갖는 대칭인 성질이 있다. 민감도와 특이도는 y에 따라 결정되며 반대 방향으로 움직이고 정밀도는 x만의 함수로 주어진다. Figure 3.1은 x와 y의 변화에 따른 HSS를 제외한 5개 측도들의 값을 나타내는 등고선 그림이다.

고려하는 모형이 랜덤모형인 경우에 측도는 x와 y에 따라 변하지 않는 고정된 상수값을 가져야 하는 것은 당연하다. 그렇지 않다면 랜덤모형임에도 서로 다른 측도의 값들을 가지게 되고 특정 랜덤모형이 다른 랜덤모형보다 더 좋은 또는 더 좋지 않은 모형으로 평가될 수 있기 때문이다. 본 논문에서는 측도가 가져야 하는 이런 조건을 제1조건이라 부르겠다. 제1조건에서 본다면 HSS를 제외한 5개의 측도들은 평가 측도로서 사용하기에 적합하지 않다. 랜덤모형인 경우 x 또는 y에 제한이 주어진다면 6개 측도들이 제1조건을 만족하는지를 살펴볼 필요가 있으며, 이 절에서는 x가 고정된 경우, y가 고정된 경우, 비교비(y/x)가 1인 경우에 측도들의 성질에 대해 살펴보았다.

3.1. 실제 1의 비율(x)을 고정시킨 경우

하나의 자료가 주어지면 실제 1의 비율인 x는 고정되며 모형 또는 분계점에 따라 y는 변하게 된다. Table 3.1과 같이 예측모형이 랜덤모형인 경우에 x가 고정되면 정밀도와 HSS는 y에 따라 변하지 않는다. 정분류율은 $(2x-1)y + c$의 형태를 가지며 x가 0.5보다 작을 때 y에 대해 감소함수이고 x가 0.5보다 클 때 y에 대해 증가함수이다. x가 0.5이면 정분류율은 상수인 0.5가 되어 y와 무관하다. 민감도는 y에 대해 증가함수이고 특이도는 감소함수이다. F-measure는 y에 대해 증가함수이고 x가 클수록 증가폭이 크다. x는 고정되고 랜덤모형일 때 정밀도와 HSS는 항상 상수값을 갖고, 정분류율은 x가 0.5일



(a) Accuracy (b) Sensitivity

(c) Specificity (d) Precision

(e) F-measure

Figure 3.1. Contour plot of measures for x and y in case of random model.

때 상수값을 갖고, 나머지 3개의 측도는 상수값을 갖지 않는다. 따라서 랜덤모형일 때 상수값을 가져야 한다는 제1조건의 측면에서 x가 고정되었을 때의 정밀도와 HSS 그리고 x가 0.5일 때의 정분류율은 적합한 측도로 판단된다.

3.2. 예측 1의 비율(y)을 고정시킨 경우

Table 3.1과 같이 예측모형이 랜덤모형인 경우에 y가 고정되면 민감도, 특이도, 그리고 HSS는 x에 따라 변하지 않는다. 정분류율은 $(2y-1)x+c$의 형태를 가지며 y가 0.5보다 작을 때 x에 대해 감소함수

이고 y가 0.5보다 클 때 x에 대해 증가함수이다. y가 0.5이면 정분류율은 상수인 0.5가 되어 x와 무관하다. 정밀도는 x에 대해 증가함수이고 F-measure는 x에 대해 증가함수이고 y가 클수록 증가폭이 크다. y는 고정되고 랜덤모형일 때 민감도, 특이도, HSS는 항상 상수값을 갖고, 정분류율은 y가 0.5일 때 상수값을 갖고, 나머지 2개의 측도는 상수값을 갖지 않는다. 따라서 랜덤모형일 때 상수값을 가져야 한다는 제1조건의 측면에서 y가 고정되었을 때의 민감도, 특이도, HSS와 y가 0.5일 때의 정분류율은 적합한 측도로 판단된다.

3.3. 비교비(y/x)가 1인 경우

비교비가 1인 경우는 실제 1의 비율과 1로 예측하는 비율이 같은 경우, 즉 $x = y$인 경우이며 Figure 3.1에서 대각선의 점선은 비교비가 1인 경우의 측도들의 값이다. 비교비가 1이면 HSS만 0으로 상수가 되고 다른 측도들은 모두 x의 함수로 표현된다. 정분류율은 $2(x - 0.5)^2 + 0.5$으로 표현되며 $x = 0.5$일 때 최솟값인 상수 0.5를 가진다. 민감도, 정밀도, F-measure들의 값은 모두 x가 되어 x에 대해 증가함수이고 특이도는 $1 - x$가 되어 x에 대해 감소함수이다. 비교비가 1이고 랜덤모형일 때 HSS만 항상 상수값을 갖고, 정분류율은 x가 0.5일 때 상수값을 갖고, 나머지 4개의 측도는 상수값을 갖지 않는다. 따라서 랜덤모형일 때 상수값을 가져야 한다는 제1조건의 측면에서 비교비가 1일 때의 HSS와 x가 0.5일 때의 정분류율은 적합한 측도로 판단된다.

4. 일반모형인 경우

고려하는 모형이 랜덤모형이 아닌 일반적인 모형인 경우에 측도들은 Table 2.4와 같이 세 변수들 (x, y, z)로 이루어진 다소 복잡한 형태의 식을 갖는다. 고정된 x와 y에 대해 z가 클수록 좋은 모형인 것은 당연하며 고려하는 측도가 사용하기 적합한 측도가 되기 위해서는 좋은 모형일수록 측도의 값은 커져야 한다. 즉 특정 측도가 모형이 우수한가를 평가할 수 있는 적합한 측도가 되기 위해서는 측도의 식이 세 변수들(x, y, z) 모두로 이루어지고 z에 대해서 증가함수이고 x와 y에 대해서 감소함수이어야 한다. 본 논문에서는 측도가 가져야 하는 이 조건을 제2조건이라 부르겠다. 제2조건에서 본다면 정분류율, 특이도, F-measure, HSS는 적합한 측도이고, 민감도와 정밀도는 적합하지 않은 측도이다.

일반적인 모형에 대해서 x 또는 y에 제한이 주어졌을 때 6개 측도들이 제2조건을 만족하는지를 살펴볼 필요가 있으며, 이 절에서는 x가 고정된 경우, y가 고정된 경우, 비교비(y/x)가 1인 경우에 측도들의 성질에 대해 살펴보았다. x와 y가 주어졌을 때 z는 $\max(0, x + y - 1)$와 $\min(x, y)$ 사이의 값을 가질 수 있다.

4.1. 실제 1의 비율(x)을 고정시킨 경우

Table 4.1은 x가 상수(c)일 때 각 측도들의 y와 z에 대한 식을 보여주는 표이며, 민감도는 z만의 함수이고 나머지 5개 측도들은 y와 z에 대한 함수임을 볼 수 있다. y가 고정되었을 때 z가 증가할수록 모든 측도들의 값은 증가하며 z가 고정되었을 때 y가 증가할수록 민감도를 제외한 모든 측도들은 감소한다. 따라서 x가 고정되었을 때 민감도를 제외한 5개 측도들은 제2조건을 만족하는 적합한 측도이다.

4.2. 예측 1의 비율(y)을 고정시킨 경우

Table 4.2는 y가 상수(c)일 때 각 측도들의 x와 z에 대한 식을 보여주는 표이며, 정밀도는 z만의 함수이고 나머지 5개 측도들은 x와 z에 대한 함수임을 볼 수 있다. x가 고정되었을 때 z가 증가할수록 모든 측

Table 4.1. Measures in case x is constant c

Measure	Formula	Measure	Formula
Accuracy	$1-c-y+2z$	Precision	$\frac{z}{y}$
Sensitivity	$\frac{z}{c}$	F-measure	$\frac{2z}{c+y}$
Specificity	$\frac{1-c-y+z}{1-c}$	HSS	$\frac{2z-2cy}{c+y-2cy}$

HSS = Heidke's skill score.

Table 4.2. Measures in case y is constant c

Measure	Formula	Measure	Formula
Accuracy	$1-c-x+2z$	Precision	$\frac{z}{c}$
Sensitivity	$\frac{z}{x}$	F-measure	$\frac{2z}{c+x}$
Specificity	$\frac{1-c-x+z}{1-x}$	HSS	$\frac{2z-2cx}{c+x-2cx}$

HSS = Heidke's skill score.

Table 4.3. Measures in case the comparability ratio is 1 ($y=x$)

Measure	Formula	Measure	Formula
Accuracy	$1-2x+2z$	Precision	$\frac{z}{x}$
Sensitivity	$\frac{z}{x}$	F-measure	$\frac{z}{x}$
Specificity	$\frac{1-2x+z}{1-x}$	HSS	$\frac{z-x^2}{x-x^2}$

HSS = Heidke's skill score.

도들의 값은 증가하며 z가 고정되었을 때 x가 증가할수록 정밀도를 제외한 모든 측도들은 감소한다. 따라서 y가 고정되었을 때 정밀도를 제외한 5개 측도들은 제2조건을 만족하는 적합한 측도이다.

4.3. 비교비(y/x)가 1인 경우

Table 4.3은 y가 비교비가 1일 때($x=y$) 각 측도들의 x와 z에 대한 식을 보여주는 표이다. 6개의 모든 측도들은 x가 고정되었을 때 z가 증가함에 따라 증가하고 z가 고정되었을 때 x가 증가함에 따라 감소한다. 따라서 비교비가 1인 경우에 6개의 측도들은 모두 제2조건을 만족한다.

5. 결론

본 논문에서는 반응변수가 이진형인 분류모형에 대한 평가측도들의 특성을 파악하고 사용하기 적합한 평가측도인가를 살펴보았다. 고려한 측도는 정분류율, 민감도, 특이도, 정밀도, F-measure, HSS의 6개이다. 이원분할표에서 각 측도들을 x(실제로 1인 비율), y(1로 예측되는 비율), z(실제와 예측이 모두 1인 비율)를 사용하여 표현하였다.

본 연구에서 평가측도가 사용하기 적합한 측도가 되기 위해 가져야 하는 조건으로 두 가지를 제안하였다. 제1조건은 랜덤모형인 경우에 평가측도들은 x와 y에 따라 변하지 않는 상수값을 가지는 것이다. HSS는 항상 상수값을 가지며 나머지 5개 측도들은 x 또는 y에 따라 변하는 값을 가지므로 HSS는 제1조건을 만족하고 나머지 5개 측도들은 만족하지 않는다. 따라서 x 또는 y에 제한을 주지 않으면 HSS를 제외한 5개의 측도들은 평가측도로서 적합하지 않다. x가 고정되었을 때의 정밀도, y가 고정되었을 때의 민감도 및 특이도, x 또는 y가 0.5일 때의 정분류율은 상수값을 가지므로 제한된 범위 안에서 제1조건을 만족하며 평가측도로 사용하기 적합하다.

적합한 측도가 되기 위한 제2조건은 평가측도의 식이 세 변수들(x, y, z) 모두로 이루어지고 z에 대해서 증가함수이고 x와 y에 대해서 감소함수이어야 한다는 것이다. 정분류율, 특이도, F-measure, HSS는 제2조건을 만족하고 민감도와 정밀도는 만족하지 못한다. 따라서 x 또는 y에 제한을 주지 않으면 민감도와 정밀도는 평가측도로서 적합하지 않다. x가 고정되었을 때의 정밀도, y가 고정되었을 때의 민감도, 그리고 비교비가 1인 경우의 민감도 및 정밀도는 제2조건을 만족하며 제한된 범위 안에서 적합한 측도이다.

고려한 6개 측도들이 제1조건과 제2조건을 모두 만족하는 범위를 본 논문에서 제한을 둔 x가 고정되었을 때, y가 고정되었을 때, 그리고 비교비가 1인 경우와 함께 살펴보았다. 정분류율은 x 또는 y가 0.5일 때, 민감도는 y가 고정되었을 때, 특이도는 y가 고정되었을 때, 정밀도는 x가 고정되었을 때 두 조건을 만족하였다. F-measure는 두 조건을 만족하는 경우가 없고, HSS는 항상 두 조건을 만족하였다. 따라서 HSS는 이진형 반응변수의 분류모형에 대한 평가측도로 항상 사용이 적합하고 다른 측도들은 특정한 범위 안에서는 사용해도 좋지만 범위를 벗어나서 사용하는 것은 신중해야 한다는 결론을 내릴 수 있다.

본 연구에서 고려하지 않았지만 아래의 식으로 표현될 수 있는 Matthew's correlation coefficient (MCC)는 본 연구에서 제시한 두 조건을 모두 만족하는 것으로 보이며 심도 깊은 논의가 필요할 것으로 여겨진다.

$$\mathrm{MCC} = \frac{z - xy}{\sqrt{xy(1-x)(1-y)}}.$$

본 연구의 결과를 바탕으로 다양한 연구들이 이어질 것으로 기대한다. 예를 들어 고려하는 모형이 랜덤모형보다 좋은 지에 대한 검정은 HSS가 0보다 큰 지에 대한 검정으로 가능할 것이다. 또한 HSS는 0과 1이 바뀌어도 같은 식을 가지므로 HSS를 확장하여 3개 이상의 범주에 대한 분류에도 사용 할 수 있을 것으로 보이므로 여기에 관한 연구는 좋은 과제가 될 것이다.

References

Bekkar, M., Djemaa, H. K., and Altiouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets, *Journal of Information Engineering and Applications*, **3**, 27–38.

Kim, B., Bae, W., Seok, K., Cho, D., and Choi, K. (2018a). *SAS EM 14.1 Data Mining Basis and Application*, Kyowoo, 293–317.

Kim, H., Shin, D., Shin, W., and Hwang, C. (2018b). Rating Information-Aided Denoising AutoEncoder for effective collaborative filtering, *The Journal of Korean Institute of Communications and Information Sciences*, **43**, 357–1367.

Kim, M., Kim, S., and Ock, C. (2015). A predictive model of problem drinking of workers using decision tree analysis, *Journal of The Korean Society of Living Environmental System*, **22**, 460–468.

Kim, S. Y. (2016). *The comparison of analytical models for risk factors of colonic adenomatous polyp* (Master Thesis), Graduate School, Chung-Ang University.

Leem, Y. M. and Ryu, C. H. (2006). A comparison of data mining techniques for predicting model of

industrial accidents. In *Proceedings for the Spring Conference 2006, Society of Korea Industrial and Systems Engineering*, 107–113.

Park, I., Kim, Y., Choi, Y., Kim, S., Kim, E., Won, S., and Kang, S. (2013). Development of advanced TB case classification model using NHI claims data, *The Journal of Digital Policy & Management*, **11**, 289–299.

Sakong, J. H. (2012). *A study on predicting stock price based on data mining techniques* (Master Thesis), Graduate School, Inje University.

Sohn, K., Lee, J., Lee, S., and Ryu, C. (2005). Statistical models for prediction of heavy rain in Honam area, *Asia-Pacific Journal of Atmospheric Sciences*, **41**, 897–907.

Sung, O. (2013). A empirical study on the relevance of technology finance supporting business for technologically innovative SMEs, *Journal of Korea Technology Innovation Society*, **16**, 303–322.

이진자료 분류모형에 대한 평가측도의 특성 비교

김병수[a,1] · 권소영[b]

[a]인제대학교 통계학과, [b]식품의약품안전처 의료기기정책과

(2019년 01월 18일 접수, 2019년 02월 15일 수정, 2019년 02월 15일 채택)

요 약

본 논문에서는 반응변수가 이진형인 분류모형에 대한 평가측도들의 특성을 파악하고 사용하기 적합한 평가측도인가를 살펴보았다. 고려한 측도는 정분류율, 민감도, 특이도, 정밀도, F-measure, HSS (Heidke's skill score)의 6개이다. 각 측도들은 이원분할표에서 x(실제로 1인 비율), y(1로 예측되는 비율), z(실제와 예측이 모두 1인 비율)을 사용하여 표현하였다. 본 연구는 평가측도가 사용하기 적합한 측도가 되기 위한 조건으로 두 가지를 제안하였다. 제1조건은 랜덤모형인 경우에 평가측도는 x와 y에 대해 상수이고, 제2조건은 평가측도의 식이 세 변수들(x, y, z) 모두로 이루어지고 z에 대해서 증가함수이고 x와 y에 대해서 감소함수이어야 한다는 것이다. HSS는 두 조건을 모두 만족하므로 이진형 반응변수의 분류모형에 대한 평가측도로 항상 사용이 적합하고, 다른 측도들은 제한된 범위 내에서만 사용하는 것이 좋다.

주요용어: 이진변수, 평가측도, 분류모형, 랜덤모형, 하이드게 점수

이 논문은 2017년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2017R1D1A3B03034032).

[1]교신저자: (50834) 경남 김해시 인제로 197, 인제대학교 통계학과, 통계정보연구소. E-mail: statkbs@inje.ac.kr