The Korean Journal of Applied Statistics
2021, Vol. 34, No. 5, 711–721

DOI: https://doi.org/10.5351/KJAS.2021.34.5.711
Print ISSN1225-066X / Online ISSN 2383-5818

Selecting the optimal threshold based on impurity index in imbalanced classification

Shuin Jang[a], In-Kwon Yeo[1,a]

[a]Department of Statistics, Sookmyung Women's University

Abstract

In this paper, we propose the method of adjusting thresholds using impurity indices in classification analysis on imbalanced data. Suppose the minority category is Positive and the majority category is Negative for the imbalanced binomial data. When categories are determined based on the commonly used 0.5 basis, the specificity tends to be high in unbalanced data while the sensitivity is relatively low. Increasing sensitivity is important when proper classification of objects in minority categories is relatively important. We explore how to increase sensitivity through adjusting thresholds. Existing studies have adjusted thresholds based on measures such as *G*-Mean and F1-score, but in this paper, we propose a method to select optimal thresholds using the chi-square statistic of CHAID, the Gini index of CART, and the entropy of C4.5. We also introduce how to get a possible unique value when multiple optimal thresholds are obtained. Empirical analysis shows what improvements have been made compared to the results based on 0.5 through classification performance metrics.

Keywords: imbalanced data, binomial classification, threshold moving, impurity index

1. 서론

불균형 자료(imbalanced data)란 범주 간 개체 수의 차이가 큰 자료이다. 특정 범주의 개체가 다른 범주의 개체에 비해 현저히 많거나 적은 자료를 의미한다. 이때, 개체의 수가 많은 범주를 다수 범주, 개체의 수가 적은 범주를 소수 범주라고 한다. 이러한 불균형 자료는 다양한 분야에서 관측되고 있는데, 희귀질병 감염여부 판정, 제조 과정에서의 불량여부 판정, 신용카드의 이상거래여부 탐지, 보험청구에서 사기여부 탐지 등이 불균형 자료의 예이다. 이러한 불균형 자료에 대해 분류 분석을 할 때 소수범주의 확률이 낮게 추정되어 0.5와 같이 통상적인 기준으로 분류 판정을 하면 다수범주의 개체뿐만 아니라 대부분의 소수범주 개체들도 다수범주로 분류되는 경향이 있다. 하지만 전체 자료의 대부분을 차지하는 다수범주 개체가 제대로 분류되기 때문에 전체적으로는 오분류율이 낮아 분류 성능이 높은 것으로 잘못 인식될 수 있다 (Longadge 등, 2013). 앞에서 언급한 희귀병 감염, 불량품 생산, 신용카드 이상거래, 보험사기와 같은 불균형 자료에서는 소수범주에 속한 개체를 소수범주로 분류하는 것이 중요하기 때문에 오분류율만으로 분류 성능을 평가하는 것은 적절하지 않다. 불균형 자료에서 이러한 문제를 해결하기 위한 다양한 연구가 진행되어 왔으며 대표적인 방법으로 범주 내 개체의 수를 임의로 변형하는 샘플링 방법과 분류확률의 임계값을 조정하여 소수 범주로 배치될 가능성을 높이는 임계값 조정 방법 등이 있다.

샘플링은 모형추정 과정에서 다수 범주 또는 소수 범주의 개체 수를 조정하여 균형을 맞추는 방법이다. 샘플링에서 다수와 소수 범주의 비율을 1:1로 만들 필요는 없으며 R이나 Python에서 제공하는 패키지에서는 이

[1] Corresponding author: Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul 04310, Korea. E-mail: inkwon@sookmyung.ac.kr

Published 31 October 2021 / journal homepage: http://kjas.or.kr

비율을 조절 할 수 있도록 하고 있다. 샘플링 방법에는 다수 범주를 소수 범주의 개체 수만큼 적게 추출하는 언더샘플링(under-sampling, down-sampling)과 소수 범주를 다수 범주의 개체 수만큼 많이 추출하는 오버샘플링(over-sampling, up-sampling)이 있으며 소수범주의 근접개체들 사이의 임의의 지점에 새로운 가상 개체들을 생성하는 SMOTE (Chawla 등, 2002) 방법 등이 사용된다. 언더샘플링은 전체 자료의 크기가 감소하기 때문에 모형의 적합 속도가 빨라 효율적이지만 많은 양의 개체가 제거되면서 분류 추정에 중요한 영향을 미치는 개체가 제거될 수 있으며, 전체적으로 정보 손실이 발생할 수 있다. 오버샘플링은 개체를 제거하지 않기 때문에 정보 손실이 발생하지 않지만 자료가 커지면서 모형의 적합 시간이 길어지고, 소수 범주의 복원추출로 자료가 중복되어 과적합 문제가 발생하기도 한다. SMOTE는 소수범주의 관측값을 그대로 복제하는 것이 아니기 때문에 과적합될 가능성이 오버샘플링보다 낮고 정보의 손실이 발생하지 않는다는 장점을 가지고 있어 많이 적용되는 방법이다. 각 샘플링 방법을 적용하는데 있어 주의해야 할 사항에 대해서는 Kim과 Lee (2017)을 참고하기 바란다. 샘플링 방법에 따른 분류성능은 자료의 특성에 따라 차이가 있기 때문에 어떤 방법이 더 우위에 있다고 말하기 어려우며 샘플링을 얼마나 많이 했는지에 따라 분류 결과가 달라진다. 또한 샘플링 자료를 통해 추정된 분류 확률은 원자료의 추정 확률과 다르기 때문에 결과 활용에 한계가 존재한다. 샘플링 방법에 대한 자세한 활용 및 비교결과는 Kim 등 (2015)과 Kim과 Lee (2017)를 참고하기 바란다.

본 논문에서는 또 다른 방식인 임계값을 조정하는 방법에 대해 알아본다. 샘플링 방법은 모형적합 전에 자료 변형을 가하지만 임계값 조정방법은 원자료를 그대로 적합한 후 분류할 때 최적화시키는 분류 임계값을 선택한다. 임계값 조정방법은 분류 향상과 동시에 추정한 분류 확률을 그대로 사용할 수 있다는 장점이 있다. 임계값 조정에 대한 연구로 Voigt 등 (2014), Zou 등 (2016), Collell 등 (2018)이 있으며 이 임계값 조정 방법을 이해하기 위해서는 분류성능측도를 알아야 하기 때문에 관련 설명은 2절에서 다루기로 한다. 본 논문에서는 이들이 제안한 방법 이외 의사결정나무에서 분지 기준으로 사용되는 엔트로피(entropy)나 지니지수(Gini's index)와 같은 불순도 지수(impurity index)와 CHAID에서 사용하는 카이제곱 통계량을 기반으로 임계값을 선택하는 방법을 제안한다. 또한 임계값 조정 방법을 사용하는데 있어 발생하는 문제와 이에 대한 해결 방법을 알아본다. 비대칭 자료 관련 연구에서 사용되는 실제자료를 이용하여 제안 방법의 성능을 비교해 본다.

2. 임계값 조정을 통한 분류

범주 분류에서의 임계값(threshold)이란 개별 자료에 대해 어느 범주로 할당할 것인지를 결정하는 분류 기준값으로 이 값보다 큰 분류확률을 가지는 범주에 해당 자료를 할당한다. 범주 불균형을 처리하는 기법 중 하나인 임계값 조정(threshold-moving) 방법은 분류 학습 전 자료 전처리를 통해 불균형 문제를 해결하는 샘플링방법과 다르게 완성된 분류 모형의 결과를 이용한다 (Collell 등, 2018). 이항 분류의 경우 기본 임계값으로 0.5를 사용하는데 불균형 자료의 분류에서 소수범주 개체에 대해 해당 범주에 속할 확률 추정값이 작아 0.5를 임계값으로 설정하면 거의 대부분의 소수범주 개체가 다수범주로 분류되는 경향이 있다. 임계값 조정 방법은 이 임계값을 낮추어 소수 범주로 분류될 가능성을 높인다. 그러나 임계값을 낮추면 다수 범주의 개체를 소수범주로 잘못 분류할 수 있기 때문에 범주별 성능 점수의 균형을 고려하여 최적의 임계값을 선택해야 한다.

2.1. 분류성능측도

불균형 자료 연구에서는 기본적으로 분류범주가 2개 즉, 이항(binomial) 또는 이진(binary)인 상황을 가정한다. 본 논문에서도 범주가 2개인 이항 불균형 자료라고 가정하고 다수 범주는 Negative (N) 소수 범주는 Positive (P)로 표시하여 분석한다. 범주를 숫자로 표시했을 때에는 N은 0, P는 1로 표시한다. 이항 분류에서의 분류결과에 따라 다음과 같은 혼동 행렬(confusion matrix)을 기반으로 성능 지표을 유도한다.

	분류 N	분류 P
실제 N	True Negative (TN)	False Positive (FP)
실제 P	False Negative (FN)	True Positive (TP)

이 행렬에서 TP, FP, FN, TN는 해당 셀의 빈도를 의미하며 성능평가에서는 다음과 같은 측도가 많이 사용된다.

$$\begin{aligned}
\text{정확도(Accuracy)} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}},\\
\text{민감도(Sensitivity)} &= \frac{\text{TP}}{\text{TP} + \text{FN}},\\
\text{특이도(Specificity)} &= \frac{\text{TN}}{\text{TN} + \text{FP}},\\
\text{정밀도(Precision)} &= \frac{\text{TP}}{\text{TP} + \text{FP}}.
\end{aligned}$$

정확도는 정분류율이라고도 하며 전체자료에서 정분류된 자료의 비율로 이와 반대되는 개념인 오분류율(misclassification rate)은 '1-정확도'로 표시된다. 민감도는 소수(Postive)범주의 자료에 대해 정분류된 비율, 특이도는 다수(Negative)범주의 자료에 대해 정분류된 비율을 의미하며 소수 범주로 분류된 자료 중에서 정분류된 비율을 정밀도라고 한다. 민감도를 Recall^+ 또는 Acc^+, 특이도를 Recall^- 또는 Acc^-로 표시하기도 한다.

분류모형에 대한 분류성능은 여러 임계값을 기준으로 민감도를 y축, '1-특이도'인 false positive rate (FPR)를 x축에 표시하여 변화를 시각화한 receiver operating characteristic (ROC) 곡선으로 비교하거나 이를 수치하하기 위해 이 곡선의 아래 면적인 area under curve (AUC)를 계산하여 이 값이 1에 얼마나 가까운지로 비교한다. 이 기준에 의한 비교는 여러 분류모형에 대한 성능을 비교하기 위한 것이며 최적의 임계값을 정하는 방법은 아니다.

불균형 자료의 경우 분류 모형은 절대적으로 많은 다수 범주 자료를 다수로 분류하기 때문에 특이도는 높은 반면 소수 범주 자료도 다수범주로 분류하는 경향이 높아 일반적으로 민감도는 낮게 나온다. 따라서 불균형 자료에서 소수범주에 대한 판별이 상대적으로 중요할 때에서 적절한 수준으로 특이도를 낮추더라도 민감도는 높이는 임계값 조정이나 모형을 선택하는 것이 적절하다고 할 수 있다. 불균형 자료에 대한 분류모형의 성능비교의 다양한 기준은 Akosa (2017)에 잘 정리되어 있으며 그 중 문헌 상에 많이 언급되는 측도가 균형정확도(balanced accuracy), G-mean 그리고 $F1$-score이다. 균형정확도는 민감도와 특이도의 산술평균인 (민감도+특이도)/2이다. G-mean는 민감도와 특이도의 기하평균으로 동일한 균형정확도를 가지는 상황에서는 특이도와 민감도가 비슷한 값을 가질 때 커지기 때문에 민감도를 높게 만드는 임계값을 선택하는 근거를 제공한다.

$$G\text{-Mean} = \sqrt{\text{민감도} \times \text{특이도}}$$

균형정확도와 G-mean은 민감도와 정확도를 기반한 측도인데 반하여 F1-score의 민감도와 정밀도의 조화평균으로 표시된다.

$$F1\text{-score} = \frac{2 \times \text{정밀도} \times \text{민감도}}{\text{정밀도} + \text{민감도}}$$

민감도와 정밀도가 0이면 $F1$-score는 0으로 정의한다. Akosa (2017)에 의하면 이 측도는 균형자료일 때 성능이 좋고 불균형일 때에는 수정된 F-측도(Adjusted F-measure)를 통해 개선시킬 수 있다고 한다. 일반적인 F-측도는 다음과 같이 정의하고 있으며

$$F\text{-measure} = \frac{\left(\beta^2 + 1\right) \times \text{민감도} \times \text{정밀도}}{\text{민감도} + \beta \times \text{정밀도}}, \quad \beta \geq 0,$$

$F1$-score는 F-측도에서 $\beta = 1$인 값이다 (Espindola와 Ebecken, 2005). F-측도에서 가중값을 측도에 포함한 것처럼 가중값 $\alpha, \beta > 0$을 반영한 균형정확도나 G-mean도 고려할 수도 있다.

$$\text{가중정확도} = \frac{\alpha \times \text{민감도} + \beta \times \text{특이도}}{\alpha + \beta}, \qquad \text{가중 } G\text{-Mean} = \left(\text{민감도}^{\alpha} \times \text{특이도}^{\beta}\right)^{\frac{1}{(\alpha+\beta)}}$$

위의 측도를 기준으로 다양한 임계값 조정 방법 연구들이 이루어졌는데 Collell 등 (2018)은 보정된 사후확률 추정량을 얻어 $F1$-Score을 최대화하는 적절한 임계값을 설정하는 방법을 제안하였다. Yu 등 (2015)은 SVM을 응용한 검색 알고리즘을 통해 의사결정 임계값을 자동으로 찾아내는 연구를, Zou 등 (2016)은 최상의 $F1$-Score과 AUC를 얻기 위해 임계값을 조정하는 샘플링 기반의 임계값 자동 조정 방법을 제안했다. Voigt 등 (2014)은 오분류 비용을 알 수 없는 불균형 자료의 경우, 추정된 소수 범주에 대한 MSE를 최소화하는 임계값을 선택하는 방법을 연구하였다.

2.2. 제안 임계값 조정 방안

이 논문에서는 위에서 언급한 민감도, 특이도, 정밀도가 아닌 의사결정나무의 분지과정에서 적용되는 엔트로피와 지니지수와 같은 불순도 측도나 Chi-square automatic interaction detection (CHAID)에 적용되는 카이제곱 통계량을 이용하여 임계값을 조정하는 방법을 제안한다.

CHAID는 AID나 THeta automatic interaction detection (THAID)를 개선한 알고리즘으로 Kass (1980)에 의해 발표된 의사결정나무의 일종이다. 모형 적합에 사용된 표본크기를 n이라고 하고, 임계값을 t로 설정했을 때 분류된 자료의 혼동행렬을 아래와 같이 표시하자.

	분류 N	분류 P
실제 N	TN(t)=$O_{00}(t)$	FP(t)=$O_{01}(t)$
실제 P	FN(t)=$O_{10}(t)$	TP(t)=$O_{11}(t)$

각 범주에 속하는 관측크기를 $O_{i+}(t) = O_{i0}(t) + O_{i1}(t)$, 각 범주로 분류된 크기를 $O_{+j}(t) = O_{0j}(t) + O_{1j}(t)$라고 하면, 카이제곱통계량은 다음과 같이 정의된다.

$$\chi^2(t) = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{\left(O_{ij}(t) - \hat{E}_{ij}(t)\right)^2}{\hat{E}_{ij}(t)} = \frac{n\,(O_{00}(t)O_{11}(t) - O_{01}(t)O_{10}(t))^2}{O_{0+}(t)O_{1+}(t)O_{+0}(t)O_{+1}(t)}.$$

여기서 $\hat{E}_{ij}(t) = O_{i+}(t)O_{+j}(t)/n$로 범주형자료에 대한 동질 또는 독립 귀무가설 하에서 해당 칸의 기대도수 추정값이고 $O_{+j}(t) = 0$이면 $\chi^2(t) = 0$으로 정의한다. 위 표에서 대각원소의 빈도가 커진다는 것은 그만큼 분류가 잘 된다는 것을 의미하는데 이 경우 $\chi^2(t)$는 커진다. 따라서 이 통계량이 커지면 분류가 잘 되었다고 볼 수 있으므로 이 통계량을 최대로 만드는 임계값 t를 선택하는 방법을 고려할 수 있다. 또한 이 통계량은 기대값이 작은 칸에서의 영향력을 상대적으로 많이 반영하기 때문에 불균형 자료에서 소수 범주의 분류 결과를 많이 반영하는 형태를 가진다.

지니지수는 Brieman 등 (1984)의 CART에서 적용된 불순도 측도이며 위의 혼동행렬의 값을 이용하여 지니지수 $G(t)$를 표시하면 다음과 같다.

$$G_j(t) = 1 - \sum_{i=0}^{1}\left(\frac{O_{ij}(t)}{O_{+j}(t)}\right)^2, \qquad G(t) = \sum_{j=0}^{1}\frac{O_{+j}(t)}{n}G_j(t).$$

여기서 $O_{+j}(t) = 0$이면 $G_j(t) = 1$로 정의한다. 두 그룹의 분류가 잘되면 $G_j(t)$는 0에 가까워지기 때문에 $G(t)$가 작으면 작을수록 두 그룹으로 분류가 잘 되었다고 볼 수 있다.

*ID*3 (Quinlan, 1986), *C*4.5 (Quinlan, 1993)와 *C*5.0 등의 알고리즘에서 사용된 엔트로피 $E(t)$는 의사결정나무의 대표적인 불순도 측도로 다음과 같이 계산된다.

$$E_j(t) = -\sum_{i=0}^{1}\left(\frac{O_{ij}(t)}{O_{+j}(t)}\right)\log_2\left(\frac{O_{ij}(t)}{O_{+j}(t)}\right), \qquad E(t) = 1 - \sum_{j=0}^{1}\frac{O_{+j}(t)}{n}E_j(t).$$

여기서 $O_{+j}(t) = 0$이면 $E_j(t) = 1$로 정의하며 두 분류의 분류가 잘되면 $E_j(t)$는 0에 가까운 값이 되며 $E(t)$가 커지면 커질수록 두 그룹으로 분류가 잘 되었다고 볼 수 있다. 본래 엔트로피에서는 $E(t)$의 오른쪽 식에 있는 1 대신 분류 전의 엔트로피 값을 사용하지만 이 문제에서는 해당 부분이 없기 때문에 1로 대체하였다. 이 논문에서는 $\chi^2(t)$나 $E(t)$와 동일한 방향성을 가지도록 $G(t)$ 대신 $\underline{G}(t) = 1 - G(t)$를 사용하며 이들은 분류임계함수(classifying threshold function, CTF)라고 부르겠다.

2.3. 임계값 조정 방법 적용

샘플링 방법에서는 샘플링을 위해 원자료가 있어야 하지만 임계값 조정 방법에서는 원자료가 있는 경우뿐만 아니라 표본에 대해 각 범주에 속할 확률만 있어도 사용할 수 있는 장점이 있다. 원자료가 있는 경우에도 모형추정을 통해 각 자료에 대해 범주에 속할 확률을 추정하고 결국에는 이 분류확률을 이용하기 때문에 추정확률만 있는 경우를 가정하고 설명하고자 한다. 임계값 방법을 실제 분석에 적용할 때에는 다음과 같이 0과 1사이의 적용가능한 임계값 집합을 먼저 구성한다.

$$\mathcal{T} = \{t_1, t_2, \ldots, t_T\}.$$

각 t에 대해 $\chi^2(t)$, $\underline{G}(t)$, 또는 $E(t)$를 계산하여 해당 분류임계함수를 가장 크게하는 t를 선택한다. 경우에 따라서로 다른 t값에 대해 동일한 혼동행렬을 유지하여 분류임계함수의 값이 동일한 경우가 나올 수 있다.만약 최적 분류임계함수 값이 여러 개 나오는 경우에는 해당 임계값들 중 가운데 있는 값을 최적 임계값으로 선택하는 것도 한 방법이 될 수 있다. 왜냐하면 실증분석에서의 Figure 1에서 보듯이 대부분의 분류임계함수는 일봉(uni-modal) 형태를 가지기 때문이다. 하지만 이에 대한 이론적 근거는 없기 때문에 이 논문에서의 이 중간값 대신 일종의 붓스트랩 방법을 통해 보다 안정적으로 임계값을 선택하는 방법을 제안한다.

일반적인 붓스트랩 방법에서는 재표집된 표본으로 모형을 다시 추정하여 분류확률을 계산하는데 표본크기가 큰 경우 모형추정에 시간이 많이 소요되거나 부가적인 작업이 부담스러운 경우도 있을 수 있다. 그래서 이 논문에서는 원자료를 이용하여 계산된 분류확률에서 무작위로 n개를 추출하여 결과를 도출한 것을 추천한다. 실증분석에서도 표본을 재표집한 결과와 확률을 재표집한 결과 간에 차이가 거의 없었고 훨씬 수월하게 분석을 진행할 수 있는 장점이 있다. 추출한 분류확률에 대해 $\mathcal{T}$의 각 임계값을 적용하여 혼동행렬을 구한 후 분류임계 측도를 계산한다. 예를 들어, 분류임계함수로 $\chi^2(t)$를 사용한다고 하고 j번째 표집에 대해 i번째 임계값을 적용한 분류임계함수를 $\chi_j^2(t_i)$라고 하자. 이러한 작업을 B번 반복하고 각 t별로 B개의 분류임계 측도값의 평균을 계산한다.

$$\bar{\chi}^2(t_i) = \frac{1}{B}\sum_{j=1}^{B}\chi_j^2(t_i), \qquad i = 1, 2, \ldots, T.$$

그리고 이 값을 최대로 만드는 $t^* = \arg\max \bar{\chi}^2(t)$를 선택한다. 만약 이런 과정을 거쳤는데도 불구하고 최적 임계값이 여러개 나오는 경우에는 이들 중 무작위 선택하거나 가운데에 있는 임계값을 선택할 수 있다.

3. 임계값 조정 실증분석

이 논문에서 제안하고자 하는 방법은 분류모형 간의 성능을 비교하는 것이 아니라 각 모형에서 추정된 분류확률에 대해 임계값을 조정하여 분류 성능을 개선하는 것이기 때문에 어떤 분류모형을 사용했는지는 중요하지

않다. 이에 실험에서는 이항 또는 이진 분류의 기본 모형인 로지스틱 회귀모형을 이용하여 분류확률을 추정하고 임계값 0.5 대비 분류임계함수에서 유도된 임계값 기준의 결과가 분류성능측도 관점에서 어떤 개선이 이루어졌는지 비교하고자 한다.

불균형 비율(imbalanced ratio, IR)은 소수 범주의 개체 수(#(P))에 대한 다수 범주의 개체 수(#(N))의 비율로 #(N)/#(P)로 얼마나 불균형이 심하지를 표시하는 측도이다. 실증실험에서는 Kaggle 등에서 제공하는 다양한 불균형 비율을 가지는 데이터를 적용해 보았다. 이들 자료의 구조는 다음과 같이 정리할 수 있다.

Dataset	Length	Attributes	#(N)	#(P)	IR
Creditcard	284807	30	284315	492	577.9
Mammography	11183	6	10923	260	42.0
Forestcover	38501	54	35754	2747	13.0
Satimage	6430	36	5805	625	9.3

Creditcard는 유럽의 카드소지자에 의해 2013년 9월 중 이틀동안 이루어진 온라인 거래에서 정상과 사기 여부에 대한 데이터로 극단적인 불균형 자료 분류 연구에서 많이 언급되는 대표적인 데이터 중 하나이다. Mammography는 Woods 등 (1993)이 유방촬영술을 통해 유방에 석회화 된 부분이 있는지 여부를 검출하는 패턴인식 방법을 비교하기 위해 사용한 데이터이다. Blake와 Merz (1998)에 등록되어 있는 Forestcover는 본래 581,012개 자료로 7개의 범주로 구성되어 있으나 Chawla 등 (2002) 분석할 때 사용한 것과 같이 2개의 범주(Ponderosa Pine(35754개)와 Cottonwood/Willow(2747개))로만 재구성하여 분석하였다. Satimage는 위성사진에서 주위에 있는 3×3 픽셀의 다중 스펙트럴 값을 기반으로 중심픽셀을 판별하기 위한 데이터로 Blake와 Merz (1998)에 등록되어 있으며 본래는 6개의 범주로 구성되어 있는데 Chawla (2003)와 같이 가장 적은 범주를 소수범주, 나머지를 묶어 다수범주로 처리하여 분석하였다.

임계값 집합은 $\mathcal{T} = \{0.01, 0.02, \ldots, 0.99\}$로 0.01부터 0.99까지 0.01씩 증가시킨 99개의 임계값으로 구성하였다.

3.1. 전체자료 분석

먼저 동일한 최적 분류임계함수(classifying threshold function, CTF) 값을 가지는 임계값이 어떻게 구성되어 있는지를 알아보기 위해 제안방법을 원자료에 적용해 보았다. 분류임계함수 값을 시각적으로 비교하기 위해 분류임계함수별로 최소값(min)과 최대값(max)을 이용하여 정규화, (함수값-min)/(max – min) 하여 0과 1사의 값으로 만들어 최적 임계값의 위치와 전체 패턴을 비교해 보았다. 흥미로운 것은 네 데이터 모두 모든 t에 대해 $\chi^2(t)$와 $\underline{G}(t)$의 정규화 값이 동일한 것으로 나타났으며 확인결과 $\chi^2(t)$와 $\underline{G}(t)$의 상관계수가 1인 것으로 나타났다. 실제로 이진분류에서 두 분류임계함수의 결과 항상 동일한지는 더 연구를 진행할 필요가 있으며 이러한 결과 때문에 $\chi^2(t)$와 $\underline{G}(t)$를 같이 표시할 필요가 없어 이 논문에서는 $\chi^2(t)$와 $E(t)$를 기준으로 결과를 비교했다. Figure 1는 네 데이터에 대해 정규화한 두 분류임계함수를 비교한 것으로 분류임계함수에 따라 최적 임계값에 차이가 있는 것으로 볼 수 있다.

분석결과 모든 데이터의 최적임계값은 하나만 있는 것으로 나타났으며 이를 Table 1에 'Threshold' 열에 표시하였다. 2절에서 설명 각 분류성능 측도를 계산하고 각 데이터별로 균형정확도(BA), G-Mean, $F1$-score가 가장 큰 값에는 밑줄을 표시하였다. Satimage 데이터를 제외한 나머지 데이터에서는 엔트로피의 BA와 G-Mean 결과가 좋은 것으로 나타났으며 $F1$-score 관점에서는 χ^2 방법이 전체 데이터에서 우수한 것으로 나타났다. Satimage에서는 임계값을 0.5에서 최적 임계값으로 낮추었을 때 특이도가 0.9에서 0.5대로 낮아진

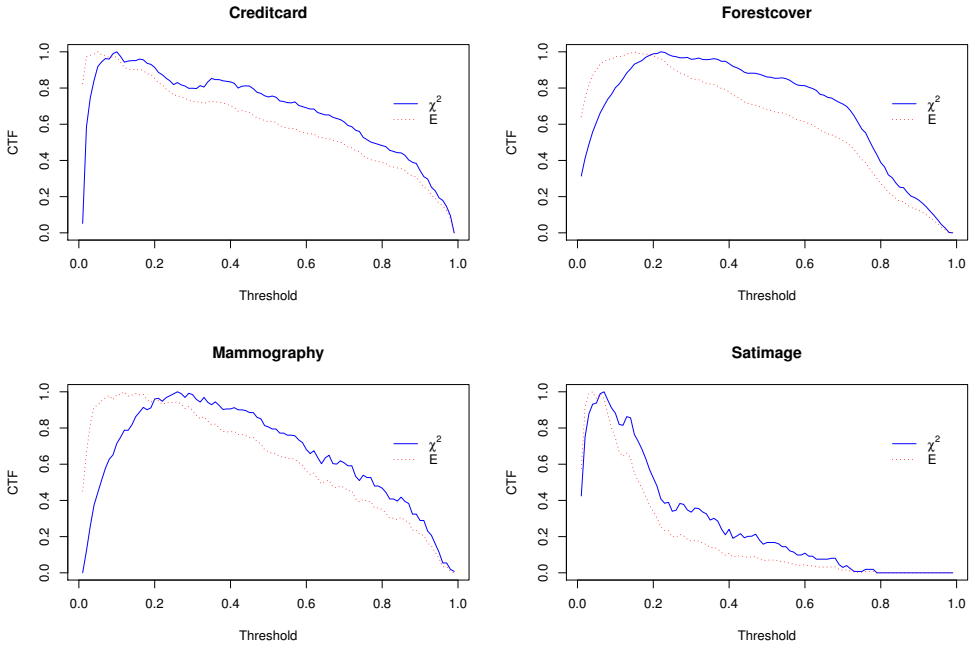

Figure 1: *Comparisons of $\chi^2(t)$ and $E(t)$.*

Table 1: Performance metrics of the proposed methods

Dataset	Method	Threshold	Accuracy	Sensitivity	Specificity	Precision	BA	G-Mean	$F1$-score
Creditcard	0.5	0.50	0.9992	0.6260	0.9999	0.8800	0.8129	0.7912	0.7316
	χ^2	0.10	0.9993	0.7907	0.9996	0.7939	0.8951	0.8890	0.7923
	E	0.05	0.9992	0.8171	0.9995	0.7376	0.9083	0.9037	0.7753
Forestcover	0.5	0.50	0.9463	0.4492	0.9845	0.6902	0.7169	0.6650	0.5442
	χ^2	0.22	0.9316	0.7157	0.9482	0.5151	0.8320	0.8238	0.5990
	E	0.15	0.9129	0.7991	0.9216	0.4393	0.8603	0.8582	0.5669
Mammography	0.5	0.50	0.9837	0.4038	0.9975	0.7955	0.7007	0.6347	0.5357
	χ^2	0.26	0.9836	0.5731	0.9934	0.6742	0.7832	0.7545	0.6195
	E	0.12	0.9739	0.7000	0.9804	0.4596	0.8402	0.8284	0.5549
Satimage	0.5	0.50	0.9039	0.0240	0.9986	0.6522	0.5113	0.1548	0.0463
	χ^2	0.07	0.5678	0.9344	0.5283	0.1758	0.7314	0.7026	0.2959
	E	0.04	0.5151	0.9760	0.4655	0.1643	0.7207	0.6740	0.2812

반면, 민감도는 0.024에서 0.9대로 급격히 증가하는 것을 볼 수 있으며 나머지 데이터에서는 민감도는 급격히 증가하면서도 특이도는 조금 낮아지는 결과를 보이고 있다.

비록 현재 데이터에서는 최적임계값이 하나만 나왔으나 여러 개의 최적 임계값이 존재하는 경우를 대비하여 붓스트랩을 통해 최적 임계값을 유도하는 방법을 적용한 실험을 적용해 보았다. 반복추출을 통한 분석에서 Creditcard 자료의 경우 표본크기가 매우 커 동일한 수만큼 무작위 추출하여 모형을 적합하는데 시간이 많이 걸려 원자료에서 재표집하는 대신 원자료의 분류확률에서 무작위로 동일한 수만큼의 추정확률을 추출하여

Table 2: Optimal thresholds based on Bootstrap

Bootstrap	Method	Creditcard	Forestcover	Mammography	Satimage
Probability	χ^2	0.10	0.22	0.26	0.07
	E	0.05	0.15	0.12	0.04
Data	χ^2	-	0.22	0.26	0.07
	E	-	0.15	0.14	0.04

Table 3: Mode of optimal thresholds for train data

Method	Creditcard	Forestcover	Mammography	Satimage
χ^2	0.09	0.22	0.26	0.06
E	0.05	0.15	0.14	0.04

성능을 비교하였다. Mammography, Forestcover, Satimage의 경우에는 Creditcard에 적용한 방법과 더불어 원자료에서 동일 수만큼 표본을 무작위 추출하여 모형을 새로 적합하여 각 임계값별로 분류임계함수를 계산하는 방법도 적용해 보았다.여기서 반복수는 $B = 1,000$으로 설정하였다. 즉, 각 임계값에 대해 분류임계함수의 값이 1000개가 생성되고 이 1000개 값의 평균을 해당 임계값의 분류임계함수 값으로 사용한다. Table 2에는 각 데이터별로 평균 분류임계함수 값 중 가장 큰 평균을 가지는 임계값을 선택한 것이다. 표에서 Bootstrap이 Probability로 표시된 부분은 분류확률을 무작위 추출한 방법의 결과이고 Data로 표시된 부분은 전체자료에서 매번 표본을 무작위 추출하여 모형을 추정하는 방법의 결과이다. 실험결과를 보면 Mammography의 Data에서 엔트로피가 0.12가 0.14로 바뀐 것을 제외하고 Table 1의 최적임계값과 동일한 것을 볼 수 있으며 모형적합 후 추정된 확률에서 붓스트랩하여 임계값을 선택해도 문제가 없는 것을 볼 수 있다.

3.2. 훈련자료와 검증자료 분석

분류 모형을 만들고 검증하기 위해 원자료에서 매 반복마다 70% 훈련자료(train data)의 비복원 추출하고 나머지 30%는 검증자료(test data)로 사용하였다. 이를 통해 일종의 out-of-bag (OOB) 오류가 얼마나 되는지를 알아보고자 하였다. 이 분석에서는 앞의 실험과 다르게 원자료에서 훈련자료와 검증자료를 추출하였으며 추출과정에서 원자료 내에서 소수범주와 다수범주의 비율을 유지하기 위해 다수범주의 70%, 소수범주의 70%가 훈련자료에 포함되도록 일종의 층화추출을 적용하였다. 모형적합에 사용되는 표본크기가 줄어들어 Creditcard 자료에 대해서도 표본을 직접 주출하는 방법을 적용하여 분석하였다. 이 분석에도 반복수는 $B = 1,000$으로 설정하였으며 매 반복마다 훈련자료로 최적의 임계값을 각각 구하여 검증자료에 적용하여 분류성능측도를 계산하였다. 각 반복 실험에서 최적분류임계값이 여러개 나오는 경우 가운데에 있는 값을 최적임계값으로 설정하였다.

Table 3은 각 훈련자료에 대해 선택된 1,000개의 최적임계값 중 빈도가 가장 많은 최빈값으로 Table 1과 Table 2의 결과와 차이가 거의 없는 것을 확인할 수 있다.

각 반복마다 훈련자료에 의해 선택된 최적임계값을 기준으로 검증자료를 분류하여 혼동행렬을 유도하고 이를 기반으로 분류성능측도를 계산하였으며 Table 4는 1,000개의 분류성능측도의 평균을 표시한 것이다. Table 1과 비교해 보면 전반적인 패턴은 동일하며 BA, G-Mean, $F1$ score에서는 최대 0.0251정도 조금 낮게 나와 큰 차이를 보이지 않고 있다. Satimage를 제외한 데이터에서는 BA와 G-Mean 기준에서 엔트로피가, $F1$-score 기준에서는 카이제곱 방법의 결과가 좋은 것으로 나타나 원자료 전체로 분석했을 때와 동일한 결과를 얻었다.

현재 분류임계값 조정 방법은 다음과 같은 앙상블 기법을 적용할 수 있다.

Table 4: Performance metrics for test data

Dataset	Method	Accuracy	Sensitivity	Specificity	Precision	BA	G-Mean	$F1$-score
Creditcard	0.5	0.9992	0.6212	0.9998	0.8725	0.8105	0.7877	0.7247
	χ^2	0.9992	0.7666	0.9996	0.7836	0.8831	0.8752	0.7743
	E	0.9991	0.8070	0.9994	0.7083	0.9032	0.8979	0.7526
Forestcover	0.5	0.9366	0.4466	0.9743	0.6726	0.7104	0.6587	0.5308
	χ^2	0.9218	0.6885	0.9397	0.5021	0.8141	0.7987	0.5889
	E	0.9031	0.7727	0.9131	0.4274	0.8429	0.8339	0.5579
Mammography	0.5	0.9837	0.4119	0.9973	0.7856	0.7046	0.6397	0.5384
	χ^2	0.9832	0.5521	0.9935	0.6721	0.7728	0.7396	0.6038
	E	0.9727	0.6730	0.9798	0.4650	0.8264	0.8107	0.5409
Satimage	0.5	0.9034	0.0227	0.9979	0.5563	0.5103	0.1467	0.0437
	χ^2	0.5658	0.9075	0.5291	0.1717	0.7183	0.6910	0.2884
	E	0.5180	0.9652	0.4700	0.1637	0.7176	0.6730	0.2799

1. 원자료 또는 분류확률을 재표집하여 분류임계함수 기준으로 최적임계값을 B개 구한다.

$$\mathcal{T}^* = \{t_1^*, t_2^*, \ldots, t_B^*\}$$

2. 새로운 관측자료에 대한 분류확률을 계산한다.
   - 원자료를 재표집한 경우: 각 추정된 모형을 이용하여 분류확률 $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_B$ 계산
   - 분류확률을 재표집한 경우: 원자료로 추정한 모형을 이용하여 분류확률 $\hat{p}$ 계산
3. 분류확률이 최적임계값보다 큰 비율이 아래와 같이 50% 이상이면 해당 관측값은 소수범주, 미만이면 다수범주로 분류한다. 여기서, $I(\cdot)$는 지시함수를 의미한다.
   - 원자료를 재표집한 경우: $B^{-1}\sum_{b=1}^{B} I(t_b^* < \hat{p}_b) \geq 0.5$
   - 분류확률을 재표집한 경우: $B^{-1}\sum_{b=1}^{B} I(t_b^* < \hat{p}) \geq 0.5$

원자료를 재표집하는 경우에는 모형추정 과정에서 변수선택 등의 방법을 추가로 적용할 수도 있다.

4. 결론

이 논문에서는 이항 불균형 자료에 대해 CHAID의 카이제곱통계량, CART의 지니지수, $C4.5$의 엔트로피와 같은 불순도 지수를 활용하여 분류 기준점을 조정하는 방법을 제안하였다. 이 분류임계함수는 소수범주의 개체수가 작으면 보다 가중될 수 있는 형태를 가지고 있어 소수범주 자료를 소수범주로 분류하는 것이 상대적으로 더 중요한 경우에 사용될 수 있다. 제안 방법은 개별 관측값에 대해 분류 확률을 얻을 수 있는 모든 분류모형에 적용할 수 있으며 분류확률만 있는 경우에도 적용할 수 있다. 최적임계값이 여러 개 발생하는 경우에 대한 해결방법도 제시하였다.

실증분석에서는 로지스틱 회귀모형을 적용하여 일반적으로 사용하는 0.5 기준의 분류임계값과 제안 분류임계함수의 최적임계값에 의한 분류 결과를 비교하였다. 분석에 사용된 데이터에서는 0.5보다 작은 최적임계값을 선택했으며 이를 통해 특이도는 조금 낮아진 반면 민감도는 현저히 커져 균형정확도, G-Mean, $F1$-score 등이 상승하는 것으로 나타났다. 엔트로피는 균형정확도와 G-Mean 기준에서, 카이제곱 통계량과 지니지수는 $F1$-score 기준에서 더 좋은 임계값을 제공하는 경향이 있었다. 실증분석 과정에서 카이제곱 분류임계함수의 값과 지니 분류임계함수의 값 간의 상관계수가 1인 것으로 나타났는데 우연인지 아니면 필연인지에 대한 연구가 필요한 것으로 보인다.

References

Akosa J (2017). Predictive accuracy: a misleading performance measure for highly imbalanced data, *SAS Papers*, **942**, 1–12.

Blake C and Merz C (1998). *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine.

Breiman L, Friedman JH, Olshen RA, and Stone CJ (1984). *Classification and Regression Trees*, Chapman & Hall, New York.

Chawla NV, Bowyer KW, Hall LO, and Kegelmeyer WP (2002). SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence research*, **16**, 321–357.

Chawla NV, Lazarevic A, Hall LO, and Bowyer KW (2003). SMOTEboost: improving prediction of the minority class in boosting. In *7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 107–119.

Collell G, Prelec D, and Patil KR (2018). A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data, *Neurocomputing*, **275**, 330–340.

Espindola RP and Ebecken N (2005). On extending f-measure and G-Mean metrics to multi-class problems, *WIT Transactions on Information and Communication Technologies*, **35**, 25–34.

Kass GV(1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, **29**, 119–127.

Kim D, Kang S, and Song J (2015). Classification analysis for unbalanced data, *The Korean Journal of Applied Statistics*, **28**, 495–509.

Kim HY and Lee W (2017). On sampling algorithms for imbalanced binary data: performance comparison and some caveats, *The Korean Journal of Applied Statistics*, **30**, 681–690.

Longadge R, Dongre SS, and Malik L (2013). Class imbalance problem in data mining: review, *International Journal of Computer Science and Network*, **2**.

Quinlan JR (1986). Induction of decision trees, *Machine Learning*, **1**, 81–106.

Quinlan JR (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

Voigt T, Fried R, Backes M, and Rhodc W (2014). Threshold optimization for classification in imbalanced data in a problem of gamma-ray astronomy, *Advances in Data Analysis and Classification*, **8**, 195–216.

Woods K, Doss C, Bowyer K, Solka J, Priebe C, and Kegelmeyer P (1993). Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography, *International Journal of Pattern Recognition and Artificial Intelligence*, **7**, 1417–1436.

Yu H, Mu C, Sun C, Yang W, Yang X, and Zuo X (2015). Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data, *Knowledge-Based Systems*, **76**, 67–78.

Zou Q, Xie S, Lin Z, Wu M, and Ju Y (2016). Finding the best classification threshold in imbalanced classification, *Big Data Research*, **5**, 2–8.

Received June 15, 2021; Revised July 9, 2021; Accepted July 15, 2021

불균형 자료에서 불순도 지수를 활용한 분류 임계값 선택

장서인[a], 여인권[1,a]

[a]숙명여자대학교 통계학과

요 약

이 논문에서는 불균형 자료에 대한 분류 분석에서 불순도지수를 이용하여 임계값을 조정하는 방법에 대해 알아본다. 이항자료에 대한 분류에서는 소수범주를 Positive, 다수범주를 Negative라고 하면, 일반적으로 사용하는 0.5 기준으로 범주를 정하면 불균형 자료에서는 특이도는 높은 반면 민감도는 상대적으로 낮게 나오는 경향이 있다. 소수범주에 속한 개체를 제대로 분류하는 것이 상대적으로 중요한 문제에서는 민감도를 높이는 것이 중요한데 이를 분류기준이 되는 임계값을 조정을 통해 높이는 방법에 대해 알아본다. 기존연구에서는 G-mean이나 F1-score와 같은 측도를 기준으로 임계값을 조정했으나 이 논문에서는 CHAID의 카이제곱통계량, CART의 지니지수, C4.5의 엔트로피를 이용하여 최적임계값을 선택하는 방법을 제안한다. 최적임계값이 여러 개 나올 수 있는 경우 해결방법을 소개하고 불균형 분류 예제로 사용되는 데이터 분석을 통해 0.5를 기준으로 ?(무엇?)을 때와 비교하여 어떤 개선이 이루어졌는지 등을 분류성능측도로 알아본다.

주요용어: 불균형 자료, 이항 분류, 분류 기준점 조정, 불순도 지수

[1]교신저자: (04310) 서울시 용산구 청파로47길 100, 숙명여자대학교 통계학과. E-mail: inkwon@sookmyung.ac.kr