The Korean Journal of Applied Statistics
2021, Vol. 34, No. 4, 599–609

DOI: https://doi.org/10.5351/KJAS.2021.34.4.599
Print ISSN1225-066X / Online ISSN 2383-5818

Discriminant analysis for unbalanced data using HDBSCAN

Bo-Hui Lee[a], Tae-Heon Kim[b], Yong-Seok Choi[1,b]

[a]Department of Advertising and Public Relations, Silla University;
[b]Department of Statistics, Pusan National University

Abstract

Data with a large difference in the number of objects between clusters are called unbalanced data. In discriminant analysis of unbalanced data, it is more important to classify objects in minority categories than to classify objects in majority categories well. However, objects in minority categories are often misclassified into majority categories. In this study, we propose a method that combined hierarchical DBSCAN (HDBSCAN) and SMOTE to solve this problem. Using HDBSCAN, it removes noise in minority categories and majority categories. Then it applies SMOTE to create new data. Area under the roc curve (AUC) and F1 scores were used to compare performance with existing methods. As a result, in most cases, the method combining HDBSCAN and synthetic minority oversampling technique (SMOTE) showed a high performance index, and it was found to be an excellent method for classifying unbalanced data.

Keywords: unbalanced data, discriminant analysis, HDBSCAN, SMOTE

1. 서론

현대사회에서는 정상 제품과 불량 제품을 분류하는 문제나 종양이 악성인지 아닌지 분류하는 문제와 같이 여러 가지 분류 문제들이 존재한다. 분류 문제에서 사용되어지는 자료들은 균형이 맞는 자료도 있지만 균형이 맞지 않는, 즉 군집 간의 비율의 차이가 큰 자료들을 자주 볼 수 있다. 예를 들면 사람들의 당뇨병 여부 문제에서 당뇨병 검사 결과가 음성인 사람은 많고, 양성인 사람은 적어 대표적인 불균형자료라고 할 수 있다. 다른 예로는 스팸메일 여부, 물품의 불량 판별 여부, 온라인 거래가 사기인지 아닌지 판별 여부 등이 있다. 불균형 자료는 모델의 성능을 저하하는 요인으로 알려져 있는데 불균형자료를 분석하는 많은 경우에 작은 군집이 큰 군집으로 오분류 되는 특징을 보인다. 이는 의사결정 나무(decision tress, DT)와 로지스틱 회귀(logistic regression, LR) 같은 분석 방법에서 작은 군집의 특징을 노이즈(noise)로 간주해 종종 무시되기 때문이다.

일반적으로 불균형 자료의 경우 큰 군집을 작은 군집으로 오분류 하는 것 보다, 그 반대 경우가 더 큰 손실(loss)을 불러일으키는 경우가 많다. 예를 들면, 병이 없는 사람을 병이 있다고 진단하는 것 보다 병이 있는 사람을 병이 없다고 진단하는 것이 일반적으로 더 큰 문제를 발생시킨다. 기존 연구들은 이러한 문제를 해결하기 위한 방법으로 군집의 균형을 맞추는 기법인 오버샘플링(oversampling), 언더샘플링(undersampling)과 이 둘을 결합한 하이브리드샘플링(hybridsampling) 기법을 이용하여 성능을 높이려는 시도가 활발하게 진행되고

This research was supported by the BK21 FOUR funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

[1] Corresponding Author: Department of Statistics, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-Gu, Busan 46241, Korea. E-mail: yschoi@pusan.ac.kr

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

있다. 이런 샘플링 기법을 사용한 방법론 중에서 Ijaz 등 (2018)은 밀도 기반 군집분석(density based spatial clustering of application with noise, DBSCAN)과 Chawla 등 (2002)이 제안한 synthetic minority oversampling technique (SMOTE)를 결합한 방법을 제안하였다.

본 연구에서는 군집의 크기가 다른 경우에 효과적으로 군집을 형성해 주는 DBSCAN을 개선한 알고리즘인 계층적 밀도 기반 군집분석(hierarchical DBSCAN, HDBSCAN)을 통해 군집별로 노이즈를 탐색하고 제거한 후, 오버샘플링 기법을 적용했을 때 분류 모델 성능에 영향을 미치는지 실험해 보고자 한다. 작은 군집의 자료를 제거함으로써 중요한 정보를 잃을 수 있지만, 작은 군집의 노이즈를 제거한 후 오버샘플링을 진행한다면 양질의 새로운 자료를 얻을 수 있는 장점이 있다. 판별 분석을 위해 실험에 사용된 모델은 많은 곳에서 사용되고 있는 로지스틱 회귀, 의사결정 나무와 랜덤 포레스트(random forest, RF)를 사용하였다. 분석에 사용한 자료는 미국 버지니아 주에 거주하는 아프리카계 미국인들의 비만, 당뇨병 및 기타 심혈관 위험 요인을 알기 위해 얻어진 당뇨병에 대한 자료와 knowledge extraction evolutionary learning (KEEL)에서 제공하는 불균형자료 4개를 이용하였다. 실험에 사용한 자료는 모두 이진 분류 문제이며 평가지표로는 area under the roc curve (AUC)와 F1 점수를 사용하였다.

2. 선행연구

2.1. 불균형자료

불균형자료 문제 해결을 위한 방법으로는 훈련자료에서 손실을 계산할 때 가중치를 조절하여 특정 군집에 더 큰 손실 값을 갖도록 하는 가중치 조정 방법과 자료의 일부를 제거하거나 가상의 값을 새로 만드는 재표집 방법이 있다. 재표집 방법에는 소수 범주의 자료를 추가 생성하여 다수 범주의 자료와 균형을 맞추는 오버샘플링 기법, 다수 범주의 자료를 제거하여 소수 범주의 자료와 균형을 맞추는 언더샘플링 기법과 위 두 가지 방법을 모두 사용하는 하이브리드샘플링 방법이 있다. 본 연구에서는 불균형자료 문제 해결을 위해 재표집 방법을 이용하였다. 군집 C_k의 자료행렬 X_k를 개체에 관한 행 표현과 변수에 관한 열 표현으로 나타내면 $k = 1, \ldots, g;\ i = 1, \ldots, n_k;\ j = 1, \ldots, p$에 대해 식 (2.1)과 같다.

$$X_k = (x_{kij}) = \begin{bmatrix} \mathbf{x}_{k1}^T \\ \vdots \\ \mathbf{x}_{ki}^T \\ \vdots \\ \mathbf{x}_{kn_k}^T \end{bmatrix} = \left[\mathbf{x}_{(k1)}, \ldots, \mathbf{x}_{(kj)}, \ldots, \mathbf{x}_{(kp)}\right]. \tag{2.1}$$

식 (2.1)에서 x_{kij}는 군집 C_k에 속하는 자료행렬 X_k의 i번째 개체에서 j번째 변수의 관측 값을 나타낸다. 개체인 각 행은 통계적으로 서로 독립이며 변수인 각 열은 서로 연관이 되어 있다. 행벡터 $\mathbf{x}_{ki}$는 군집 C_k에 속하는 X_k의 i번째 개체에 대한 p개 변수의 관측치를 나타내며, 열벡터 $\mathbf{x}_{(kj)}$는 군집 C_k에 속하는 X_k의 j번째 변수에 대한 n_k개의 개체들의 관측치를 나타낸다. g개의 모든 군집의 각각의 자료행렬 $X_1, \ldots, X_g$의 결합(joint) 자료행렬을 X라 하면 $i = 1, \ldots, n;\ j = 1, \ldots, p$에 대해 식 (2.2)와 같이 나타낼 수 있다.

$$X = \begin{bmatrix} X^1 \\ \vdots \\ X_k \\ \vdots \\ X_g \end{bmatrix} = (x_{ij}) = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \left[\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(j)}, \ldots, \mathbf{x}_{(p)}\right]. \tag{2.2}$$

식 (2.2)에서 x_{ij}는 자료행렬 X의 i번째 개체에서 j번째 변수의 관측 값을 나타낸다. 개체인 각 행은 통계적으로 서로 독립이며 변수인 각 열은 서로 연관이 되어 있다. 자료행렬 X의 i번째 행 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$는 i번째 개체와 p개 변수의 관측치를 나타내며 j번째 열 $\mathbf{x}_{(j)} = (x_{1j}, \ldots, x_{nj})^T$는 X의 j번째 변수에 대한 n개의 개체들의 관측치를 나타낸다. 자료행렬 X_k 중 다른 자료행렬에 비해 개체 수가 매우 많은 자료행렬과 개체 수가 매우 작은 자료행렬을 X_+와 X_-로 (2.3)과 (2.4)와 같이 정의하기로 하자.

$$X_+ = \text{Matrix of data in majority category}, \tag{2.3}$$

$$X_- = \text{Matrix of data in minority category}. \tag{2.4}$$

여기서 X_+는 n_+개의 개체와 p개의 변수, X_-는 n_-개의 개체와 p개의 변수로 이루어진 다수범주와 소수범주 자료행렬이고 $n_+ \gg n_-$을 만족하며 이는 n_+가 n_-에 비해 상대적으로 크기가 매우 큼을 의미한다.

2.2. 오버샘플링

오버샘플링 기법은 자료행렬 X_-에서 개체를 추출하여 새로운 자료를 만들고 이를 기존 자료에 더하는 과정으로 이뤄진다. 오버샘플링 기법 중 랜덤 오버샘플링은 X_-에서 개체를 추출할 때 무작위로 개체를 추출하는데 이 방법은 무작위로 개체를 복사하여 성능의 편차가 크게 나타나고, 개체의 수가 늘어나 계산에 필요한 시간이 커지거나 분류기가 과적합 할 수 있는 단점이 존재한다. 이러한 단점을 해결하기 위해 제안된 대표적인 오버샘플링 기법으로 SMOTE가 있다. SMOTE는 소수 범주에 속하는 개체에 대해 h개의 최근접 이웃을 찾고, 개체와 그 이웃과 사이에 직선을 그은 다음 직선 위의 임의의 점을 설정한 비율이 될 때까지 자료를 생성하는 기법이다. 이렇게 새로 생성된 개체들은 소수 범주 자료의 성향을 반영하고 있다. 더 나아가 Han 등 (2005)은 SMOTE를 수정한 기법인 borderline SMOTE (BSM)을 제안했다. SMOTE는 소수 범주에 속하는 모든 개체를 대상으로 기법을 적용한 반면에, BSM은 범주의 결정 영역(decision region)에 있는 개체들에 한해서만 기법을 적용시키는 방법이다. 또 다른 오버샘플링 기법으로 He 등 (2008)이 제안한 adaptive synthetic sampling (ADASYN)있다. SMOTE와 동일한 방식으로 새로운 개체를 생성한 후 노이즈를 무작위로 추가하여 점들 사이의 선에서 벗어난 더 사실적인 점을 만드는 기법이다. Table 1은 SMOTE 알고리즘을 단계별로 나타낸 것이다.

식 (2.5)에서 X_{create}는 X_-의 i번째 개체 $\mathbf{x}_{-i}$와 이것의 최근접 이웃 중 m번째 개체 $\mathbf{x}_{-\text{im}}$사이의 직선의 임의의 점을 의미한다.

2.3. HDBSCAN

DBSCAN 알고리즘은 두 개의 매개변수 군집의 크기와 최소 이웃 자료의 수에 아주 민감하게 반응하는 단점이 있다. 또한 군집들 간에 밀도가 다르다면 그에 대한 정보를 반영해 줄 수 없고, 자료들의 계층적 구조를 반영한 군집 형성이 불가능하다. 이러한 단점을 보완한 알고리즘이 McInnes와 Healy (2017)가 제안한 HDBSCAN이다. HDBSCAN 알고리즘은 계층적 군집분석 방법 중 하나인 단일연결법에서부터 시작된다.

Table 1: SMOTE algorithm

[Step 1]	Select the data X_- in the minority class.
[Step 2]	Find the i^{th} observation of X_-, h recent neighbors of $\mathbf{x}_{-i} = (x_{-i1}), \ldots, (x_{-ip})$, $i = 1, \ldots, n_-$. The m^{th} object of the h recent neighbors of $\mathbf{x}_{-i}$ is defined as $\mathbf{x}_{-im}$, $m = 1, \ldots, h$.
[Step 3]	The expression (2.5) is prepared according to the predetermined SMOTE ratio. $$X_{create} = \mathbf{x}_{-i} + \lambda(\mathbf{x}_{-im} - \mathbf{x}_{-i}),\ \lambda \in [0, 1] \quad (2.5)$$

군집분석은 유사성이 높은 개체를 하나의 군집으로 묶는 통계적 분석 방법이며 크게 계층적 군집분석과 비계층적 군집분석으로 나뉜다 (Choi, 2018). 계층적 군집분석은 개체들을 순차적으로 병합(agglomeration) 또는 분할(division)하는 과정을 통해 진행된다. 계층적 군집분석에는 단일연결법, 완전연결법, 평균연결법, 중심연결법과 중위수연결법 등이 있다. 본 연구에서 활용할 단일연결법은 다른 군집에 속한 가장 가까운 두 점 사이의 유클리드 거리를 군집 간의 거리로 측정하는 방법이다. 하지만 단일연결법은 노이즈에 민감한 단점이 있다. 이러한 단점을 해결하기 위한 방법으로 로버스트 단일연결법(robust single linkage)이 있다. 로버스트 단일연결법은 노이즈에 영향을 덜 받기 위해서 유클리드 거리 대신에 새로운 거리지표 mutual reachability (mr) 를 정의한다.

$$d_{mr}(\mathbf{x}_r, \mathbf{x}_s) = \max[\text{core}_{\ g}(\mathbf{x}_r), \text{core}_{\ g}(\mathbf{x}_s), d(\mathbf{x}_r, \mathbf{x}_s)]. \quad (2.6)$$

식 (2.6)에서 $\mathbf{x}_r$과 $\mathbf{x}_s$는 자료행렬 X의 r번째, s번째 개체를 의미하며, $\text{core}_{\ g}(\mathbf{x}_r)$는 $\mathbf{x}_r$과 $\mathbf{x}_r$의 최근접 이웃 g개 간의 유클리드 거리 중 최댓값, $d(\mathbf{x}_r, \mathbf{x}_s)$는 $\mathbf{x}_r$과 $\mathbf{x}_s$의 유클리드 거리를 의미한다. mr은 $\text{core}_g(\mathbf{x}_r)$, $\text{core}_g(\mathbf{x}_s)$, $d(\mathbf{x}_r, \mathbf{x}_s)$ 중 최댓값을 사용한다. 밀도가 높은 지점의 자료는 $\text{core}_g()$의 값이 작기 때문에 $d(\mathbf{x}_r, \mathbf{x}_s)$값을 사용하고, 밀도가 낮은 지점의 자료에서 우연히 두 개체의 거리가 가깝더라도 $\text{core}_g()$와 비교하여 거리를 좀 더 로버스트하게 만들어준다.

특히, 군집나무에 의한 군집분석에서는 자료가 큰 경우 로버스트한 단일연결법을 적용하더라도 분석하기 어려운 문제가 있다. 군집나무를 단순화시키기 위해서 최소 군집 크기를 만족하지 못하는 군집들을 노이즈라고 생각하고 제거한다. 이러한 과정을 가지치기라고 한다. 노이즈로 예상되는 가지를 쳐낸 군집나무는 가지 수가 적어지면서 군집나무가 단순화되지만 여전히 실제 군집 구조에 대한 충분한 정보를 포함하고 있다. 군집나무를 단순화 시킨 다음으로 적절한 군집을 구성시켜준다. 계층적 군집분석의 경우 적절한 군집을 구성하기 위해 적절한 mr값을 선택하여 나무를 잘라 군집을 형성하는 것이 일반적이다. 하지만 밀도가 다른 군집을 고려하고자 할 때, 적절한 mr은 군집마다 다르게 된다. 예를 들어 어떤 군집에서는 mr의 값이 0.3이 큰 값일 수 있지만 또 다른 군집에서는 크지 않은 값일 수 있다. 그러므로 군집을 선택하기 위해 다른 방법을 사용해야 한다.

밀도가 다른 군집을 이상적으로 나누는 방법은 가장 큰 거리 범위를 가지는 군집들을 선택하는 것이다. 즉 군집이 형성될 때 가장 오래 유지되고 있는 군집을 결정하는 것이다. 이를 위해서는 식 (2.7)에 의한 군집의 지속성 측정 지표가 필요하다.

$$\lambda = \frac{1}{mr}. \quad (2.7)$$

λ를 mr의 역수로 정의한 이유는 나무를 자르는 기준을 mr의 역수로 바꾸어 줌으로서 작은 거리에는 민감하게 반응하고 큰 거리에는 덜 민감하게 반응하게 만들어 국소밀도를 더 잘 표현하기 위해서이다.

군집의 지속성 측정을 위해 군집이 형성될 때와 가지치기 될 때 그리고 분할될 때의 λ값을 정의한다. 군집나무에 속해있는 군집 C_k는 자료행렬 X_k로 구성되어 있고 X_k는 n_k개의 개체와 p개의 변수를 가지고 있다. 군집 C_k가 최초로 형성될 때의 λ값을 λ_{kb}, C_k가 가지치기 될 때의 λ값을 λ_{kl}, 군집 C_k가 새로운 군집 2개로 분할될 때의 λ값을 λ_{kd}라고 정의한다. 만약에 군집 C_k의 자식 군집 C_{kc}가 최초로 형성될 때의 λ값을 λ_{kcl}이라 하고 새로운 군집 2개로 분할될 때 λ값을 λ_{kcd}로 정의한다. 앞에서 정의한 것들을 이용하여 군집 C_k의 지속성 점수(persistence score)를 식 (2.8)과 같이 정의한다.

$$PS(C_k) = \lambda_{kd} - \lambda_{kb}. \tag{2.8}$$

군집 C_k의 지속성 점수 $PS(C_k)$는 군집 C_k가 새로운 군집들로 분할될 때 가지는 λ값에서 군집 C_k가 처음으로 형성될 때의 λ값을 뺀 것이다. C_k의 자식 군집 C_{kc}들의 지속성 점수는 식 (2.9)로 정의한다.

$$\sum_c PS(C_{kc}) = \sum_c (\lambda_{\mathrm{kcd}} - \lambda_{\mathrm{kcb}}), \tag{2.9}$$

$$\sum_c PS(C_{kc}) > PS(C_k). \tag{2.10}$$

식 (2.9)는 군집 C_k의 자식군집들의 지속성 점수를 합한 것이다. 만약 식 (2.10)을 만족하면 군집을 c개로 유지하고 식 (2.10)을 만족하지 못한다면 c개의 군집 대신에 부모 군집인 C_k를 1개의 군집으로 인정한다. 다음의 과정을 군집나무에 있는 모든 군집에 진행하게 되면 $\sum_{k=1}^{g} PS(C_k)$값이 최대가 되는 g개의 군집들이 형성된다. 이 군집들은 지속성 점수가 가장 높고 가변 밀도를 반영한 군집들이다.

3. 연구 방법

불균형자료 문제 해결을 위해서 소수 범주의 자료를 늘리는 오버샘플링 기법과 다수 범주의 자료를 줄이는 언더샘플링 기법이 자주 활용되고 있다. 오버샘플링 기법으로 자주 활용되는 SMOTE는 노이즈에 민감한 단점이 있다. 이러한 점을 개선하기 위해 Ijaz 등 (2018)은 DBSCAN을 이용해 노이즈를 제거하고, 오버샘플링을 진행하는 하이브리드샘플링 방법을 제안했다. 하지만 이 방법은 다수 범주에 대한 노이즈만 제거하고 소수 범주에 대한 노이즈에 관해서는 영향을 주지 못한다. 이러한 점을 개선하기 위해서 본 연구에서는 HDBSCAN을 이용하여 노이즈를 제거하고 오버샘플링을 적용하는 방법을 사용하였다.

3.1. 전처리 자료

HDBSCAN 알고리즘은 군집별로 밀도가 다르더라도 각 군집의 노이즈를 찾아낼 수 있다. 이를 이용하여 소수 범주의 자료와 다수 범주의 자료의 노이즈를 찾아 제거하고 SMOTE 기법을 적용하여 소수 범주의 자료를 생성하였다. 성능 비교를 위해 4가지 전처리 방법을 적용한 자료를 (1)~(4)라고 정의한다.

(1) 원본 자료(Original)

(2) SMOTE를 적용한 자료(Smote)

(3) HDBSCAN을 이용해 노이즈를 제거한 자료(Hdbscan)

(4) HDBSCAN을 이용해 노이즈를 제거한 후에 SMOTE를 적용한 자료(Hdbsm)

Table 2: Confusion matrix

Actual	Predicted	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

3.2. 평가지표

대부분의 불균형자료의 경우 소수 범주에 속하는 자료를 소수 범주로 판단하는 것이 중요하다. 그렇기 때문에 본 연구에서는 평가지표로 정확도(accuracy)를 사용하는 것 보다 F1 점수를 평가지표로 사용하였다. Table 2는 모델의 성능을 평가할 때 사용되는 지표인 혼동행렬을 나타낸 것이다. TP는 예측을 Positive로 하고 실제로도 Positive인 경우를 의미하고, FP는 예측을 Positive로 했지만 실제로는 Negative인 경우를 의미한다. TN은 예측을 Negative로 하고 실제로도 Negative인 경우이고 FN은 예측을 Negative로 했지만 실제로는 Positive인 경우를 의미한다.

식 (3.1)은 정밀도를 의미하며 예측을 Positive라고 분류한 것 중에서 실제 Positive인 것의 비율이다. 식 (3.2)는 재현율을 의미하고 실제 Positive인 것 중에서 예측을 Positive라고 한 것의 비율이다. 정밀도와 재현율 모두 높으면 좋겠지만 이 둘 사이에는 서로 트레이드-오프(trade-off) 관계가 존재하고 있다. 식 (3.3)은 F1 점수를 의미하며 정밀도와 재현율의 조화평균으로 두 가지 모두 고려할 수 있다. F1 점수는 불균형자료의 경우 모델의 성능을 파악하기 위해서 AUC와 함께 정확도 대신 많이 사용되어지고 있다.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3.2}$$

$$\text{F1} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3.3}$$

Figure 1은 receiver operating characteristic curve (ROC Curve)에서의 AUC를 나타내는 그림이다. x축의 1-Specificity는 Table 2를 참고하면 FP/(FP+TN)을 의미하며 실제 Negative인 것을 Positive로 예측한 정도를 나타낸다. y축의 Sensitivity는 Recall과 동일한 의미이다. AUC는 ROC 그래프의 아래 면적을 나타내는 값이며 최대가 1 최소가 0.5로 값이 클수록 좋은 모형을 의미한다.

4. 연구 결과

본 장에서는 연구에 사용될 자료들에 대해서 소개하고 이 자료들에 4가지 전처리 방법 을 적용하여 원본 자료(Original), SMOTE를 적용한 자료(Smote), HDBSCAN을 이용해 노이즈를 제거한 자료(Hdbscan)와 HDBSCAN을 이용해 노이즈를 제거하고 SMOTE를 적용한 자료(Hdbsm)에 대해서 랜덤 포레스트(RF)와 의사결정나무(DT) 그리고 로지스틱 회귀(LR) 분류기를 적용하여 성능을 비교하였다.

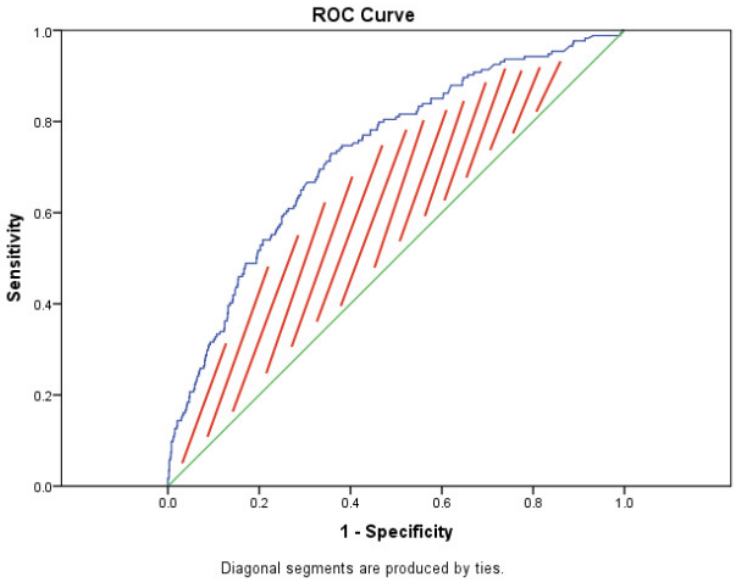

Figure 1: *AUC in ROC Curve.*

Table 3: Explanation of diabetes dataset variable

Variable	Explanation
glyhb	Glycosylated Hemoglobin (>7 : diabetes, ≤7 : normal)
stab.glu	Stabilized Glucose (mg/dL)
age	Age (years)
ratio	Cholesterol/High Density Lipoproteins (HDL) Ratio
waist	Waist (inches)
chol	Total Cholesterol (mg/dL)
bp.1s	First Systolic Blood Pressure (mmHg)

4.1. 자료

첫 번째 자료는 미국 버지니아 주에 거주하는 아프리카계 미국인들의 비만, 당뇨병 및 기타 심혈관 위험 요인을 알기 위해 얻어진 당뇨병에 대한 자료이다. 총 개체 수는 403개이고 변수는 19개로 이루어진 자료이다. RF를 이용하여 변수 중요도를 기준으로 유의미하다고 판단되는 7개의 변수를 추출하였고 Table 3에 정리되어 있다.

반응 변수가 되는 'glyhb'의 값이 7이 넘으면 양성, 넘지 않으면 음성으로 판단한다. 'glyhb' 변수에 결측치가 있는 개체 21개를 제거한 382개의 개체를 이용하였다. 'negative'인 개체는 322개, 'positive(diabetes)'인 개체는 60개로 이루어진 이진 분류 자료이다.

두 번째부터 다섯 번째 자료들은 knowledge extraction evolutionary learning (KEEL)에서 제공하는 불균형자료이다. 4가지 자료 모두 범주형 변수는 'class'이며 'negative', 'positive'로 이루어진 2진 분류 자료이다. Table 4는 자료의 특성에 대해 정리한 표이다.

Table 4: Data characteristics

Dataset	Number of Observation	Number of Variable	Number of majority category	Number of minority category	Imbalanced ratio
Diabetes	382	7	322	60	5.37
Glass	214	9	138	76	1.82
Vehicle	846	18	630	216	2.90
Yeast	1484	8	1321	163	8.10
Wisconsin	683	9	444	239	1.86

Table 5: Diabetes data classification results

Classifier	Preprocessing method	AUC	F1 score
RF	Original	0.925	0.604
	Smote	0.908	0.600
	Hdbscan	0.931	0.693
	Hdbsm	**0.932**	**0.758**
DT	Original	0.755	0.612
	Smote	0.732	0.556
	Hdbscan	0.814	0.689
	Hdbsm	**0.842**	**0.700**
LR	Original	**0.947**	0.667
	Smote	0.926	**0.680**
	Hdbscan	0.868	0.667
	Hdbsm	0.892	0.655

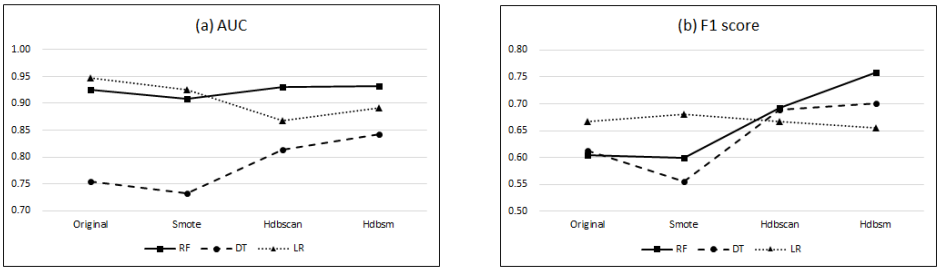

Figure 2: *Performance comparison of preprocessing methods and classifiers in diabetes data.*

4.2. 결과 및 해석

본 연구에서는 훈련 자료와 시험 자료의 비율을 7:3으로 설정하여 실험을 진행하였다. 4가지의 전처리 방법과 3가지의 분류기를 조합하여 총 12개의 결과를 비교하였다.

Table 5와 Figure 2는 Diabetes 자료에 대해 3가지의 분류기와 4가지의 전처리 방법에 대한 AUC와 F1 점수를 비교해 놓은 표와 그림이다. Figure 2의 (a)는 전처리 방법과 분류기에 따른 AUC를 비교하는 그림이고 (b)는 F1 점수를 비교하는 그림이다. Figure 2의 (a)를 보면 RF 분류기에서는 Original에 비해 Smote의 AUC는 떨어지고 Hdbscan에서는 AUC가 조금 올랐다. 이는 오버샘플링만 진행했을 때는 성능이 안 좋아졌지만 군집별로 노이즈를 제거 했을 때는 성능이 좋아졌다. Hdbsm의 경우 Hdbscan과 비교했을 때 미미하게 성능이 향상했다. 노이즈를 제거하고 오버샘플링을 적용한 것이 노이즈만 제거한 것 보다 더 좋은 성능을 보이고

Table 6: Wisconsin data classification results

Classifier	Preprocessing method	AUC	F1 score
RF	Original	0.994	**0.971**
	Smote	0.995	0.970
	Hdbscan	**0.998**	0.946
	Hdbsm	0.997	0.947
DT	Original	0.938	0.920
	Smote	**0.944**	**0.935**
	Hdbscan	0.912	0.816
	Hdbsm	0.866	0.837
LR	Original	**0.997**	**0.958**
	Smote	**0.997**	**0.958**
	Hdbscan	**0.997**	0.897
	Hdbsm	0.993	0.909

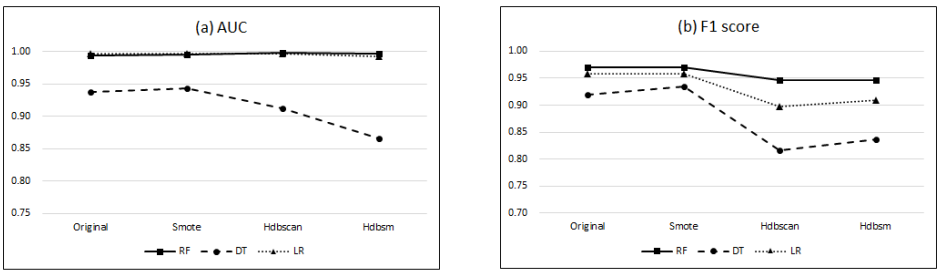

Figure 3: *Performance comparison of preprocessing methods and classifiers in Wisconsin data.*

있다. DT 분류기도 RF 분류기와 비슷한 양상을 보였으나 LR 분류기에서는 Original에 비해 오히려 AUC값이 떨어지는 것을 발견할 수 있다.

Figure 2의 (b)를 보면 RF 분류기와 DT 분류기에서 4가지 전처리 방법 중 Hdbsm에서 가장 높은 F1 점수를 볼 수 있었다. RF 분류기의 경우 Original은 Smote와 F1 점수는 차이가 거의 없었고 Hdbscan의 경우에는 F1 점수가 Original에 비해 0.089만큼 높아진 것을 볼 수 있었다. 특히 Hdbsm에서 눈에 띄게 F1 점수가 높아진 것을 볼 수 있었다. Hdbsm은 Original에 비해 F1 점수가 0.154만큼 높아진 것을 확인 할 수 있었다. 이는 자료를 그냥 사용하는 것 보다 HDBSACN 알고리즘을 적용하여 노이즈를 제거한 자료를 사용하는 것이 성능 향상에 도움이 된다는 것을 보여주는데 이 방법보다 노이즈를 제거한 자료에 SMOTE 알고리즘을 적용한 방법이 더 좋은 성능을 보이는 것을 확인하였다. Hdbsm은 AUC보다 F1 점수에서 눈에 띄게 성능이 향상하는 것을 확인할 수 있는데 이는 다른 전처리 방법보다 Hdbsm이 소수 범주의 자료를 더 잘 예측한다는 것으로 볼 수 있다.

하지만 LR 분류기의 경우 Original에서 AUC가 가장 높았고 Smote에서 F1 점수가 가장 높았다. 이 결과에서 분류기에 따라 성능 차이가 나는 단점을 발견할 수 있었다. 하지만 전체 결과 중 Hdbsm에 RF 분류기를 사용했을 때의 F1 점수가 가장 높았으므로 Diabetes 자료에서는 Hdbsm이 가장 우수한 성능을 보였다고 할 수 있다.

Table 6과 Figure 3은 Wisconsin 자료에 대해 3가지의 분류기와 4가지의 전처리 방법에 대한 AUC와 F1 점수를 비교해 놓은 표와 그림이다. Wisconsin 자료는 앞의 4가지 자료와 다르게 모든 결과가 Hdbsm 전처리 방법을 사용했을 때 성능이 떨어지는 것을 확인 할 수 있다. 이 자료는 다른 자료들과는 다르게 아무런 전처리 방법을 적용하지 않았을 때 AUC와 F1 점수가 상대적으로 높은 편이다. 즉 상대적으로 분류하기 쉬운 자료라

고 말할 수 있다. 전체 결과 중 Hdbsan에 RF 분류기를 사용했을 때 AUC가 가장 높고 Original에 RF 분류기를 사용했을 때 F1 점수가 가장 높으므로 Wisconsin 자료에서는 Hdbsm은 저조한 성능을 보였다.

5. 결론

불균형자료의 판별분석에서 다수 범주의 개체를 잘 분류하는 것 보다 소수 범주의 개체를 잘 분류하는 것이 더 중요하다. 기존의 많은 연구들은 소수 범주의 개체를 늘리는 오버샘플링 기법과 다수 범주의 개체를 줄이는 언더샘플링 기법을 많이 이용하고 있다. 하지만 소수 범주의 자료에 노이즈가 존재한다면 오버샘플링 기법으로 새로 만들어진 소수 범주의 개체가 노이즈에 영향을 받아 오히려 성능이 떨어지는 일이 빈번하다.

본 연구에서는 이를 해결하기 위한 전처리 방법으로 HDBSCAN을 활용하였다. HDBSCAN은 밀도가 다른 군집들도 잘 분류하는 특징이 있어 소수 범주의 노이즈와 다수 범주의 노이즈 모두 찾아낼 수 있는 장점이 있다. HDBSCAN을 이용하여 노이즈를 제거하고 오버샘플링 기법중 하나인 SMOTE를 적용하는 전처리 방법을 제시했다. 그 결과 총 5개의 불균형자료 중 4개의 자료에서 HDBSCAN과 SMOTE를 결합한 전처리 방법을 이용했을 때 가장 뛰어난 성능을 보였다. 특히 F1 점수에서 눈에 뛰는 성능 향상을 보였다.

본 연구의 한계점은 다음과 같다. 첫째, 모든 분류기에서 HDBSCAN과 SMOTE를 결합한 전처리 방법이 가장 좋은 성능을 보이지 않는다. 본 연구에서는 3가지의 분류기만을 사용하였는데 다른 분류기들을 사용했을 때 본 연구에서 제시한 방법보다 다른 방법이 더 높은 성능을 보일 수도 있다. 둘째, 상대적으로 분류가 쉬운 자료에 적용하면 성능이 떨어지는 경우가 있다. 노이즈로 예상되는 개체가 매우 적고 아무런 전처리 없이도 상대적으로 높은 성능을 보이는 자료에서는 성능이 떨어지는 경우가 많다. 이러한 점들이 본 연구의 한계점으로 판단된다.

본 연구에서는 이진 분류 문제에 관해서만 다루었지만 HDBSCAN은 군집이 3개 이상이더라도 각 군집의 노이즈를 찾아낼 수 있는 측면을 생각하면 다중 범주 분류에 있어서도 좋은 성능을 보일 것으로 예상된다. 또한 오버샘플링 기법으로 SMOTE가 아닌 다른 기법들과 HDBSCAN을 결합하여 성능을 평가 해보는 것도 연구해야 할 과제 중 하나이다. 본 연구에서는 개체수가 상대적으로 작은 자료에 대해서만 실험을 하였는데 개체수가 매우 큰 자료에 대해서도 실험을 해볼 필요성이 있다.

References

Chawla NV, Hall LO, Bowyer KW, and Kegelmeyer WP (2002). Smote: Synthetic minority oversampling technique, *Journal of Artificial Intelligence Research*, **16**, 321–357.

Choi YS (2018). *Multivariate Data Analysis with R*, Kyungmoon, Seoul.

Han H, Wang W, and Mao B (2005). Borderline smote: Anew over sampling method in imbalanced data sets learning. In *Proceedings of International Conference on Intelligent Computing*, 878–887.

He H, Bai Y, Garcia EA, and Li S (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of International Joint Conference on Neural Networks*, 1322–1328.

Ijaz M, Alfian G, Syafrudin M, and Rhee J (2018). Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and Random Forest, *Applied Sciences*, **8**, 1325.

McInnes L and Healy J (2017). Accelerated hierarchical density based clustering, *IEEE International Conference on Data Mining Workshops (ICDMW)*.

Received April 5, 2021; Revised April 13, 2021; Accepted April 14, 2021

불균형자료를 위한 판별분석에서 HDBSCAN의 활용

이보희,[a] 김태헌,[b] 최용석[1,b]

[a]신라대학교 광고홍보학과; [b]부산대학교 통계학과

요 약

군집간의 개체 수의 차이가 큰 자료들을 불균형자료라고 한다. 불균형자료의 판별분석에서 다수 범주의 개체를 잘 분류하는 것 보다 소수 범주의 개체를 잘 분류하는 것이 더 중요하다. 그러나 개체 수가 상대적으로 작은 소수 범주의 개체를 개체 수가 상대적으로 많은 다수 범주의 개체로 오분류하는 경우가 많다. 본 연구에서는 이를 해결하기 위해 HDBSCAN과 SMOTE를 결합한 방법을 제안한다. HDBSCAN을 이용하여 소수 범주의 노이즈와 다수 범주의 노이즈를 제거하고 SMOTE를 적용하여 새로운 자료를 만들어낸다. 기존의 방법들과 성능을 비교하기 위하여 AUC와 F1 점수를 이용하였고 그 결과 대부분의 경우에 HDBSCAN과 SMOTE를 결합한 방법이 높은 성능 지표를 보였고, 불균형자료를 분류하는데 있어 뛰어난 방법으로 나타났다.

주요용어: 불균형자료, 판별분석, HDBSCAN, SMOTE

본 논문은 교육부 및 한국연구재단의 4단계 두뇌한국(BK)21 사업으로 지원된 연구임.

[1]교신저자: (46241) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교. E-mail: yschoi@pusan.ac.kr