The Korean Journal of Applied Statistics
2024, Vol. 37, No. 4, 455–465

DOI: https://doi.org/10.5351/KJAS.2024.37.4.455
Print ISSN1225-066X / Online ISSN 2383-5818

SMOTE by Mahalanobis distance using MCD in imbalanced data

Jieun Jung[a], Yong-Seok Choi[1,a]

[a]Department of Statistics, Pusan National University

Abstract

SMOTE (synthetic minority over-sampling technique) has been used the most as a solution to the problem of imbalanced data. SMOTE selects the nearest neighbor based on Euclidean distance. However, Euclidean distance has the disadvantage of not considering the correlation between variables. In particular, the Mahalanobis distance has the advantage of considering the covariance of variables. But if there are outliers, they usually influence calculating the Mahalanobis distance. To solve this problem, we use the Mahalanobis distance by estimating the covariance matrix using MCD (minimum covariance determinant). Then apply Mahalanobis distance based on MCD to SMOTE to create new data. Therefore, we showed that in most cases this method provided high performance indicators for classifying imbalanced data.

Keywords: imbalnaced data, Mahalanobis distance, MCD, SMOTE

1. 서론

불균형 자료(imbalanced data)란, 분류 문제에서 반응변수의 값 분포가 심하게 불균형한 경우를 가리킨다. 주로 제조 분야와 같은 여러 분야에서, 정상 관측치에 비해 비정상 관측치의 수가 매우 적게 수집된다. 이러한 불균형한 자료의 분포로 인해, 비정상 관측치를 예측하는 분류 규칙은 무시되는 경향이 있다. 현실에서는 주로 소수 군집을 식별하는 것이 주요 목표인 경우가 많다. 불균형 자료를 처리하기 위해서는 기본적으로 반응변수의 분포를 조정하여 더 균형 잡힌 분포로 만드는 데 중점을 둔다 (Menardi와 Torelli, 2010). 주로 과소추출법(under-sampling)과 과대추출법(over-sampling) 두 가지 방법을 사용한다. 과소추출법은 다수 군집에 해당하는 관측치를 제거하여 학습의 계산 비용을 줄이지만, 관측치의 정보를 손실이 발생할 수 있다 (Wing 등, 2015). 반대로, 과대추출법은 소수 군집의 관측치를 증가시키는 방법을 말한다. Chawla 등 (2002)이 제안한 SMOTE (synthetic minority over-sampling technique)는 계산 비용이 적으며, 소수 군집의 관측치를 무작위로 복제하지 않아 널리 사용되어 다양한 관측치 생성 방법이 제안되었다. Abdi와 Hashemi (2016)는 유클리드 거리를 기반으로 하는 대신 마할라노비스 거리를 활용한 SMOTE인 MDO (Mahalanobis distance-based over-sampling technique)를 제안하였다.

본 연구에서는 자료의 이상치가 존재할 경우 마할라노비스 거리 계산에 미치는 영향을 고려하여 최소 공분산 행렬식(minimum covariance determinant; MCD)을 이용하여 공분산 행렬을 추정한다. 이후, 마할라노비스 거리를 기반으로 SMOTE를 적용하여 분류모형 성능 향상 여부를 확인한다.

[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 2024 / journal homepage: http://kjas.or.kr
© 2024 The Korean Statistical Society. All rights reserved.

Table 1: Data structure

Cluster	Matrix	x_1	x_2	$\cdots$	x_p
C_1	X_1	x_{111}	x_{112}	$\cdots$	x_{11p}
		x_{121}	x_{122}	$\cdots$	x_{12p}
		$\vdots$	$\vdots$		$\vdots$
		x_{1n_11}	x_{1n_12}	$\cdots$	x_{1n_1p}
C_2	X_2	x_{211}	x_{212}	...	x_{21p}
		x_{221}	x_{222}	$\cdots$	x_{22p}
		$\vdots$	$\vdots$		$\vdots$
		x_{2n_21}	x_{2n_22}	$\cdots$	x_{2n_2p}

2. 자료구조 및 분석기법

2.1. 불균형 자료

p개 확률변수들에 의한 확률벡터 $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$에 의해서 얻어진 n_1개와 n_2의 관측치를 갖는 두 개의 군집 C_1과 C_2를 고려하자. 군집 $C_g, g = 1, 2$에 속하는 자료 행렬 X_g를 식 (2.1)과 같이 정의할 수 있다.

$$X_g = \begin{bmatrix} x_{g11} & \cdots & x_{g1j} & \cdots & x_{g1p} \\ \vdots & & \vdots & & \vdots \\ x_{gi1} & \cdots & x_{gij} & \cdots & x_{kip} \\ \vdots & & \vdots & & \vdots \\ x_{gn_g1} & \cdots & x_{gn_gj} & \cdots & x_{gn_gp} \end{bmatrix} = \left(x_{gij}\right). \tag{2.1}$$

자료 행렬 X_g의 n_g개의 개체들의 관측치 중 C_1(다수 군집)에 속하는 n_1개의 개체를 가지는 자료 행렬을 X_1, C_2(소수 군집)에 속하는 n_2개의 개체로 이루어진 자료 행렬을 X_2라 정의한다. n_1이 n_2보다 상대적으로 매우 클 때 불균형 자료(imbalanced data)라 한다. Table 1은 본 연구에서 사용된 자료의 형태이다.

이러한 자료의 불균형 문제를 해결하기 위해 과소추출법 및 과대추출법의 다양한 추출방법들이 개발되었다. 이러한 방법의 목표는 군집의 개수를 균형 잡힌 분포를 갖는 자료로 조정하여, 분류모형이 군집을 더 효과적으로 분류할 수 있도록 하는 것이다.

2.2. SMOTE

Chawla 등 (2002)가 제안한 SMOTE 알고리즘은 불균형 자료에서 소수 군집의 가상 관측치를 생성하는 방법이다. SMOTE는 소수 군집의 관측치를 증가시켜 군집의 분포가 더 균형 있게 되며 관측치를 단순히 복제하지 않기 때문에 과적합 문제를 해결할 수 있다. SMOTE 알고리즘의 가상 관측치 생성 방법을 Algorithm 1에서 설명한다.

Algorithm 1의 [Step 3]에서 x_{syn}은 x_{2i}와 x_{2im} 사이의 임의의 점을 나타낸다. x_{2i}와 x_{2im} 사이에 0부터 1까지 균일하게 무작위로 선택된 값을 곱하여 가중치를 부여한다. 무작위로 선택된 가중치를 통해 다양한 관측치가 생성된다. 이러한 과정은 과대추출된 관측치 수 n_2^*가 n_1과 같아질 때까지 SMOTE 과정을 반복하게 된다. 1부터 k 사이에서 무작위로 선택된 m 값을 이용하여 가까운 이웃을 찾을 때, 식 (2.2)와 같이 유클리드 거리가 사용된다.

Algorithm 1 : SMOTE Algorithm

[Step 1] Select the data matrix X_2 belonging to C_2.

[Step 2] Choose a random integer m from 1 to the total number of observations, denoted as k.
Use the selected value of m to find the m^{th} nearest observation, denoted as $\mathbf{x}_{2im}$,
for each i^{th} observation $\mathbf{x}_{2i} = (x_{2i1}, x_{2i2}, \ldots, x_{2ip})^T, i = 1, \ldots, n_2$ in X_2

[Step 3] Observations are generated according to the defined SMOTE formula.

$$\mathbf{x}_{\text{syn}} = \mathbf{x}_{2i} + u(\mathbf{x}_{2im} - \mathbf{x}_{2i}), \quad u \in [0, 1]$$

[Step 4] Repeat [Step 2] and [Step 3] for all observations in X_2.

$$d(\mathbf{x}_{2i}, \mathbf{x}_{2im}) = \|\mathbf{x}_{2i} - \mathbf{x}_{2im}\| = \sqrt{(\mathbf{x}_{2i} - \mathbf{x}_{2im})^T (\mathbf{x}_{2i} - \mathbf{x}_{2im})}. \tag{2.2}$$

2.3. 최소 공분산 행렬식

식 (2.2)는 두 관측치 사이의 유클리드 거리를 계산한 것이며, 식 (2.3)으로 정의되는 마할라노비스 거리는 각 변수 간의 상관관계를 고려한 거리이다.

$$m(\mathbf{x}_{2i}, \mathbf{x}_{2im}) = \|\mathbf{x}_{2i} - \mathbf{x}_{2im}\|_{S^{-1}} = \sqrt{(\mathbf{x}_{2i} - \mathbf{x}_{2im})^T S^{-1} (\mathbf{x}_{2i} - \mathbf{x}_{2im})}. \tag{2.3}$$

식 (2.3)의 S는 자료 행렬 X의 공분산 행렬이다. 여기서 특히, 자료가 비대칭 분포이거나 이상치가 포함된 자료일 경우 위와 같은 전통적인 방식으로 마할라노비스 거리를 계산하게 된다면 이상치를 정상 관측치와 동일하게 판별되는 감춤효과(masking effect)가 발생한다. 감춤효과를 제거하기 위해서는 평균벡터와 공분산 행렬의 로버스트 추정량을 활용해 마할라노비스 거리를 계산해야 한다 (Rousseeuw와 van Zomeren, 1990).

이를 위해 지금까지 제안된 대표적인 방법은 Aelst와 Rousseeuw (2009)가 제안한 최소 부피 타원체(minimum volume ellipsoid)와 Hubert와 Debruyne (2010)이 제안한 최소 공분산 행렬식이다. MCD는 $n^{-1/2}$의 속도로 정규 분포로 수렴한다는 장점을 가지고 있으며, 특히 Fast-MCD가 개발된 이후로 평균과 공분산의 강력한 추정치로 MCD를 선호한다 (Hubert와 Debruyne, 2010). 본 연구에서는 이러한 장점을 가진 Fast-MCD를 활용하여 마할라노비스 거리를 계산한다. 먼저 Fast-MCD 알고리즘의 기반이 되는 MCD를 고려하고자 한다.

MCD를 설명하기 위해서 2.1절에서 설명한 자료 행렬 X_1, X_2를 합친 자료 행렬을 X라 할 때 식 (2.4)와 같이 X는 n_1과 n_2의 개체를 합친 n개의 개체와 p개의 변수를 갖게 된다.

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}. \tag{2.4}$$

식 (2.4)에서 공분산 행렬식을 최소로 만드는 h개 표본 관측치를 뽑아 부분집합 행렬 H의 평균벡터 $\bar{\mathbf{x}}_h = (1/h)\sum_{i=1}^{h} \mathbf{x}_i$와 공분산 행렬 $S_h = (1/h)\sum_{i=1}^{h}(\mathbf{x}_i - \bar{\mathbf{x}}_h)(\mathbf{x}_i - \bar{\mathbf{x}}_h)^T$를 구하면 이를 MCD 추정치라 한다. h는 $(n + p + 1)/2 \le h \le n$를 만족하는 정수이며 h가 $(n + p + 1)/2$일 경우 MCD 추정치가 가장 로버스트(robust)하다. 이 MCD 추정치는 유사등변성(affine equivariance) 특성과 관측치의 50% 가까이가 오염되어 있더라도 영향을 받지 않는다 (Davies, 1992). 그러나 MCD 추정치를 얻기 위해서는 각 관측치를 계속 추출하여 공분산 행렬을 계산해야 하므로, 관측치가 많은 자료의 경우 막대한 계산이 요구되어 효율성이 낮다는 단점을 가진다.

이에 Rousseuw와 Driessen (1999)은 계산적으로 효율적인 Fast-MCD 알고리즘을 구축했다. Algorithm 2은 Fast-MCD 알고리즘 설명이다.

Algorithm 2 : Fast-MCD Algorithm

[Step 1] Construct a subset matrix H_1, containing h sampled observations from the data matrix X, and compute its mean vector $\bar{\mathbf{x}}_1$ and covariance matrix S_1.

[Step 2] Compute the Mahalanobis distance $r(\mathbf{x}_i, \bar{\mathbf{x}}_1)$ for each observation in the entire data matrix X using the $\bar{\mathbf{x}}_1$ and S_1.

$$r(\mathbf{x}_i, \bar{\mathbf{x}}_1) = \|\mathbf{x}_i - \bar{\mathbf{x}}_1\|_{S_1^{-1}}, \quad i = 1, \ldots, n.$$

[Step 3] Construct a subset matrix H_2 using the h observations corresponding to the smallest distances in $r(\mathbf{x}_i, \bar{\mathbf{x}}_1)$, and compute its $\bar{\mathbf{x}}_2$ and S_2. (In this case, $|S_1| \geq |S_2|$ is guaranteed.)

[Step 4] Repeat [Step 2] and [Step 3] until $|S_l| = 0$ or $|S_l| = |S_{l+1}|$.

[Step 5] Using the $\bar{\mathbf{x}}_l$ and S_l of the subset matrix H_l obtained through iteration, compute the Mahalanobis distance $r(\mathbf{x}_i, \bar{\mathbf{x}}_l)$ for the entire data matrix X.

$$r(\mathbf{x}_i, \bar{\mathbf{x}}_l) = \|\mathbf{x}_i - \bar{\mathbf{x}}_l\|_{S_l^{-1}}, \quad i = 1, \ldots, n; l =, \ldots, n-1.$$

Table 2: Confusion matrix

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

Algorithm 2을 통해 구해진 평균벡터 $\bar{\mathbf{x}}_{mcd}$와 공분산 행렬 S_{mcd}를 사용하여 마할라노비스 거리는 식 (2.5)와 같이 적용된다.

$$r(\mathbf{x}_i, \bar{\mathbf{x}}_{mcd}) = \|\mathbf{x}_i - \bar{\mathbf{x}}_{mcd}\|_{S_{mcd}^{-1}} = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}_{mcd})^T S_{mcd}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{mcd})}. \tag{2.5}$$

3. 연구방법

Abdi와 Hashemi (2016)은 마할라노비스 거리를 기반으로 하는 SMOTE 방법인 MDO 방법을 제안했다. 그러나 자료에 이상치가 존재할 경우, 전통적인 마할라노비스 거리를 사용하면, 관측치 식별과정에서 감춤효과가 발생할 수 있어 결과가 정확하지 않을 수 있다. 따라서 본 연구에서는 평균벡터와 공분산 행렬을 Fast-MCD를 활용하고 이를 마할라노비스 거리에 적용한 SMOTE 방법을 제안한다. 이 절에서는 본 방법을 포함한 기존의 관측치 증가 방법들과의 성능 차이를 비교하기 위해 사용된 성능 지표와 분석에 활용된 실제 자료를 소개한다.

3.1. 평가지표

일반적으로 분류 성능을 평가할 때 정확도(accuracy)가 사용되지만, 불균형 자료에서는 정확도가 적절한 척도가 아니라는 문제가 있다. 이는 소수 군집이 정확도에 미치는 영향이 상대적으로 적기 때문이다 (Sun 외, 2009). 따라서 Table 2를 이용해 불균형 자료에서 선호되는 지표에 대해 설명한다.

Table 3: Data sets

Data set	Number of observation	Positive rate	Number of variable
Yeast 3	1484	10.98%	8
Pima	768	34.90%	8
Haberman	306	26.47%	3
Page-blocks	5472	10.21%	10

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}. \tag{3.1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}. \tag{3.2}$$

식 (3.1)의 재현율(recall)은 분류모형이 실제 positive인 관측치들 중에서 올바르게 positive로 예측한 관측치의 비율을 나타낸다. 반면, 식 (3.2)의 정밀도(precision)는 분류모형이 positive로 예측한 관측치 중에서 실제 positive인 관측치의 비율을 나타낸다. 이 두 지표는 주로 positive 값의 성능을 평가하는 데 중점을 두는 지표이다. 그러나 재현율과 정밀도는 서로 상충하는 관계에 있기 때문에 이러한 지표를 결합한 식 (3.3)의 점수를 사용한다.

$$F1 = \frac{2}{(1/\text{recall}) + (1/\text{precision})} = 2 \times \frac{\text{precison} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{3.3}$$

또한, 이진 분류 성능을 평가에 ROC (reciver operating characteristic) 곡선과 이를 기반으로 한 AUC (area under the curve) 점수가 가장 널리 사용된다. ROC 곡선은 재현율과 1-특이성(specificity)으로 그려진 곡선이다. 특이성은 분류모형이 실제 negative로 예측한 비율을 나타내며, TN/(FP+TN)으로 계산된다. 1에서 특이성을 빼면 FP/(FP+TN)이 된다. ROC 곡선의 x축은 FPR (false positive rate)이며 FPR이 변할 때 TPR (true positive rate)이 어떻게 변하는지를 보여주는 곡선이다. 이러한 ROC 곡선 아래 영역인 AUC 점수는 ROC 곡선의 성능을 요약한 값으로, 이 값이 1에 가까울수록 더 우수한 성능을 의미한다.

3.2. 실증분석

성능 비교를 위해 4가지 증가 방법을 적용한 자료를 다음과 같이 정의한다.

(1) 원본 자료(original)

(2) SMOTE를 적용한 자료(SMOTE)

(3) 마할라노비스 거리 기반 SMOTE 자료(MDO)

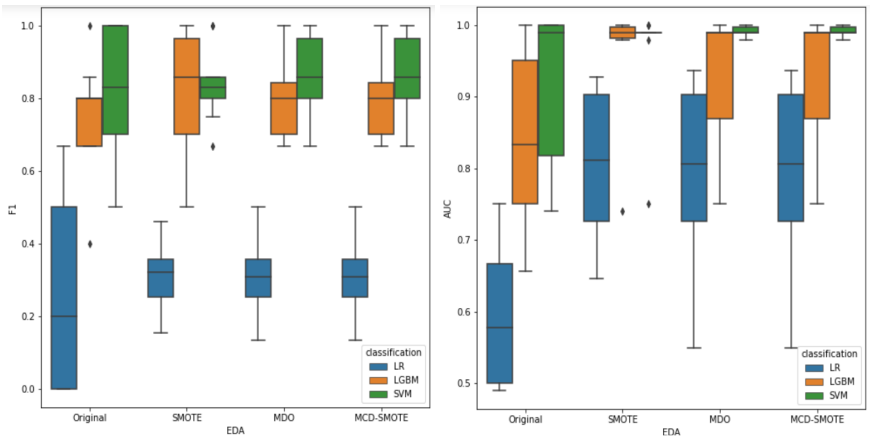
(4) Fast-MCD 추정량을 통해 마할라노비스 거리 기반 SMOTE 자료(MCD-SMOTE)

모든 자료는 knowledge extraction evolutionary learning(http://www.keel.es/)에서 제공하는 불균형 자료를 사용하였다. Table 3는 자료의 특성에 대해 정리한 표이다.

Table 3 자료의 모든 반응변수 y는 class이며 값으로 negative와 positive로 이루어진 이진 분류 자료이며 모두 결측치가 없는 자료이다. 반응변수는 0과 1로 인코딩(encoding)되었으며, 성능 평가를 위해 F1 점수와 AUC 점수를 평가지표로 사용하였다. 이 자료들을 이용해 로지스틱 회귀모형(LR), light gradient boosting model (LGBM), 서포트 벡터 머신(SVM)을 이용하여 성능을 평가하였다.

Table 4: Classification results of simulation

Classifier	Method	F1 Score	AUC Score
LR	Original	0.2567	0.5896
	SMOTE	**0.3093**	**0.8084**
	MDO	0.3017	0.7938
	MCD-SMOTE	0.3017	0.7938
LGBM	Original	0.7457	0.8385
	SMOTE	**0.8205**	**0.9656**
	MDO	0.7914	0.9333
	MCD-SMOTE	0.7914	0.9333
SVM	Original	0.8195	0.9114
	SMOTE	0.8388	0.9666
	MDO	**0.8588**	**0.9905**
	MCD-SMOTE	**0.8588**	**0.9905**



(a) F1 Scores cross validation (b) AUC Scores cross validation

Figure 1: *Performance comparison in simulation.*

4. 연구결과

Table 3의 자료 중 불균형 정도가 10% 미만인 자료가 없으므로, 4.1절에서는 모의 실험을 통해 결과를 확인하고자 한다. 4.2절에서는 Table 3의 4가지 실증분석 자료를 사용한다. 모의 실험과 실증 분석 모두 3.2절에서 정의한 (1)부터 (4)의 방법을 사용하여 3가지 분류 모형의 성능을 비교한다. 모형의 성능을 평가하고 일반화 능력을 확인하기 위해 10-fold 교차 검증을 수행하였다.

4.1. 모의 실험

모의 실험 자료는 500개의 관측치와 2개의 변수를 가진 자료를 준비했다. 이 중 5%만이 소수 군집에 속하도록 설정했다. 그 후, 25개의 소수 군집 중 무작위로 2개를 선택하여 이상치로 만들었다.

Table 4과 Figure 1은 모의 실험 자료에 대한 3가지 분류 모형과 4가지 추출 방법에 대한 F1 점수와 AUC 점수 값을 비교한 결과를 보여준다. Table 4에서는 SVM을 제외하고 모두 SMOTE에서 두 점수 모두 가장

Table 5: Classification results of Yeast3 data

Classifier	Method	F1 Score	AUC Score
LR	Original	0.2369	0.5698
	SMOTE	0.6700	0.8881
	MDO	0.6768	0.8942
	MCD-SMOTE	**0.6796**	**0.8949**
LGBM	Original	0.7187	0.8376
	SMOTE	0.7482	**0.8717**
	MDO	0.7404	0.8657
	MCD-SMOTE	**0.7500**	0.8695
SVM	Original	0.4668	0.6580
	SMOTE	0.6690	0.8942
	MDO	0.6686	**0.8944**
	MCD-SMOTE	**0.6746**	0.8936

Table 6: Classification results of Pima data

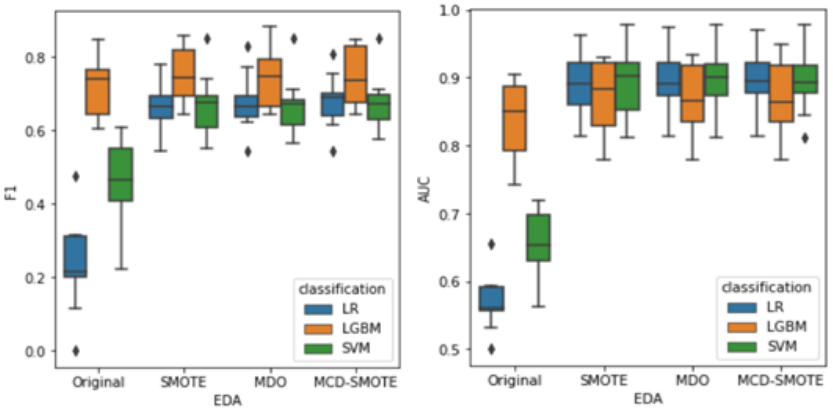
Classifier	Method	F1 Score	AUC Score
LR	Original	0.6133	0.7160
	SMOTE	0.6653	0.7431
	MDO	0.6698	0.7460
	MCD-SMOTE	**0.6740**	**0.7497**
LGBM	Original	0.6205	0.7098
	SMOTE	0.6418	0.7250
	MDO	0.6271	0.7144
	MCD-SMOTE	**0.6499**	**0.7300**
SVM	Original	0.6123	0.7139
	SMOTE	**0.6742**	**0.7504**
	MDO	0.6631	0.7407
	MCD-SMOTE	0.6692	0.7454

높았다. 그러나 Figure 1의 (a)와 (b)에서 SMOTE의 상자 그림은 이상치가 보인다. 이는 특정 fold에만 높은 점수나 낮은 점수를 보이며 일관된 결과를 보이지 않기 때문이다. MDO와 MCD-SMOTE를 통해 변수의 상관관계를 고려하여 이 문제를 해결하였다. MDO와 MCD-SMOTE의 점수가 차이가 없는 것은 모의 실험 자료의 이상치가 2개로 매우 적기 때문에 이상치의 영향이 전체 자료에 미치는 영향이 적기 때문이다.

4.2. 실증자료 결과

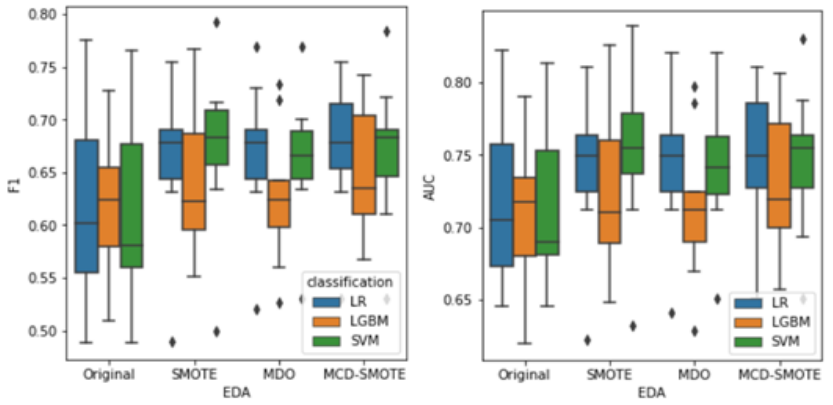
Table 5부터 Table 6은 자료별 교차 검증 점수들의 평균이다. 각 과대추출 방법에 따른 점수들을 시각적으로 확인하기 위해 Figure 2부터 Figure 3을을 제시한다.

Table 5과 Figure 2는 Yeast3 자료에 대해 3가지 분류모형과 4가지 추출 방법에 대한 F1 점수와 AUC 점수 값을 비교한 결과를 제시한 표와 그림이다. Figure 2의 (a)와 (b)는 F1 점수와 AUC 점수의 교차 검증 점수들을 상자 그림을 통해 시각적으로 보여준다. 표를 통해 MCD-SMOTE가 3가지 분류모형 모두에서 가장 높은 F1 점수를 기록했음을 볼 수 있다. 이는 재현율과 정밀도 모두에서 성능이 향상되었음을 시사한다. LGBM은 F1 점수에서 0.7500으로 가장 우수한 성과를 보였으며, AUC 점수 값은 분류모형에 상관없이 유사한 값을 나타내었지만 LR에서 0.8949로 가장 뛰어난 성능을 기록했다. 뿐만 아니라, Figure 2를를 통해 교차 검증 별 F1



(a) F1 Scores cross validation (b) AUC Scores cross validation

Figure 2: *Performance comparison in Yeast3 data.*



(a) F1 Scores cross validation (b) AUC Scores cross validation

Figure 3: *Performance comparison in Pima data.*

점수와 AUC 점수의 변동성이 감소했음을 확인할 수 있다.

Table 6과 Figure 3은 Pima 자료에 대한 결과이다. Table 6에서 볼 수 있듯이 LR과 LGBM은 모두 MCD-SMOTE가 가장 우수한 성능을 보였다. 특히, LGBM의 경우 MDO를 사용했을 때 F1 점수와 AUC 점수가 0.6271 및 0.7144로 감소하며 SMOTE보다 낮은 성능을 보였다. 그러나 MCD-SMOTE를 사용하면 이러한 점수가 약 0.02 정도 향상되며 SMOTE보다 높은 성능을 관찰할 수 있었다. 이는 이상치를 제외하고 마할라노비스 거리를 계산한 SMOTE가 더 좋은 성능을 제공한다는 것을 시사한다. Figure 3의 (a)와 (b)를 살펴보면 Original과 SMOTE를 사용한 경우 SVM을 통해 분류할 때 성능 값이 매우 흩어져 있으며, F1 점수와 AUC 점수가 중위수 주변에 흩어져 있음을 보여준다. 이러한 경우 결과의 신뢰성을 보장하기 어렵다. 반면 MDO와 MCD-SMOTE를 사용한 경우, SVM의 성능은 일반적으로 감소하지만, F1 점수와 AUC 점수는 중앙값 주변에 집중되어 있음을 관찰할 수 있다. 특히, MCD-SMOTE는 MDO보다 높은 성능을 나타내었다. 이러한 결과는 MCD-SMOTE가 예측의 일관성을 강화하고 분류모형 선택 시에도 효율적인 모형을 선택할 수 있음을

강조한다.

5. 결론

본 연구에서는 불균형 자료의 분류 성능을 향상시키기 위해 Fast-MCD를 이용하여 평균 벡터와 공분산 행렬을 추정하고, 이를 기반으로 마할라노비스 거리를 계산하여 SMOTE에 적용하였다. Fast-MCD를 통해 이상치가 전체 관측치의 추정에 큰 영향을 미치지 않도록 공분산 행렬이 감소하므로, 마할라노비스 거리가 이상치에 민감하지 않게 되었다. 이를 통해 MDO와 SMOTE보다 더 적합한 관측치를 선택하여 최근접 이웃을 결정할 수 있었다. 불균형 자료의 문제는 다양한 분야에서 공통적으로 발생하며, 대중적으로 사용되는 분류모형은 불균형 자료에 부적절하여 많은 관심을 받고 있다. 특히, 적은 수의 이상치를 가진 자료에서 MCD-SMOTE가 적절하게 활용될 것으로 기대된다.

실증자료를 통해 기존에 제안된 방법과 비교했을 때, MCD-SMOTE가 모든 경우에서 더 나은 성능을 보여주지는 않았다. 그러나 확인된 바와 같이 SMOTE는 특정 관측치의 부분집합에 과도하게 적합되어 불안정한 성능을 보이며 중위수를 중심으로 많이 퍼져 있는 반면, MCD-SMOTE를 통해 교차 검증 점수의 변동이 줄어든 것을 확인할 수 있었다. 이는 MCD-SMOTE가 일관된 예측을 수행하며 모형 선택 시에도 효율적인 모형을 선택할 수 있게 된다는 것을 시사한다.

본 연구의 한계점은 다음과 같다. 먼저, Algorithm 2의 Fast-MCD 알고리즘은 자료의 오염도가 높더라도 공분산 행렬을 추정하는 데 사용할 수 있지만, 표본으로 추출된 관측치의 개수 h가 변수의 수 p보다 훨씬 적을 경우 MCD 추정치를 계산하는 것이 어려워지며, h가 p보다 작을 경우 l단계의 부분집합 행렬 H_l의 공분산 행렬이 단위행렬이 되는 문제가 발생될 수 있다. 또한, 본 연구에서의 자료들은 모두 이진 분류 자료로 다중 분류 자료의 결과에 대한 적용 가능성은 알 수 없다. 따라서 향후 연구에서는 다중 분류 자료를 통해 MCD-SMOTE 방법의 성능을 확인할 필요가 있다. 뿐만 아니라, SMOTE 알고리즘을 기반으로 하는 과대추출법들의 단점을 극복할 수 있는 다른 방법들과 다양한 로버스트 추정법들과의 조합으로 추가적인 연구를 기대할 수 있다.

References

Abdi L and Hashemi S (2016). To combat multi-class imbalanced problems by means of over-sampling techniques, *IEEE Transactions on Knowledge and Data Engineering*, **28**, 238–250.

Aelst SV and Rousseeuw P (2009). Minimum volume ellipsoid, *WIREs Computational Statistics*, **1**, 71–82.

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

Davies PP (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator, *The Annals of Statistics*, **20**, 1828–1843.

Hubert M and Debruyne M (2010). Minimum covariance determinant, *WIREs Computational Statistics*, **2**, 36–43.

Menardi G and Torelli N (2010). Training and assessing classification rules with imbalanced data, *Data Mining and Knowledge Discovery* **28**, 92–122.

Rousseeuw PJ and Driessen KV (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212–223.

Rousseuw PJ and van Zomeren BC (1990). Unmasking multivariate outliers and leverage points, *Journal of the Statistical Association*, **85**, 633–639.

Sun Y, Wong AKC, and Kamel MS (2009). Classification of imbalanced data : A review, *International Jorunal of Pattern Recognition*, **23**, 687–719.

Wing WY Ng, Hu J, Yeung DS, Yin S, and Roli F (2015). Diversified sen-sitivity-based undersampling for imbalance classification problems, *IEEE Transactions on Cybernetics*, **45**, 2402–2412.

Received January 27, 2024; Revised March 08, 2024; Accepted March 10, 2024

불균형 자료에서 MCD를 활용한 마할라노비스 거리에 의한 SMOTE

정지은[a], 최용석[1,a]

[a]부산대학교 통계학과

요 약

불균형 자료 문제에 대한 해결책으로 SMOTE (synthetic minority over-sampling technique)가 가장 많이 사용되고 있다. SMOTE는 유클리드 거리를 기반으로 가장 가까운 이웃을 선택한다. 그러나 유클리드 거리의 단점 중 하나는 변수들 간의 상관관계를 고려하지 않는다는 것이다. 이에 대한 대안으로 변수 간의 공분산을 고려하는 마할라노비스 거리가 제안되었다. 그러나 이상치가 존재하는 경우, 대개 마할라노비스 거리를 계산하는 데 영향을 미친다. 이 문제를 해결하기 위해 최소 공분산 행렬 MCD (minimum covariance determinant)를 사용하여 공분산 행렬을 추정하여 마할라노비스 거리를 사용한다. 이후 MCD를 활용한 마할라노비스 거리를 SMOTE에 적용하여 새로운 관측치를 생성한다. 대부분의 경우 이 방법이 불균형 자료를 분류하는 데 높은 성능 지표를 제공함을 보여주었다.

주요용어: 불균형 자료, 마할라노비스 거리, MCD, SMOTE

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