creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

김 경 희 교수 지도
석사학위 청구논문

레이블 노이즈가 존재하는 불균형 자료의 분류분석 연구

2019

성신여자대학교 대학원
통계학과
권 소 영

레이블 노이즈가 존재하는 불균형 자료의 분류분석 연구

김 경 희 교수 지도

이 논문을 석사학위논문으로 제출함

2018년 11월

성신여자대학교 대학원

통계학과

권 소 영

인 준 서

권소영의 석사학위 논문으로 인준함.

2018년 11월

심사위원장 ..... (서명 또는 인)

심사위원 ..... (서명 또는 인)

심사위원 ..... (서명 또는 인)

성신여자대학교 대학원

논 문 개 요

관측된 범주형 종속변수에 오류가 포함된 것을 의미하는 레이블 노이즈(label noise)와 두 집단의 자료 수가 불균형한 자료를 의미하는 불균형자료(imbalanced data)는 실데이터에서 접하기 쉬운 문제이며 분류 성능을 낮출 수 있다. 따라서 본 논문에서는 이를 개선하기 위해 레이블 노이즈가 존재하는 불균형자료에서의 분류분석(classification analysis)에 대해 비교연구를 진행하였다. 특히, 레이블 노이즈의 발생형태, 샘플링방법, 분류방법에 따른 분류성능을 Accuracy, G-mean, AUC를 이용하여 살펴보았다. 이를 통해 데이터의 형태와 레이블 노이즈의 발생형태에 따라 상황별로 적합한 샘플링방법과 분류방법을 제안하고자 한다.

목 차

그 림 목 차

표 목 차

I. 서 론

분류문제에서 레이블은 중요하지만, 정확하고 신뢰할만한 레이블을 얻는 것은 어렵고 많은 비용과 시간이 필요하다. 레이블이 오염될 수 있는 요인으로는 불충분한 정보, 전문가의 실수, 인코딩 오류 등이 있으며, 학습데이터의 오염된 레이블을 처리하지 않으면 후에 예측 정확도가 떨어질 수 있다. 본 논문에서는 이처럼 관측된 범주형 종속변수에 오류가 포함된 것을 레이블 노이즈라고 부르며, 두 그룹을 가지는 자료에서 그룹 1을 그룹 2라고 관측하거나 그룹 2를 그룹 1로 잘못 관측하는 경우를 다룬다.

자료의 불균형 또한 분류 성능에 영향을 줄 수 있는 중요한 요인이다. 불균형자료는 파산감지, 불량품 감지 등 일상생활에서 접하기 쉽고 이러한 불균형자료를 분류 분석할 경우 소수 그룹이 다수 그룹에 포함되는 형태로 잘못 분류될 수 있다. 따라서 이를 해결하기 위해 불균형자료에 관한 연구들이 이루어지고 있다. Kim et al.(2014)는 필기체 인식의 경우도 불균형 데이터 문제로 보고, 불균형 문제를 고려하여 필기체 인식기의 성능을 향상시킬 수 있는 과표본화 기반의 앙상블 학습 기법을 제안하였다. Park과 Bang(2015)은 불균형자료의 분류정확도 개선을 위해 다양한 샘플링 기법을 이용한 로지스틱 회귀분석 방법론을 연구하였다. Kim et al.(2015)은 샘플링 기법을 이용하여 불균형자료에 대한 분류방법들의 성능을 비교 분석하였다.

자료의 불균형을 해결하기 위한 방법으로는 크게 Up-샘플링방법과 Down-샘플링방법이 있다. Up-샘플링방법은 Down-샘플링방법과 비교했을 때 모든 데이터를 사용할 수 있는 장점이 있지만, 계산이 오래 걸리는 단점이 있다. Down-샘플링방법은 데이터 크기가 매우 클 때 계산 시간이 단축되므로 효과적이지만 데이터의 일부가 손실되는 단점이 있다.

최근 레이블 노이즈와 불균형자료에 관한 연구들은 많이 진행되고 있지만, 불균형자료에 존재하는 레이블 노이즈의 형태에 따라 샘플링방법이 분류 성능에 미치는 영향 비교한 연구는 찾기 힘들다. 따라서 본 연구에서는 레이블 노이즈가 존재하는 불균형자료에서 레이블 노이즈의 형태에 따라 샘플링방법과 분류방법이 분류 성능에 어떠한 영향을 미치는지 모의실험을 통해 살펴보고자 한다.

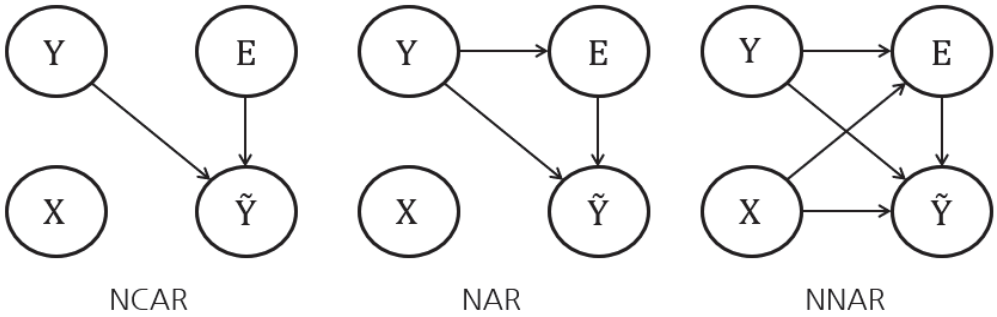
본 연구에서 사용한 분류방법으로는 분류문제에서 많이 사용되는 Linear Discriminant Analysis classifiers(LDA; Fisher, 1936), Quadratic Discriminant Analysis classifiers(QDA), k-nearest neighbour(KNN; Altman, 1992), Support Vector Machine (SVM; Cortes와 Vapnik, 1995)를 고려하였다. 분류성능을 판단하는 기준으로는 학습데이터와 평가데이터를 나눈 후 Accuracy, G-mean, AUC를 계산하였다.

본 논문에서는 분류 성능을 비교하기에 앞서 레이블 노이즈와 샘플링방법에 대해 소개한 후, 분류 알고리즘과 모형평가방법에 대해 논의한다. 논의 후에는 모의실험을 통해 레이블 노이즈의 형태에 따라 샘플링방법과 분류방법이 분류 성능에 미치는 영향을 면밀히 살펴보고자 한다. 이를 통해 데이터의 형태와 레이블 노이즈의 발생형태에 따라 상황별로 적합한 샘플링방법과 분류방법을 제안하고자 한다.

II. 본 론

II.1. 레이블 노이즈

레이블 노이즈는 관측된 범주형 종속변수에 오류가 포함된 것을 의미하며, 본 연구에서는 두 그룹을 가지는 자료에서 그룹 1을 그룹 2로 관측하거나 그룹 2를 그룹 1로 관측하는 경우를 다룬다. Frénay와 Verleysen(2014)에 따르면 레이블 노이즈는 완전 임의 노이즈(noisy completely at random: NCAR), 임의 노이즈(noisy at random: NAR), 비 임의 노이즈(noisy not at random: NNAR) 세 가지로 분류된다. NCAR은 레이블 노이즈가 종속변수와 설명변수의 영향을 받지 않는다. 따라서 각 그룹의 레이블 노이즈 발생 개체의 수가 같다. NAR은 레이블 노이즈가 종속변수의 영향을 받는 경우이며, 특정 그룹의 개체가 레이블 노이즈 발생이 더 쉬울 경우 레이블 노이즈의 비대칭을 허용한다. 마지막으로 NNAR은 레이블 노이즈의 발생이 종속변수와 설명변수의 영향을 모두 받는다. 예를 들면, 분류의 경계나 낮은 밀도를 가지는 곳에서 레이블 노이즈가 발생할 경우 NNAR로 모형화될 수 있다. [그림 1]은 NCAR, NAR, NNAR을 도식화하여 나타낸 것이다. X는 설명변수의 벡터이고 Y는 레이블의 참값, $\tilde{Y}$은 관측된 레이블, E는 Y와 $\tilde{Y}$가 일치하는지를 나타내는 이진변수이다. 아래에 나오는 화살표는 종속관계를 의미한다.



[그림 1] NCAR, NAR, NNAR의 도식화(Frénay와 Verleysen, 2014)

II.2. 불균형자료에서의 샘플링방법

불균형자료는 두 집단의 자료의 수가 불균형한 자료를 의미하며, 불균형자료를 분류 분석할 경우 소수 그룹이 다수 그룹에 포함되는 형태로 잘못 분류될 수 있다. 따라서 불균형자료에서의 분류 성능을 개선하고자 다양한 샘플링방법이 제안되어왔다. 본 연구에서는 Up-샘플링방법과 Down-샘플링방법을 사용하고자 한다.

(1) Up-샘플링방법

Up-샘플링방법은 소수 그룹의 데이터를 다수 그룹의 크기에 맞추어 추출함으로써 불균형자료의 분류문제를 해결하는 것이다. 예를 들면, 900:100의 비율을 가지는 불균형자료의 경우 100개의 데이터를 9번 반복해서 추출한다. 이때, 추출방법은 붓스트랩과 같이 복원추출을 허용하는 방법과 노이즈를 더해주는 방법이 있다(Kim et al., 2015). 본 연구에서는 복원추출을 허용

하여 추출하는 붓스트랩 방법을 사용하였다.

(2) Down-샘플링 방법

Down-샘플링 방법은 다수 그룹의 데이터를 소수 그룹의 크기에 맞추어 추출함으로써 불균형자료의 분류문제를 해결하는 것이다. 예를 들면, 900:100의 비율을 가지는 불균형자료의 경우 900개의 데이터 중 100개를 무작위로 추출한다.

II.3. 분류 알고리즘

본 연구에서는 불균형자료의 분류분석을 시행할 때 가장 광범위하게 사용되는 분류기 중 네 개인 LDA(Fisher, 1936), QDA, KNN(Altman, 1992), SVM(Cortes와 Vapnik, 1995)를 고려하였다.

(1) LDA(선형판별분석)

먼저, 두 그룹을 가지고 설명변수가 한 개일 경우로 가정했을 때 LDA(선형판별분석) 방법은 π_j, μ_j, σ^2에 대한 추정값을 이용하여 식(2.1)가 최대가 되는 그룹에 관측치 $X=x$를 할당하는 방법이다.

$$\Pr(Y=j|X=x)=\frac{\pi_j\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_j)^2\right)}{\sum_{l=1}^{2}\pi_l\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)} \tag{2.1}$$

여기서, Y는 범주형 종속변수를 의미하고 2개의 서로 다르고 순서가 없는 값을 가진다. π_j는 관측치가 그룹 j에 속하는 사전확률을 의미하며, μ_j는 그

룹 j에 대한 평균이다. σ^2는 두 그룹에 대한 공통의 분산이다. μ_j, σ^2의 추정에는 식(2.2), 식(2.3)의 추정치들이 사용된다. π_j는 그 값을 알고 있을 때는 그 값을 직접 사용하고 알려지지 않은 경우는 식(2.4)를 통해 추정한다.

$$\hat{\mu_j} = \frac{1}{n_j} \sum_{i:\, y_i = j} x_i \tag{2.2}$$

$$\hat{\sigma^2} = \frac{1}{n-2} \sum_{j=1}^{2} \sum_{i:\, y_i = j} (x_i - \hat{\mu_j})^2 \tag{2.3}$$

$$\hat{\pi}_j = \frac{n_j}{n} \tag{2.4}$$

여기서, n은 훈련데이터 관측치 수이고 n_j는 그룹 j의 훈련데이터 관측치 수이다.

만약 설명변수의 개수가 2개 이상이라면 식(2.1)이 아닌 식(2.5)를 이용하여 관측치 $X = x$를 할당한다.

$$\Pr(Y = j | X = x) = \frac{\pi_j \frac{1}{(2\pi)^{p/2} |\sum|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_j)^T \sum^{-1} (x-\mu_j)\right)}{\sum_{l=1}^{2} \pi_l \frac{1}{(2\pi)^{p/2} |\sum|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_l)^T \sum^{-1} (x-\mu_l)\right)} \tag{2.5}$$

LDA는 오류율이 가장 낮은 베이즈 분류기에 근접하고자 한다. 따라서 오류가 어느 그룹에서 발생하는지와 관계없이 오분류되는 관측치의 총수가 가장 낮을 것이다.

(2) QDA(이차선형판별분석)

QDA는 LDA와 마찬가지로 각 그룹의 관측치들이 가우스분포를 따른다고 가정한 후, 모수의 추정치를 베이즈 정리에 대입하여 수행한다. 하지만 QDA는 LDA와 다르게 각 그룹의 공분산 행렬이 다르다고 가정하며 판별식이 이차함수처럼 나타난다.

(3) KNN(k-최근접이웃)

KNN은 가장 가까운 K개 이웃의 정보를 바탕으로 데이터를 예측하는 방법론이다. KNN은 먼저 훈련데이터에서 x_0에 가장 가까운 K개 점을 식별한 후, 식(2.6)에 따라 그룹 j에 대한 조건부확률을 추정하고 검정 관측지 x_0을 확률이 가장 높은 그룹에 할당한다.

$$\Pr(Y=j|X=x_0)=\frac{1}{K}\sum_{i\in N_0} I(y_i=j) \tag{2.6}$$

여기서 N_0은 x_0에 가장 가까운 K개 점을 의미한다.

(4) SVM(서포트 벡터 머신)

SVM은 최대 마진 분류기(maximal margin classifier)을 일반화한 것이다. 최대 마진 분류기는 훈련 관측치들에서 분리 초평면까지의 최소거리가 가장 먼 초평면을 기반으로 분류하는 방법으로 단순하고 직관적이지만, 그룹들이 선형 경계에 의해 구별될 수 있어야 하는 요구조건이 있다. 따라서 이를 확장한 것이 서포트 벡터 분류기(support vector classifier)이며 이를 더 확장하여 비선형의 경계를 수용하도록 한 것이 SVM이다. 즉, SVM은 커널(kernels)을 이용하여 서포트 벡터 분류기의 변수공간을 확장한 결과이다(James et al., 2016). SVM은 식(2.7)의 형태로 나타낼 수 있다.

$$f(x)=\beta_0+\sum_{i\in S}\alpha_i K(x,x_i) \tag{2.7}$$

여기서 α_i는 훈련 관측지가 서포트 벡터가 아니면 그 관측치의 α_i는 0이다. 그러므로 서포트 포인트들의 인덱스 모임이 S라고 했을 때, 식(2.7)의 형태를 가지게 된다. $K(x,x_i)$는 두 관측치들의 유사성을 수량화하는 커널이라는 함수이다.

널리 사용되는 커널 중 하나는 선형 커널과 방사커널(radial kernal)로 p개의 설명변수를 가질 때, 각각 식(2.8), 식(2.9) 형태를 가진다.

$$K(x_i, x_{i'}) = \sum_{l=1}^{p} x_{il} x_{i'l} \tag{2.8}$$

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{l=1}^{p} (x_{il} - x_{i'l})^2\right) \tag{2.9}$$

여기서 γ는 양의 상수이다. 방사커널의 경우 주어진 관측치가 훈련 관측치로부터 유클리드 거리(Euclidean distance)로 멀리 떨어져 있으면, $K(x_i, x_{i'})$가 아주 작은 값이 될 것이다. 즉, 식(2.7)의 식에 주어진 관측치가 사실상 아무 역할을 하지 않는다는 것을 의미한다. 따라서 이는 주변 관측치들만이 검정 관측치들의 그룹에 영향을 준다는 점에서 방사커널은 국소적인 방식으로 시행된다.

II.4. 모형 평가방법

분류 모형의 성능을 평가하기 위해 여러 가지 추정치가 사용된다. 추정치를 구하기 위해서는 오분류표가 사용되는데 두 그룹을 가지는 자료의 분류를 위해 사용되는 오분류표는 [표 1]과 같다.

[표 1] 오분류표

	Predicted 0	Predicted 1
True 0	TN (True Negative)	FP (False Positive)
True 1	FN (False Negative)	TP (True Positive)

(1) Accuracy

분류 모형의 성능을 평가하기 위한 가장 대표적인 방법으로는 Accuracy

가 있다. Accuracy는 식(2.10)에 의해 구해진다.

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP}. \quad (2.10)$$

하지만 Accuracy를 불균형자료의 성능을 평가하기 위해 활용하면, 다수 그룹으로만 예측하더라도 높은 정확도를 보이기 때문에 불균형자료의 성능을 평가하기 위해서는 적합하지 않다. 따라서 소수 그룹의 정확도 또한 높일 수 있는 다른 측도를 이용하여 모형을 평가해야 한다.

(2) G-mean

불균형자료의 경우 소수 그룹에 속한 관측치를 다수 그룹으로 분류하는 경향을 보이므로 특이도 값은 크지만, 민감도 값은 매우 작게 된다. 따라서 불균형자료에서 모형의 성능을 평가하기 위해 민감도와 특이도의 기하평균인 식(2.13)의 G-mean이 이용된다(Kim et al., 2015).

$$\text{민감도}(Sensitivity) = \frac{TP}{FN + TP} \quad (2.11)$$

$$\text{특이도}(Specificity) = \frac{TN}{TN + FP} \quad (2.12)$$

$$G-mean = \sqrt{sensitivity \times specificity} \quad (2.13)$$

(3) AUC(Area Under the Curve; ROC Area)

AUC는 ROC 곡선 아래의 면적으로 분류기의 성능 평가를 위해 자주 사용되며, 그 값이 1에 가까울수록 예측력이 우수하다고 할 수 있다. ROC 곡선은 x축은 $1-Specificity$이고 y축은 $Sensitivity$로, 절단 값을 변화시키며 구한 점들을 연결한 것이다.

III. 모의실험

본 절에서는 레이블 노이즈가 존재하는 불균형자료에서 샘플링방법과 레이블 노이즈의 발생형태가 분류 성능에 미치는 영향을 모의실험을 통해 확인하고자 한다.

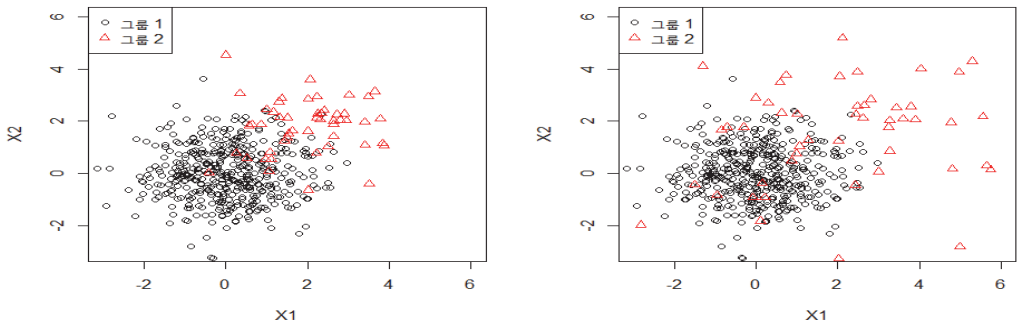
III.1. 모의실험설계

본 연구에서는 분류 기법들의 성능을 비교하는 데 있어서 시각화의 편의성을 위해 이변량 정규분포를 가정하여 자료를 그룹 1에서 450개, 그룹 2에서 50개 생성하였다. 본 연구에서 고려한 시나리오는 아래와 같으며 시나리오 1은 두 그룹의 공분산 행렬이 같고 시나리오 2는 두 그룹의 공분산 행렬이 다르다.

$$\text{시나리오 } 1 : \mu_1 = \begin{pmatrix}0\\0\end{pmatrix}, \mu_2 = \begin{pmatrix}2\\2\end{pmatrix}, \Sigma = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$$

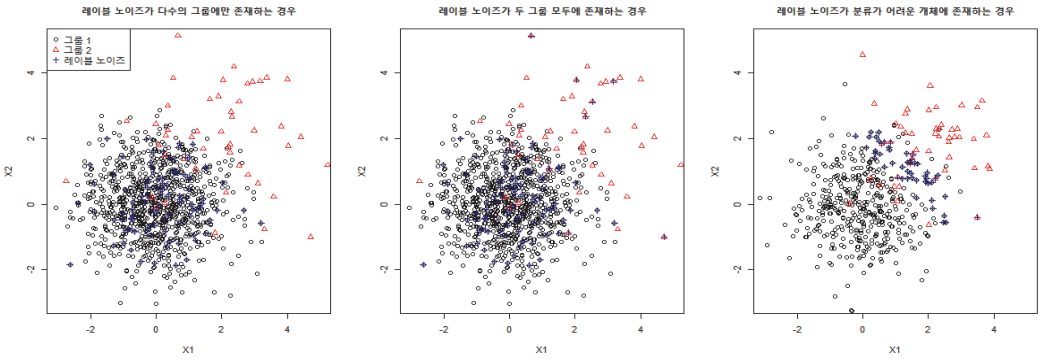
$$\text{시나리오 } 2 : \mu_1 = \begin{pmatrix}0\\0\end{pmatrix}, \mu_2 = \begin{pmatrix}2\\2\end{pmatrix}, \Sigma_1 = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}, \Sigma_2 = \begin{pmatrix}4 & 0\\0 & 4\end{pmatrix}$$

[그림 2]는 위의 시나리오 하에서 생성된 자료의 예이다.



[그림 2] 시나리오에 대한 산점도

본 논문에서는 레이블 노이즈의 발생형태를 NAR, NNAR의 형태에서 착안하여 레이블 노이즈가 다수 그룹에만 존재하는 경우, 두 그룹 모두에 존재하는 경우, 분류가 어려운 개체에 존재하는 경우 세 가지로 고려하였다. 레이블 노이즈가 두 그룹 모두에 존재하는 경우 불균형 비율과 레이블 노이즈의 비율을 동일하게 설정하였다. 그리고 분류가 어려운 개체에 레이블 노이즈가 존재하는 경우는 SVM을 활용하여 특정 그룹으로 예측될 확률이 낮은 개체를 분류가 어려운 개체로 선정하였다. [그림 3]은 본 논문에서 고려한 시나리오 1하에서 레이블 노이즈를 발생시킨 예이다. 모형을 평가할 때는 레이블 노이즈를 학습데이터에 발생시킨 후 학습데이터로부터 얻은 모형을 평가데이터에 적용하여 Accuracy, G-mean, AUC를 산출하였다. 모의실험은 이 과정을 100번 반복하여 실행하여 평균 Accuracy, G-mean, AUC를 산출하였으며 이를 바탕으로 모형의 성능을 비교하였다. 분류방법은 앞 절에서 소개한 LDA, QDA, KNN, SVM을 사용하였다. KNN의 모수 k는 $k=[n^{2/3}/2]$로 지정하였으며(Cannings et al, 2018), SVM의 모수는 10-fold 교차검증(Cross-Validation)방법을 통해 최적의 γ와 조율 파라미터 C를 구하여 지정하였다.



[그림 3] 시나리오 1하에서 발생시킨 레이블 노이즈의 예

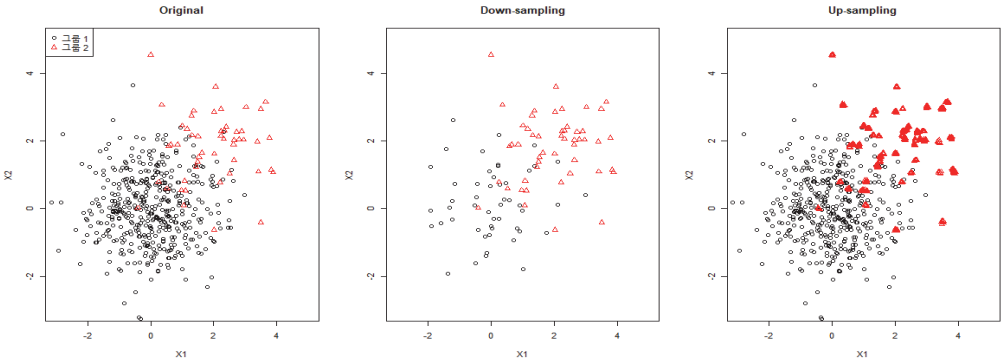
(1) 레이블 노이즈가 존재하지 않는 경우

레이블 노이즈가 존재하지 않는 불균형자료에서 분류방법에 따른 분류 성능을 비교하기 위한 절차는 다음과 같다.

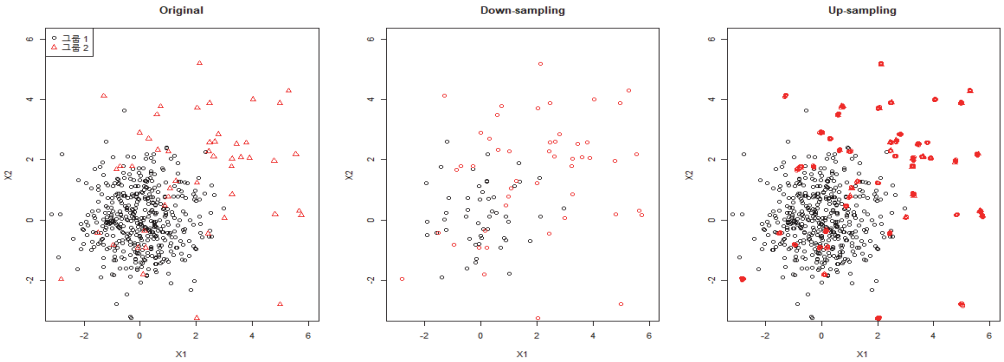
[표 2] 레이블 노이즈가 존재하지 않는 불균형자료에서 분류방법에 따른 분류 성능을 비교하기 위한 절차

[step 1] 시나리오에 따라 훈련데이터와 평가데이터 각각 500개씩 인공적으로 생성한다.
[step 2] 불균형자료이므로 Up-샘플링, Down-샘플링을 통해 두 그룹의 비율을 5:5로 맞춘다.
[step 3] 훈련데이터에 LDA, QDA, KNN, SVM을 시행한다.
[step 4] [step 3]으로부터 얻은 모형을 바탕으로 평가데이터를 예측한다. 예측된 그룹이 실제 그룹과 같은지 비교하여 Accuracy, G-mean, AUC를 구한다.
[step 5] [step 1]~[step 4]의 과정을 100번 반복한다.

[그림 4]와 [그림 5]는 [step 2]까지의 과정을 통해 얻은 자료의 예이다.



[그림 4] 시나리오 1에서 레이블 노이즈가 없는 원자료, Down-샘플링자료, Up-샘플링자료의 산점도

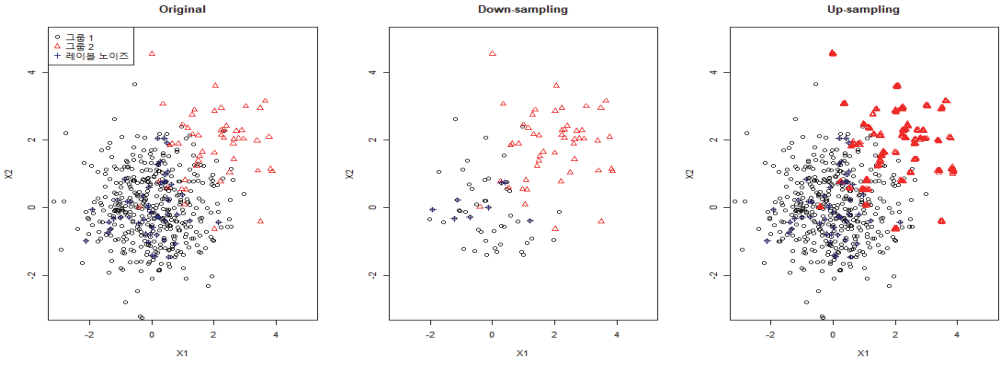


[그림 5] 시나리오 2에서 레이블 노이즈가 없는 원자료, Down-샘플링자료, Up-샘플링자료의 산점도

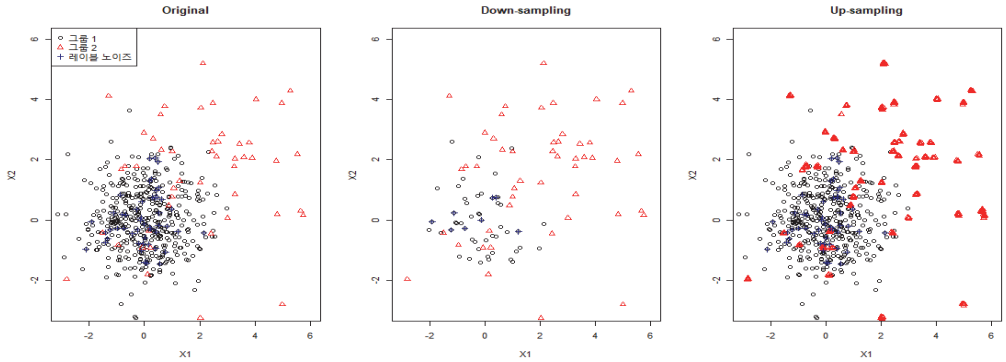
(2) 레이블 노이즈가 다수 그룹에만 존재하는 경우

레이블 노이즈가 다수 그룹에만 존재하는 경우 분류방법에 따른 분류 성능을 비교하기 위한 절차는 [표 2]의 절차에서 [step 1]을 수행한 후 다수 그룹에 레이블 노이즈를 10% 랜덤하게 발생시킨다. 예를 들면, 그룹 1이 다수 그룹일 경우 그룹 1에서 랜덤하게 50개를 뽑아 그룹 2로 변경한다.

[그림 6]과 [그림 7]은 [step 3]까지의 과정을 통해 얻은 자료의 예이다.



[그림 6] 시나리오 1에서 레이블 노이즈가 다수 그룹에만 존재하는 원자료, Down-샘플링자료, Up-샘플링자료의 산점도

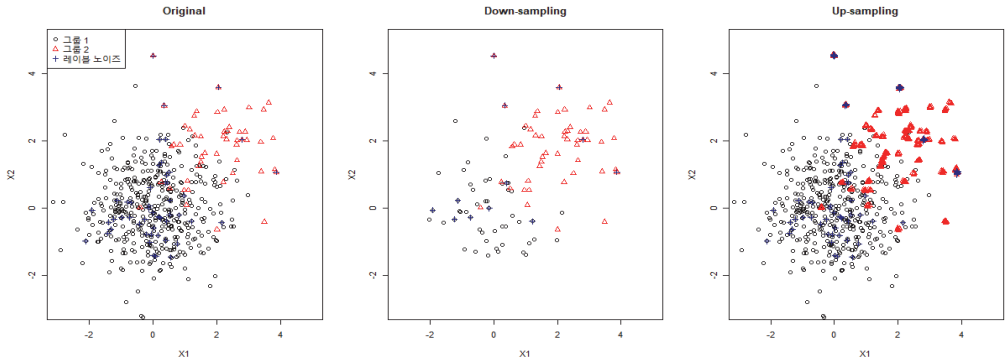


[그림 7] 시나리오 2에서 레이블 노이즈가 다수 그룹에만 존재하는 원자료, Down-샘플링자료, Up-샘플링자료의 산점도

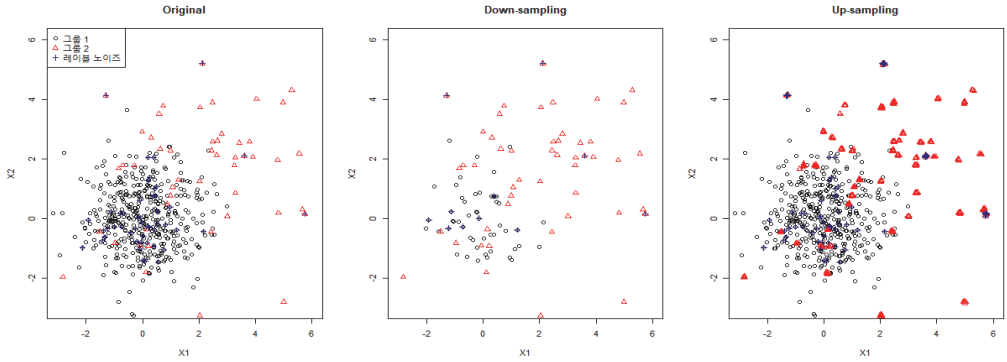
(3) 레이블 노이즈가 두 그룹 모두 존재하는 경우

레이블 노이즈가 두 그룹 모두 존재하는 경우 분류방법에 따른 분류 성능을 비교하기 위한 절차는 [표 2]의 절차에서 [step 1]을 수행한 후 불균형 비율에 맞춰 레이블 노이즈를 랜덤하게 10% 발생시킨다. 예를 들면, 그룹 1이 다수 그룹일 경우 그룹 1에서 랜덤하게 45개를 뽑아 그룹 2로 변경하고 그룹 2에서 랜덤하게 5개를 뽑아 그룹 1로 변경한다.

[그림 8]과 [그림 9]는 [step 3]까지의 과정을 통해 얻은 자료의 예이다.



[그림 8] 시나리오 1에서 레이블 노이즈가 두 그룹 모두에 존재하는 원자료, Down-샘플링자료, Up-샘플링자료의 산점도



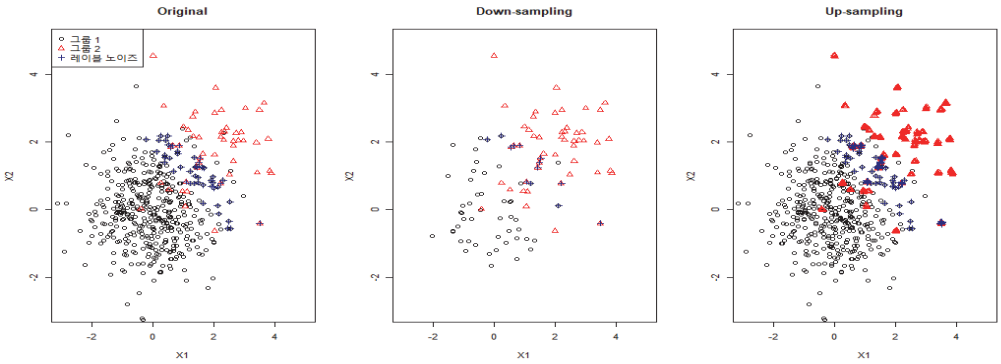
[그림 9] 시나리오 2에서 레이블 노이즈가 두 그룹 모두에 존재하는 원자료, Down-샘플링자료, Up-샘플링자료의 산점도

(4) 레이블 노이즈가 분류 어려운 개체에 존재하는 경우

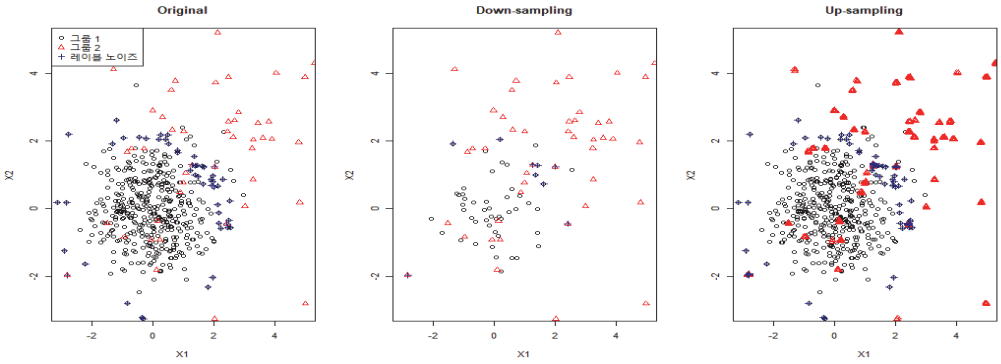
본 논문에서는 분류가 어려운 개체를 SVM을 활용하여 선정하고자 한다. SVM을 활용하여 분류가 어려운 개체에 레이블 노이즈가 존재하는 경우 분류방법에 따른 분류 성능을 비교하기 위한 절차는 다음과 같다. 먼저, [표 2]의 절차에서 [step 1]을 수행한 후, 시나리오 1은 선형 커널을 가진 SVM을 이용하여 훈련데이터의 예측확률을 구하고 시나리오 2는 RBF커널을 가

진 SVM을 이용하여 훈련데이터의 예측확률을 구한다. 그 다음, 앞에서 구해진 예측확률을 오름차순으로 정렬하고 406번째 개체부터 455번째 개체까지 개체의 레이블이 1이면 2로, 2이면 1로 변경하여 레이블 노이즈를 발생시킨다.

[그림 10]과 [그림 11]은 [step 3]까지의 과정을 통해 얻은 자료의 예이다.



[그림 10] 시나리오 1에서 레이블 노이즈가 분류 어려운 개체에 존재하는 원자료, Down-샘플링자료, Up-샘플링자료의 산점도



[그림 11] 시나리오 2에서 레이블 노이즈가 분류 어려운 개체에 존재하는 원자료, Down-샘플링자료, Up-샘플링자료의 산점도

III.2. 적용결과 및 해석

(1) 시나리오 1의 적용결과 및 해석

[표 3] 레이블 노이즈가 존재하지 않는 경우 분류 성능 비교

Model	Data	G-mean	AUC	Accuracy
LDA	Original	0.8510(0.0352)	0.8609(0.0304)	0.9608(0.0074)
	Down	0.9210(0.0258)	0.9214(0.0257)	0.9214(0.0257)
	Up	0.9196(0.0225)	0.9198(0.0224)	0.9198(0.0224)
QDA	Original	0.8517(0.0353)	0.8615(0.0305)	0.9609(0.0075)
	Down	0.9203(0.0255)	0.9207(0.0254)	0.9207(0.0254)
	Up	0.9194(0.0226)	0.9197(0.0224)	0.9197(0.0224)
KNN	Original	0.8007(0.0506)	0.8205(0.0403)	0.9576(0.0081)
	Down	0.9149(0.0323)	0.9156(0.0320)	0.9156(0.0320)
	Up	0.9148(0.0228)	0.9151(0.0227)	0.9151(0.0227)
SVM	Original	0.8285(0.0469)	0.8427(0.0385)	0.9596(0.0074)
	Down	0.9215(0.0257)	0.9219(0.0257)	0.9219(0.0257)
	Up	0.9208(0.0220)	0.9212(0.0218)	0.9212(0.0218)

[표 3]의 결과를 보면, 레이블 노이즈가 존재하지 않는 경우 Accuracy는 원자료를 사용한 결과가 가장 높았다. 하지만 G-mean, AUC는 Down-샘플링자료를 사용한 결과가 가장 높았다. 따라서 Down-샘플링자료의 결과를 비교해보면 분류방법 간의 큰 차이를 보이지 않았다.

[표 4] 레이블 노이즈가 다수 그룹에만 존재하는 경우 분류 성능 비교

Model	Data	G-mean	AUC	Accuracy
LDA	Original	0.8654(0.0361)	0.8730(0.0318)	0.9590(0.0090)
	Down	0.8581(0.0260)	0.8662(0.0234)	0.8086(0.0325)
	Up	0.8703(0.0245)	0.8628(0.0204)	0.8144(0.0204)
QDA	Original	0.8813(0.0326)	0.8864(0.0293)	0.9574(0.0089)
	Down	0.8939(0.0248)	0.8968(0.0240)	0.8631(0.0304)
	Up	0.9030(0.0208)	0.9007(0.0209)	0.8724(0.0223)
KNN	Original	0.8296(0.0415)	0.8432(0.0345)	0.9588(0.0073)
	Down	0.8809(0.0348)	0.8828(0.0340)	0.8617(0.0406)
	Up	0.8610(0.0342)	0.8580(0.0343)	0.8307(0.0358)
SVM	Original	0.6703(0.2703)	0.7599(0.1144)	0.9474(0.0206)
	Down	0.9004(0.0481)	0.9020(0.0461)	0.8879(0.0599)
	Up	0.9106(0.0233)	0.9089(0.0237)	0.8869(0.0276)

[표 4]의 결과를 보면, 레이블 노이즈가 다수 그룹에만 존재하는 경우 또한 Accuracy는 원자료를 사용한 결과가 가장 높았다. 하지만 G-mean은 KNN을 제외하고 Up-샘플링자료의 결과가 가장 높았다. AUC는 LDA는 원자료, KNN은 Down-샘플링자료, SVM과 QDA는 Up-샘플링자료가 가장 높은 AUC를 보였다. 특히 SVM에서 큰 차이를 보였다.

이 경우에는 대부분의 분류방법에서 샘플링자료를 사용한 결과가 더 높은 성능을 보이고, Up-샘플링방법을 사용한 SVM, QDA가 가장 좋은 성능을 보였다.

[표 5] 레이블 노이즈가 두 그룹 모두 존재하는 경우 분류 성능 비교

Model	Data	G-mean	AUC	Accuracy
LDA	Original	0.8178(0.0465)	0.8336(0.0375)	0.9570(0.0080)
	Down	0.8637(0.0271)	0.8713(0.0245)	0.8167(0.0340)
	Up	0.8648(0.0250)	0.8723(0.0232)	0.8184(0.0303)
QDA	Original	0.8447(0.0402)	0.8554(0.0339)	0.9583(0.0082)
	Down	0.8981(0.0245)	0.9003(0.0239)	0.8718(0.0296)
	Up	0.9044(0.0211)	0.9065(0.0208)	0.8794(0.0243)
KNN	Original	0.7919(0.0516)	0.8133(0.0407)	0.9555(0.0079)
	Down	0.8791(0.0333)	0.8773(0.0337)	0.8600(0.0403)
	Up	0.8550(0.0353)	0.8584(0.0341)	0.8278(0.0418)
SVM	Original	0.2370(0.3352)	0.5834(0.1210)	0.9158(0.0229)
	Down	0.9093(0.0229)	0.9108(0.0224)	0.8897(0.0290)
	Up	0.9095(0.0233)	0.9114(0.0227)	0.8871(0.0284)

[표 5]의 결과를 보면, 레이블 노이즈가 두 그룹 모두 존재하는 경우 또한 Accuracy는 원자료를 사용한 결과가 가장 높았다. 하지만 G-mean, AUC는 KNN을 제외하고 Up-샘플링자료의 결과가 가장 높았다. 특히 SVM에서 큰 차이를 보였는데 SVM은 한 그룹으로만 판단하는 경우가 많았기 때문으로 보인다.

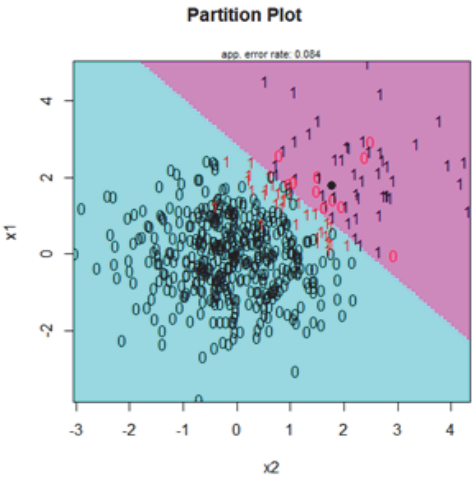
이 경우에는 샘플링자료를 사용한 결과가 높은 성능을 보이고 Up-샘플링 자료를 사용한 SVM이 가장 좋은 성능을 보였다.

[표 6] 분류가 어려운 개체에 레이블 노이즈가 존재하는 경우 분류 성능 비교

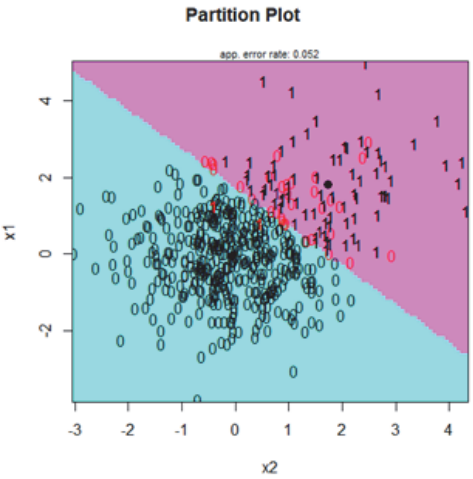
Model	Data	G-mean	AUC	Accuracy
LDA	Original	0.9077(0.0318)	0.9096(0.0298)	0.9476(0.0127)
	Down	0.7780(0.0215)	0.7937(0.0208)	0.7199(0.0232)
	Up	0.7833(0.0231)	0.7984(0.0216)	0.7231(0.0274)
QDA	Original	0.9077(0.0320)	0.9096(0.0299)	0.9477(0.0127)
	Down	0.7764(0.0212)	0.7928(0.0200)	0.7176(0.0237)
	Up	0.7814(0.0235)	0.7969(0.0220)	0.7207(0.0275)
KNN	Original	0.9122(0.0268)	0.9130(0.0259)	0.9336(0.0143)
	Down	0.7640(0.0218)	0.7845(0.0119)	0.7008(0.0243)
	Up	0.7712(0.0208)	0.7897(0.0197)	0.7067(0.0244)
SVM	Original	0.9085(0.0390)	0.9103(0.0346)	0.9332(0.0190)
	Down	0.7735(0.0224)	0.7906(0.0207)	0.7142(0.0260)
	Up	0.7443(0.0452)	0.7485(0.0472)	0.7239(0.0351)

[표 6]의 결과를 보면, 레이블 노이즈가 분류가 어려운 개체에 존재하는 경우 또한 Accuracy, G-mean, AUC 모두 원자료를 사용한 결과가 가장 높았다. [그림 12]와 [그림 13]는 레이블 노이즈가 존재하는 원자료, Up-샘플링 자료를 바탕으로 형성된 LDA의 판별구역이다. 이 그림에서 0은 그룹 1의 개체를, 1은 그룹 2의 개체를 의하며 빨간색으로 표시된 개체는 오분류된 개체를 의미한다. 이 그림을 보면 분류가 어려운 개체에 레이블 노이즈가 존재할 때 Up-샘플링자료를 사용하면 판별구역이 레이블 노이즈에 민감하게 반응하는 것을 확인할 수 있다.

이 경우에는 원자료를 사용한 결과가 높은 성능을 보이고 원자료를 사용한 KNN이 가장 좋은 성능을 보였다.



[그림 12] 레이블 노이즈가 분류 어려운 개체에 존재하는 원자료의 LDA 판별구역



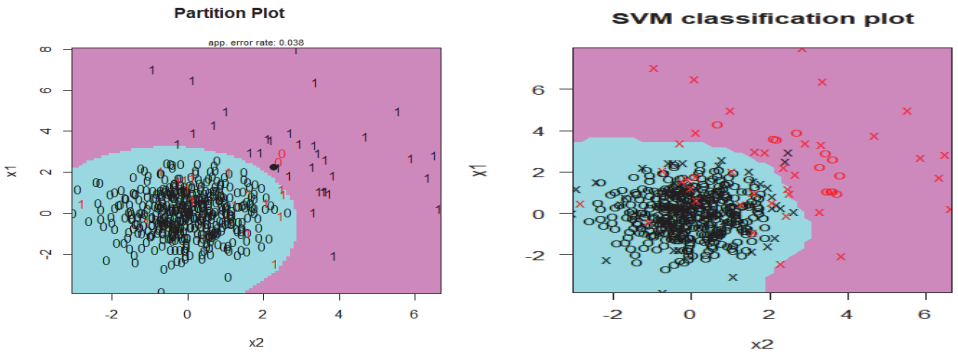
[그림 13] 레이블 노이즈가 분류 어려운 개체에 존재하는 Up-샘플링자료의 LDA 판별구역

(2) 시나리오 2의 적용결과 및 해석

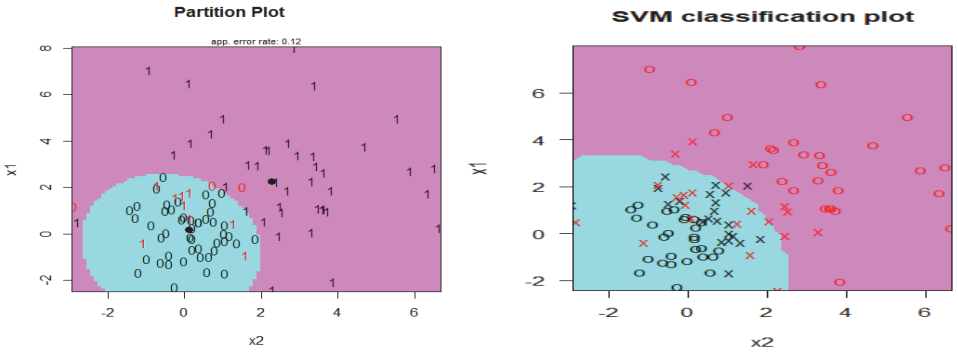
[표 7] 레이블 노이즈가 존재하지 않는 경우 분류 성능 비교

Model	Data	G-mean	AUC	Accuracy
LDA	Original	0.7493(0.0466)	0.7800(0.0349)	0.9495(0.0077)
	Down	0.8278(0.0353)	0.8331(0.0338)	0.8331(0.0338)
	Up	0.8282(0.0352)	0.8336(0.0315)	0.8336(0.0315)
QDA	Original	0.7917(0.0360)	0.8125(0.0289)	0.9556(0.0067)
	Down	0.8547(0.0356)	0.8570(0.0351)	0.8570(0.0351)
	Up	0.8603(0.0319)	0.8628(0.0298)	0.8628(0.0298)
KNN	Original	0.6175(0.0692)	0.6929(0.0422)	0.9383(0.0085)
	Down	0.8352(0.0320)	0.8414(0.0301)	0.8414(0.0301)
	Up	0.8390(0.0341)	0.8427(0.0321)	0.8427(0.0321)
SVM	Original	0.7540(0.0492)	0.7843(0.0366)	0.9522(0.0070)
	Down	0.8444(0.0351)	0.8485(0.0337)	0.8485(0.0337)
	Up	0.6946(0.0679)	0.7230(0.0534)	0.7230(0.0534)

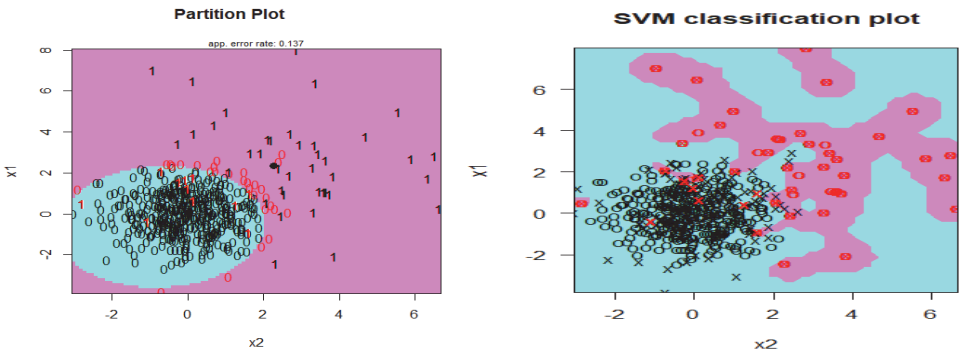
[표 7]의 결과를 보면, 레이블 노이즈가 존재하지 않는 경우 Accuracy는 원자료를 사용한 결과가 가장 높았다. 하지만 AUC는 SVM을 제외하고 Up-샘플링자료를 사용한 결과가 가장 높았다. 따라서 Up-샘플링자료를 기준으로 보면 QDA가 가장 높은 AUC 값을 가졌다. SVM의 경우는 Up-샘플링자료의 G-mean, AUC가 원자료를 사용한 결과보다 낮은데, 이는 Up-샘플링자료를 바탕으로 형성된 판별구역이 원자료와 Down-샘플링자료를 바탕으로 형성된 판별구역에 비해 그룹 2의 개체에 민감하게 형성되기 때문으로 예상된다([그림 14]-[그림 16]). SVM의 판별구역 그림 속 검은색 개체는 그룹 1의 개체이며 빨간색 개체가 그룹 2의 개체를 의미한다.



[그림 14] 레이블 노이즈가 없는 원자료의 QDA와 SVM 판별구역



[그림 15] 레이블 노이즈가 없는 Down-샘플링자료의 QDA와 SVM 판별구역



[그림 16] 레이블 노이즈가 없는 Up-샘플링자료의 QDA와 SVM 판별구역

[표 8] 레이블 노이즈가 다수 그룹에만 존재하는 경우 분류 성능 비교

Model	Data	G-mean	AUC	Accuracy
LDA	Original	0.7493(0.0466)	0.7800(0.0349)	0.9495(0.0077)
	Down	0.7954(0.0339)	0.7975(0.0336)	0.7762(0.0371)
	Up	0.7949(0.0346)	0.7970(0.0348)	0.7767(0.0335)
QDA	Original	0.7917(0.0360)	0.8125(0.0289)	0.9556(0.0067)
	Down	0.8466(0.0324)	0.8472(0.0323)	0.8499(0.0311)
	Up	0.8504(0.0356)	0.8513(0.0348)	0.8567(0.0255)
KNN	Original	0.6175(0.0692)	0.6929(0.0422)	0.9383(0.0085)
	Down	0.7979(0.0430)	0.7998(0.0427)	0.8136(0.0482)
	Up	0.7830(0.0432)	0.7844(0.0428)	0.7822(0.0424)
SVM	Original	0.7540(0.0492)	0.7843(0.0366)	0.9522(0.0070)
	Down	0.8271(0.0450)	0.8321(0.0428)	0.8605(0.0590)
	Up	0.6869(0.0781)	0.6952(0.0702)	0.7329(0.0496)

[표 8]의 결과를 보면, 레이블 노이즈가 다수 그룹에만 존재하는 경우 또한 Accuracy는 원자료를 사용한 결과가 좋았다. 하지만 G-mean, AUC은 QDA를 제외하고 Down-샘플링자료를 사용한 결과가 가장 높았다.

이 경우에는 Up-샘플링자료를 사용한 QDA 결과가 높은 성능 보였다.

[표 9] 레이블 노이즈가 두 그룹 모두에 존재하는 경우 분류 성능 비교

Model	Data	G-mean	AUC	Accuracy
LDA	Original	0.7211(0.0528)	0.7602(0.0385)	0.9478(0.0078)
	Down	0.7956(0.0366)	0.7975(0.0363)	0.7768(0.0412)
	Up	0.7958(0.0333)	0.7977(0.0336)	0.7796(0.0301)
QDA	Original	0.8110(0.0437)	0.8274(0.0361)	0.9538(0.0089)
	Down	0.8462(0.0316)	0.8470(0.0312)	0.8546(0.0309)
	Up	0.8511(0.0310)	0.8520(0.0301)	0.8605(0.0207)
KNN	Original	0.6243(0.0684)	0.6970(0.0421)	0.9387(0.0082)
	Down	0.7852(0.0406)	0.7866(0.0405)	0.7948(0.0443)
	Up	0.7855(0.0366)	0.7871(0.0368)	0.7787(0.0328)
SVM	Original	0.7290(0.0636)	0.7664(0.0452)	0.9486(0.0090)
	Down	0.8195(0.0469)	0.8244(0.0451)	0.8497(0.0451)
	Up	0.6414(0.0761)	0.6563(0.0657)	0.7154(0.0496)

[표 9]의 결과를 보면, 레이블 노이즈가 두 그룹 모두에 존재하는 경우 또한 Accuracy는 원자료를 사용한 결과가 좋았다. 하지만 G-mean, AUC은 SVM을 제외하고 Up-샘플링자료를 사용한 결과가 가장 높았다.

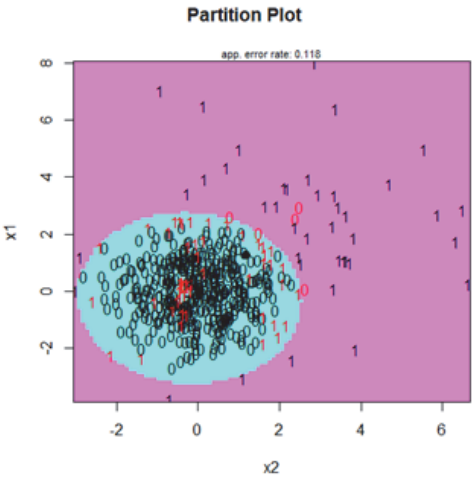
이 경우에는 Up-샘플링자료를 사용한 QDA 결과가 높은 성능를 보였다.

[표 10] 분류가 어려운 개체 레이블 노이즈가 존재하는 경우 분류 성능 비교

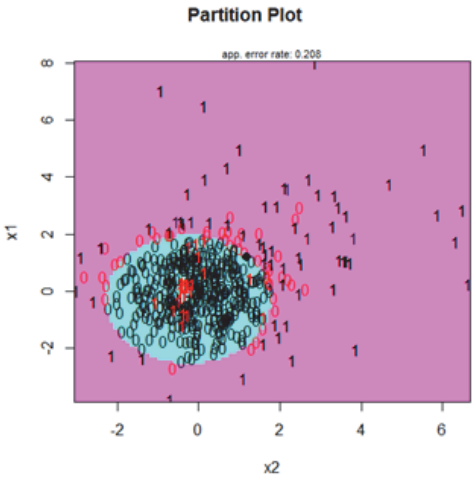
Model	Data	G-mean	AUC	Accuracy
LDA	Original	0.7664(0.0535)	0.7926(0.0409)	0.9483(0.0091)
	Down	0.7924(0.0428)	0.7964(0.0420)	0.7654(0.0494)
	Up	0.7917(0.0414)	0.7954(0.0402)	0.7658(0.0477)
QDA	Original	0.8375(0.0357)	0.8476(0.0310)	0.9478(0.0112)
	Down	0.7987(0.0451)	0.8025(0.0423)	0.7722(0.0572)
	Up	0.8033(0.0445)	0.8064(0.0431)	0.7803(0.0537)
KNN	Original	0.7324(0.0736)	0.7675(0.0526)	0.9391(0.0154)
	Down	0.7864(0.0457)	0.7891(0.0448)	0.7675(0.0520)
	Up	0.7827(0.0440)	0.7863(0.0433)	0.7568(0.0487)
SVM	Original	0.8066(0.0471)	0.8166(0.0398)	0.9044(0.0399)
	Down	0.7644(0.0451)	0.7709(0.0425)	0.7285(0.0554)
	Up	0.7431(0.707)	0.7464(0.0671)	0.7582(0.0519)

[표 10]의 결과를 보면, 분류가 어려운 개체 레이블 노이즈가 존재할 경우 또한 Accuracy는 원자료를 사용한 결과가 좋았다. 하지만 G-mean, AUC는 QDA와 SVM의 경우에만 원자료를 사용한 결과가 가장 높았다. [그림 17]과 [그림 18]은 레이블 노이즈가 존재하는 원자료, Up-샘플링자료를 바탕으로 형성된 QDA의 판별구역이다. 이 그림을 보면 분류가 어려운 개체에 레이블 노이즈가 존재하는 경우 Up-샘플링자료를 사용하면 판별구역이 레이블 노이즈에 민감하게 반응하는 것을 확인할 수 있다.

이 경우에는 원자료를 사용한 QDA 결과가 높은 성능를 보였다.



[그림 17] 레이블 노이즈가 분류 어려운 개체에 존재하는 원자료의 QDA 판별구역



[그림 18] 레이블 노이즈가 분류 어려운 개체에 존재하는 Up-샘플링자료의 QDA 판별구역

IV. 결 론

본 논문에서는 불균형자료에 레이블 노이즈가 존재할 때 샘플링방법과 레이블 노이즈의 발생 위치가 분류 성능에 어떠한 영향을 미치는지 모의실험을 통해 살펴보았다. 먼저, 시나리오 1의 자료를 바탕으로 레이블 노이즈의 발생형태, 샘플링방법, 분류방법에 따른 분류 성능을 비교하였다. 그 결과, 분류가 어려운 개체에 레이블 노이즈가 존재하는 경우는 원자료를 사용할 때 AUC가 더 높았으며 분류방법 간의 차이는 크지 않았다. 반면 다수 그룹이나 두 그룹 모두에 랜덤하게 레이블 노이즈가 존재하는 경우에는 대체로 샘플링자료를 사용할 때 높은 AUC 값을 가졌으며, 특히 SVM에서 샘플링방법 사용 유무에 따른 차이가 큰 것을 확인할 수 있었다.

다음으로 시나리오 2를 바탕으로 레이블 노이즈의 발생형태, 샘플링 방법, 분류방법에 따른 분류 성능을 비교하였다. 그 결과, 다수 그룹이나 두 그룹 모두에 레이블 노이즈가 존재하는 경우에는 Up-샘플링자료를 사용한 QDA가 가장 높은 성능을 보였지만, SVM은 Up-샘플링자료를 사용했을 때 낮은 성능을 보였다. 분류가 어려운 개체에 레이블 노이즈가 존재하는 경우에서는 QDA와 SVM은 원자료를 사용한 결과가 높은 성능을 보였고 LDA와 KNN은 Down-샘플링자료를 사용한 결과가 가장 높은 성능을 보였다. 이 중 원자료를 사용한 QDA 결과가 가장 높은 성능을 보였다.

본 논문에서는 시각화를 통해 레이블 노이즈가 불균형자료의 분류 성능에 미치는 영향을 면밀히 비교하고자 이변량 정규분포로부터 얻은 자료를 사용하였지만, 두 개 이상의 변수를 고려할 필요성이 있다. 또한 불균형 정도를 다르게 하여 불균형 정도에 따라 레이블 노이즈가 미치는 영향을 비교하는 연구도 고려할 수 있다. 따라서 향후 두 개 이상의 변수를 가지는 자료와 불균형 정도에 따라 레이블 노이즈가 분류 성능에 미치는 영향을 비교하는 연구과제를 진행하고자 한다.

참 고 문 헌

[1] Altman, N. S. (1992). An introduction to kernel and nearest neighbor nonparametric regression, *The American Statistician*, 46(3), 175-185.

[2] Cannings, T. I., Fan, Y., and Samworth, R. J. (2018). Classification with imperfect training labels, *arXiv:1805.11505*

[3] Cortes C., and Vapnik V. (1995). Support-vector networks, *Machine Learning*, 20(3), 273-297.

[4] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7(2), 179-188.

[5] Frénay, B., and Verleysen, M. (2014). Classification in the Presence of Label Noise: a Survey, *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS,* 25(5), 845-869.

[6] James, G., Witten, D., Hastie, T., and Tibshirani, R. 마이클(역). (2016). 가볍게 시작하는 통계학습 : R로 실습하는, 루비페이퍼.

[7] Kim, K. M., Jang, H. Y., and Zhang, B. T. (2014). Oversampling-Based Ensemble Learning Methods for Imbalanced Data, *KIISE Transactions on Computing Practices*, 20(10), 549-554. (in Korean).

[8] Kim, D., Kang, S., and Song, J. (2015). Classification Analysis for Unbalanced Data, *The Korean Journal of Applied Statistics*, 28(3), 495-509. (in Korean).

[9] Park, J., and Bang S. (2015). Logistic Regression with Sampling Techniques for the Classification of Imbalanced Data, *Journal of the Korean Data Analysis Society*, 17(4), 1877-1888. (in Korean).

[10] Zhu, X., and Wu, X. (2004). Class noise vs. attribute noise: A quantitative study of Their Impacts, *Artificial Intelligence Review*, 22(3), 177-210.

ABSTRACT

Classification of imbalanced data with label noise

SOYOUNG KWON

Department of Statistics

Graduate School of

Sungshin University

Label noise and imbalanced data are problematic in real data and can reduce classification performance. Therefore, in this paper, we conducted a comparative study on the classification analysis in the imbalanced data with label noise to improve classification performance. Especially, classification accuracy according to the type of label noise, sampling method, and classification method was examined using Accuracy, G-mean, and AUC. In this paper, we propose a suitable sampling method and classification method depending on the type of data and the type of label noise.