Journal of the Korean Data Analysis Society (February 2021), 23(1), 195-207.
https://doi.org/10.37727/jkdas.2021.23.1.195

불균형 자료의 분류분석 방법별 성능 비교와 접근 전략 연구

유병주[1]

요 약

불균형 자료에 대한 분류분석을 하기 위해서는 두 가지 선택의 문제에 직면하게 된다. 하나는 분류분석을 위한 모형의 선택이고 또 다른 하나는 불균형 문제를 해결하기 위한 방법의 선택이다. 그래서 이 논문에서는 훈련표본의 규모나 독립변수의 수, 불균형 정도 등과 같은 데이터의 특징을 고려한 불균형 자료에 대한 순차적인 접근 전략 문제를 다루었다. 이를 위해 이진 분류분석의 대표적인 모형인 로지스틱 회귀모형, 서포트벡터 머신, 딥러닝 방법을 자료의 특성에 따른 분류 성능을 비교하기 위한 이론적 고찰과 모의실험을 시행하였다. 그리고 자료의 불균형을 해결하기 위한 개선 방법들과 조합했을 때 Tukey의 다중비교를 통하여 분류 성능이 좋은 최적의 결과를 얻기 위한 접근 전략을 식별하기 위한 모의실험을 하였다. 모의실험 결과 자료의 특성 중 훈련표본의 수량과 불균형 여부가 지배적인 요소로 작동되는 것을 확인할 수 있었으며, 훈련표본이 적은 경우는 로지스틱 회귀모형으로 접근하여 과대추출 방법으로 자료의 불균형 문제를 해결하는 방법이 좋고, 훈련표본이 많은 경우는 딥러닝 방법으로 접근하여 가중치 방법이나 과소추출 방법으로 자료의 불균형을 개선하는 방법이 성능이 우수한 추정 결과를 얻을 수 있는 접근 전략임을 확인하였다.

주요용어 : 불균형 자료, 로지스틱 회귀모형, 서포트벡터 머신, 딥러닝, Tukey의 다중비교.

1. 서론

이진 분류분석은 다양한 특징을 정의하는 독립변수들을 이용하여 정상과 비정상, 질병의 유무, 사기 거래 여부 등과 같이 두 집단 중 어느 집단으로 분류할 것인지를 예측하는 모형을 구축하는 것이다. 이러한 모형은 일반적으로 분류의 정확도를 향상하고자 노력하는데 이는 자료의 특징과 성능평가 기준에 따라 고려할 사항이 많다. 특히 분류 성능을 평가하는 척도는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall) 등 기준이 다양하고 상반되는 경우가 있어, 자료의 특징이나 집단별 분류 우선순위, 분류 목적 등을 고려하여 기준을 적용하게 된다.

통상적인 분류분석 방법에는 전통적인 로지스틱 회귀분석, 머신러닝을 이용한 서포트벡터 머신(Support Vector Machine), 인공신경망을 활용한 딥러닝 등이 있다. 이러한 방법들은 획득한 표본의 수, 독립변수의 수, 두 집단 간 자료의 불균형 정도 등 데이터 특성에 따라 분류 성능이 변화하기 때문에 상황에 따라 분류 성능이 좋은 방법을 선택하기 위한 접근 전략이 필요하다. 특히 이진 분

[1]17018 경기도 용인시 처인구 성산로 57 사서함505-1-3, 육군 지상작전사령부 작전분석과장.
E-mail : enjoystat@gmail.com
[접수 2021년 1월 4일; 수정 2021년 1월 14일, 2021년 2월 7일; 게재확정 2021년 2월 9일]

류분석에서 불균형 자료를 이용하여 분류분석 모형을 학습시키는 경우 다수집단 표본이 최적화 알고리즘을 지배하게 되어 소수집단에 대한 오분류가 증가하는 현상이 발생할 수 있다.

그래서 데이터의 특성을 고려하여 분류분석 모형을 선택하고 불균형으로 유발되는 문제를 해결할 수 있는 적절한 방법을 선택하여 분류 성능이 최적화된 추정 방법을 선택하기 위한 접근 전략을 알아보는 것은 매우 의미 있다고 할 수 있다. 이러한 상황을 고려해서 우선 분류분석을 위한 성능평가 지표와 불균형으로 유발되는 문제점을 알아보고, 분류분석 모형들의 특징과 장단점에 관하여 연구하며, 불균형 문제를 해결하기 위한 다양한 접근 방법들을 검토할 것이다. 그리고 이러한 모든 상황을 고려한 모의실험을 시행하여 최적의 접근 전략을 제시하고자 한다.

2. 성능평가 지표와 불균형 문제

2.1. 성능평가 지표

이진 분류분석 결과를 요약하면 Table 1과 같이 표현할 수 있다. 실제와 예측의 결과가 일치된 경우를 True라고 하고 그렇지 않은 경우를 False라고 했을 때 FP 및 FN 오류를 최소화하는 것이 우리의 목표일 것이다.

Table 1. Confusion matrix

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

$$\text{Precision} = \frac{TP}{TP+FP},\ \text{Recall(TPR)} = \frac{TP}{TP+FN},\ \text{FPR} = \frac{FP}{TN+FP}$$

정밀도(Precision)는 양성으로 예측된 것 중에 실제 양성인 경우의 비율을 측정하는 지표이다. 이는 실제 음성인데 양성으로 예측한 경우를 분모에 포함하기 때문에 실제 양성과 음성의 비에 영향을 받아 다수집단과 소수집단 간 불균형으로 인한 영향에 민감한 척도라고 말할 수 있다. 재현율(Recall)은 진양성률(TPR, True Positive Rate)이라고도 하는데 실제 양성인 것 중에 양성으로 예측된 비율을 의미하고, 위양성률(FPR)은 실제 음성인 샘플 중에 양성으로 잘못 예측된 경우의 비율을 의미한다. 정확도(Accuracy)는 전체 샘플 중에 정확히 분류된 TN과 TP의 비율인데, 불균형 자료의 경우 단순히 정확도를 이용해 분류 성능을 평가하는 것은 다소 문제가 있다. 예를 들어 소수집단의 데이터가 전체의 1%라면 단순하게 모든 데이터를 음성으로 예측해도 99%의 정확도를 얻을 수 있기 때문이다. 그래서 불균형 데이터의 경우 모형의 성능을 적절히 평가하기 위한 또 다른 척도가 필요하다.

이러한 지표들을 보완하는 방안들을 Seliya et al.(2009), Wang et al.(2016), Davis, Goadrich(2006) 등이 연구하여 제시하였는데, 특히 Davis, Goadrich(2006) 주장에 따르면 ROC(Receiver Operating Characteristic) 곡선이 불균형 데이터의 분류 성능을 측정하는데 가장 적절하다고 하였다. 또한 Provost, Fawcett(1999), He, Garcia(2009)도 ROC의 유용성을 제시하였는데, ROC 곡선은 위양성률(FPR)에 대한 진양성률(TPR)을 도식하는 방법으로 올바르게 분류한 양성 표본(TP)과 오분류된 양

성 표본(FP)의 균형을 확인할 수 있기 때문이다. 그래서 Weng, Poon(2008)은 ROC 곡선을 이용해 곡선의 아래 영역을 수치적으로 계산하여 모형 간의 성능을 비교할 수 있는 척도인 AUC(Area Under the Curve)를 제시하였다. 이는 위양성률을 최소화하면서 진양성률을 증가시킬 수 있는 척도를 평가하는 데 유용하게 활용된다.

그래서 이러한 장점들을 종합적으로 고려하여 차후 모형에 대한 평가는 분류분석 결과에 대해 종합적인 척도인 ROC AUC를 기준으로 평가하기로 한다.

2.2 집단간 자료의 불균형과 접근 전략 문제

Krawczyk(2016)와 Seifert et al.(2007)은 불균형 자료에서 소수집단의 데이터 비율보다 사용 가능한 절대적인 표본 수량이 더 중요하다고 하였다. 예를 들면 백만 개의 표본을 포함하는 데이터 세트에서 소수집단이 1%에 불과한 경우라고 해도 모형을 학습시킬 수 있는 균형된 훈련표본은 10,000개가 가용하기 때문이다. Krawczyk(2016)은 데이터의 균형 또는 불균형 문제 이전에 집단 간의 특징을 잘 구별할 수 있는 독립변수들을 선택하여 모형을 구축하는 것이 선행되어야 한다고 주장했다. 다시 말해 데이터들의 위상이 겹치지 않고 명확히 구별될 수 있는 독립변수들을 선택하여 모형을 개발하면 데이터의 불균형과 관계없이 좋은 예측 결과를 얻을 수 있다.

그러나 현실적으로 독립변수들의 특징이 명확히 구분되는 모형을 개발하기도 어렵고, 어떤 경우에는 이미 모형과 수집된 데이터가 정해져 있어 새로운 자료를 수집한다는 것이 현실적으로 불가능한 때도 있다. 이 경우 소수집단과 다수집단 간 여러 가지 특징이 겹치는 모형이 수립될 수 있는데 모형의 복잡도가 증가하면 할수록 모형을 학습시킬 훈련 데이터의 불균형에 대한 민감도가 증가한다(Japkowicz, 2000).

Kubat et al.(1998), Seifert et al.(2007), Weiss(2004), Buda et al.(2018) 등은 자료의 불균형 비율에 따른 분류 성능의 차이에 관한 연구를 수행하였다. 그들에 의하면 데이터의 불균형은 분류 성능에 전반적으로 악영향을 미치며, 데이터의 불균형으로 인한 문제를 해결하는 방법은 과대추출 방법이 지배적이고 과대추출 방법은 일부 기계학습 모형과 달리 합성신경망을 이용한 딥러닝에서는 과적합 현상이 발생하지 않았다고 하였다.

그래서 접근 전략을 검토할 때는 먼저 표본의 수량과 독립변수들의 수를 고려하여 어떠한 모형을 선택할 것인가를 검토한 후 불균형 문제 해결 방법을 검토해야 할 것이다.

3. 분류분석을 위한 일반적인 모형 선택 기준

3.1. 로지스틱 회귀모형

로지스틱 회귀모형은 두 집단으로 분류가 가능한 이진형의 자료를 가지고 있는 경우 관심이 있는 어느 한쪽 그룹에 속할 확률을 종속변수로 변환하여 독립변수들의 선형결합으로 그 발생 가능성을 예측하는 데 사용되는 통계기법이다. 예를 들면 기업의 재무제표를 기준으로 기업의 도산 확률을 예측한다든지, 신용카드의 여러 가지 거래 형태와 관련된 자료를 기준으로 사기 신용거래일 확률을 예측하거나, 환자들의 다양한 의학적 데이터를 가지고 특정 암에 걸렸을 확률을 예측하는 등 다양한 분야에서 활용되고 있다.

신용카드의 거래 형태를 예를 들면 정상 거래를 0, 사기 거래를 1이라고 한다면 설명변수 X에 의존하는 사기 거래일 확률 $P(Y=1|x)$에 대한 확률 모수를 $\pi(x)$라고 하면 $\pi(x)$의 logit 변환 값은 실수 공간에 존재하게 되고, 모수 $\pi(x)$는 [0, 1] 값을 가지며 독립변수의 선형결합 형태로 표현할 수 있으며 아래와 같다.

$$\text{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha+\beta x,\ \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$

로지스틱 회귀모형은 무엇보다 계산이 간편하다는 것과 독립변수에 의한 종속변수의 변화에 대한 해석과 설명이 비교적 쉽고, 특정 집단에 속할 수 있는 확률을 예측할 수 있다는 장점이 있다. 그러나 데이터 전처리 과정이 다른 방법에 비해서 비교적 복잡하고 선형분류만 가능하다는 한계가 있다. 그런데도 이러한 방법이 선호되는 것은 비교적 계산이 쉽고 해석과 설명이 쉽다는 특징 때문에 광범위한 분야에서 활용되고 있다.

3.2. 서포트벡터 머신 모형

서포트벡터 머신은 기계학습의 하나로 패턴인식, 자료분석을 위한 지도학습 모델이며, 주로 분류분석과 회귀모형에 사용한다. 서로 다른 집단에 속한 데이터 간의 간격을 마진(Margin)이라고 하고 이 간격을 구하기 위한 기준이 되는 데이터들을 서포트벡터라고 하는데, 이 방법은 마진이 최대가 되는 선이나 평면을 찾아 이를 기준으로 어느 집단에 속할지 판단하는 비확률적 분류모형이다. 서로 다른 집단의 데이터간 간격이 최대가 되는 평면을 초평면(Maximum Margin Hyperplane)이라고 하며 선형적 분류는 물론 비선형적 분류까지 확장할 수 있다. 이런 경우 고차원으로 자료를 바꿔서 두 그룹을 분류할 수 있는 평면을 찾아내는데 이를 위해서는 적절한 알고리즘을 적용할 커널을 선택해야 한다. 커널은 데이터 간의 거리를 측정하는 방법으로 상황에 따라 선형커널, 시그모이드 커널, 가우시안 RBF(Radial Basis Function) 커널 등 다양한 커널을 적용할 수 있다.

서포트벡터 머신은 특이점에 대하여 민감하지 않고 표본 수에 비하여 독립변수의 수가 상대적으로 많거나 자료가 고차원인 경우 효과적인 방법으로 선호되고 있다. 그러나 학습 시간이 상대적으로 많이 걸리고 서로 다른 집단 간 자료들의 중첩이 많은 경우 분류가 효과적이지 못하다는 한계가 있다. 특히 효과적인 분류를 위해서 적절한 커널을 선택해야 하는데 적절한 커널을 선택하는 기준이 모호하여 다소 까다로운 부분이 있다.

3.3. 딥러닝 모형

신경망을 이용한 딥러닝 방법은 분류분석을 위한 모형을 학습시키는 과정과 분류분석을 시행하는 과정으로 이루어진다. 모델 학습과정에서 고려해야 할 사항은 레이어 구성과 활성화 함수 선택, 최적화 방법 선택, 손실함수 정의 등이 있다. Goodfellow et al(2016)에 의하면 다양한 분야에서 경험적으로 확인한 결과 레이어의 구성은 깊이가 깊어질수록 더 나은 결과를 얻을 수 있다고 하였다. 그러나 레이어의 깊이를 깊게 할수록 추정해야 할 모수가 증가하기 때문에 많은 표본이 필요하며, 모형 학습 시간이 더 오래 걸리는 단점이 있다. 이진 분류분석에 활용되는 활성화 함수는 통상적으로 Input Layer와 Hidden Layer에서는 ReLU를 Output Layer에서는 Sigmoid 함수를 사용한다.

모형에서 추정해야 할 모수를 비교하면 회귀모형은 수 개의 모수만 추정하면 되는 데 비하여

딥러닝은 추정해야 할 모수들이 레이어를 추가할 때마다 기하급수적으로 늘어나는 경향이 있다. 예를 들어 두 개의 독립변수를 가진 회귀모형의 경우 추정해야 할 모수는 3개지만, 딥러닝의 경우 16개 노드의 입력층, 16개 노드의 2개 은닉층, 1개 노드의 출력층을 구성하면 609개의 모수를 추정해야 한다.

딥러닝에 의한 분류분석은 사용자 그룹이 방대하고 다양한 알고리즘을 적용할 수 있어서 훈련 표본이 많은 모델을 학습하여 최신화시키기에 유용하다. 반면 추론 과정을 설명하거나 해석하기가 복잡하고 디버그 과정이 복잡하여 최초 사용자에게는 접근하기 어려울 뿐만 아니라 복잡한 구조를 구현하기 어려운 점이 있다.

3.4. 모형별 장단점 비교

세 가지 모형별 장단점을 종합하면 회귀모형은 설명 및 해석이 용이하여 표본이 적은 모형을 정교하게 구축할 때 사용되는 것이 좋고, 서포트벡터 머신은 다차원 분석에 유리하지만, 컴퓨팅 시간이 많이 소요되는 단점이 있다.

Table 2. Comparing Binary Classification Models

Model	Pros	Cons
LR	Efficient computation Easy to interpret and explain Thrive with little training Flexible model	Linear division boundary Higher data maintenance Regulation Overfiting possibility
SVM	Effective in the higher dimension Less impact from outliers Effective in larger amount of features	Higher training time Bad in case of overlapped classes Tricky to select the proper kernel function
DL	Easy to train large model Regular update Various algorithms Large user community	Difficult to interpret and explain Complex to debug Hard to use for new developers Difficult to implement complex architecture

새로운 모형을 개발하고자 할 때는 표본이 많고 모형에 대한 확신이 있는 경우 딥러닝 방법으로 접근하는 것이 좋을 것이다. 그러나 모형에 대한 확신이 없는 탐색적 방법론을 적용할 때 초기부터 딥러닝을 선택하면 모형에 대한 설명이나 해석이 곤란하므로 소규모 표본을 이용해 선형모형으로 모형을 정교하게 검증한 후 대규모 표본을 이용하여 딥러닝 방법으로 모형을 업데이트해 나가는 방법이 이상적일 것이다.

4. 불균형 문제 해결을 위한 추정 방법의 선택

4.1. 가중치에 의한 최적화 알고리즘 개선 방법

로지스틱 회귀모형은 등분산성 가정하에 최소제곱법으로 그 추정치를 추정한다. 그런데 등분산성 가정을 위반하는 상황에서 최소제곱법을 사용하면 회귀계수의 추정량은 불편성은 유지하지만, 그 분산이 최소화되지 않아 통계적 정확도 측면에서 추정량이 최적성을 유지하지 못하게 된다 (Chatterjee & Hadi, 2012). 그래서 이분산성을 해소할 수 있는 가중치를 적용한 최소제곱추정법을 적용하면 보다 나은 추정값을 얻을 수 있다.

로지스틱 회귀모형에서 두 집단의 분산이 다르다는 조건으로 접근하기 위해서는 양성에 해당하는 집단의 개체 수를 n_1, 양성에 해당하는 확률 모수를 $\pi(x)$라고 했을 때 양성 집단에 해당하는 가중치 w_1은 아래와 같다.

$$w_1 = n_1\pi(x)(1-\pi(x)),\ w_0 = n_0\pi(x)(1-\pi(x)).$$

그리고 음성에 해당하는 가중치 w_0는 위 식에서 음성에 해당하는 집단의 개체 수 n_0만 대체해 주면 될 것이다. 결국 두 집단 간 관찰 개체의 불균형은 두 집단 관찰 개체의 퍼짐의 정도인 분산이 서로 다르므로 등분산성 가정을 위반하는 경우 분산의 역수를 가중치로 적용하여 추정하면 된다. 이러한 방법은 서포트벡터 머신이나 딥러닝의 경우도 유사하게 적용될 수 있어 집단별 가중치를 부여하여 최적화 알고리즘에 반영하면 보다 나은 추정이 가능하다.

4.2. 과소추출 방법

과소추출 방법은 소수집단과 비교하여 다수집단의 초과된 표본 수를 제거하여 두 집단의 표본 수를 유사한 수준으로 축소하는 방법이다. 가장 쉽고 단순한 방법으로 Random Under-sampling 방법이 있으며 이는 다수집단의 개체를 랜덤으로 선택하여 제거함으로써 모형 적합 속도가 증대시키는 방법이다. 반면 이 방법은 다수의 개체가 제거됨에 따라 정보 손실이 발생할 수 있는 단점을 가지고 있다(Drummond, Holte, 2003).

소속 집단에 대한 정보가 없는 경우 데이터를 군집화할 때 사용하는 KNN(Nearist Neighbours)알고리즘은 K개의 인접한 데이터 중에 많은 수를 차지하는 군집으로 군집화시키는 것을 말한다(Cover, 1967). 그래서 데이터를 군집화하는 KNN 알고리즘을 이용하여 다수집단에서 과소추출하는 방법으로 고안한 것이 ENN(Edited Nearist Neighbours) 방법이다. 이는 다수집단의 K개의 관찰치를 선택하고, 선택된 표본 중에 오분류되거나 상대 집단과 가장 근접한 표본을 우선적으로 제거하여 다수집단 자료를 과소추출하는 방법이다(Wilson 1972).

4.3. 과대추출 방법

과대추출 방법은 과소추출 방법과 반대로 소수집단의 부족한 표본수를 다수집단과 균형을 맞추기 위해 소수집단 표본을 반복 추출하는 방법이다. 단순하게 소수집단에서 부족한 표본 수만큼 임의로 반복 추출하여 다수집단 표본 수만큼 확보하는 방법을 생각할 수 있는데, 이는 임의 과대추출 방법으로 Random Over-sampling이라고 한다. 그러나 임의 과대추출 방법은 소수집단의 개체에 대하여 반복 추출하기 때문에 과적합 문제를 초래할 수 있다.

이러한 과적합 문제를 해결할 수 있는 다른 과대추출 방법은 Chawla et al.(2002)이 제안한 SMOTE(Synthetic Minority Over-sampling Technique)이 있다. 이 방법은 소수집단에서 최근접한 K개를 선택하여 그 개체들 사이의 가상 공간상에 새로운 개체를 생성시켜 반영하는 방법이다.

4.4. 과대 및 과소추출 혼합방법

Chawla et al.(2002)에 의하면 위음성(FN) 오류는 위양성(FP) 오류보다 훨씬 비용이 큰 경우가 많으며 다수집단에서 과소추출하는 방법은 소수집단 분류 민감도를 높이는 좋은 수단이 되며, 특히

소수집단에서 과대추출을 하고 다수집단에서 과소추출을 하는 방법을 조합하면 성능이 더 좋아진다고 하였다.

혼합방법은 위에서 제시한 소수집단에 대한 과대추출과 다수집단에 대한 과소추출을 결합하여 시행함으로써 두 방법의 장점을 극대화하는 방법이다. 대표적으로 적용하는 방법은 SMOTE-ENN 방법(Batista et al., 2003)과 SMOTETomek 방법(Batista et al., 2004)이 있다. SMOTE-ENN 방법은 소수집단은 SMOTE 방법으로 과대추출을 하고, 다수집단은 ENN 방법으로 과소추출을 하는 방법이며, Tomek 과소추출 방법은 다수집단과 소수집단 표본 간 인접한 거리를 측정하여 거리가 작은 다수집단의 표본을 제거하는 방법이다.

5. 모의실험과 데이터별 접근 전략

5.1. 모의실험 방법과 조건

앞에서 설명한 분류분석 방법을 간단하게 회귀모형(LR), 서포트벡터 머신(SVM), 딥러닝(DL)으로 표기하여 기본적인 접근방법(B)을 기준으로 불균형에 대한 개선된 추정 방법을 가중치 적용(W), 과소추출(U), 과대추출(O), 혼합방법(C)으로 표기하고, 두 가지를 결합한 형태를 디자인하면 아래의 Table 3과 같이 15가지 추정 방법으로 정리할 수 있다.

Table 3. Estimation Method Matrix

Model	Base	Weighted estimation	Under-sampling	Over-sampling	Combined method
LR	LB	LW	LU	LO	LC
SVM	SB	SW	SU	SO	SC
DL	DB	DW	DU	DO	DC

그리고 표본의 변화, 독립변수의 변화, 불균형 유무 등에 대한 추정 방법별 성능을 비교하기 위해 Table 4와 같이 모의실험을 계획하였다. 여기서 표본수를 4가지 경우로 적용한 것은 추정 방법별 차이가 있을 수 있는 최소/최대 표본을 포함하고, 빈번하게 있을 수 있는 중간 상황들을 비교하기 위해 계획하였으며, 불균형 정도는 다수집단과 소수집단의 비율이 50:50으로 불균형이 없는 경우와 90:10으로 불균형이 심각한 상황을 고려하였다.

Table 4. Experimental Design Matrix

Number of Training Samples		Variables	Imbalanced ratio
100	5,000	2	50:50
50,000	100,000	20	90:10

3가지 조건에 대하여 요인설계(Factorial Design)를 적용하면 총 16가지 실험 Case가 만들어 지고, 이를 Table 3에서 제시한 추정 방법과 결합하여 모형별 기본방법을 적용하고, 불균형 데이터인 경우 모형별 불균형 해소방법 4가지를 적용하여 30회씩 반복하여 실험하였다.

예를 들면 훈련 표본수 100개, 독립변수 2개, 불균형 정도 90:10인 경우를 설명하면 Sklearn 라이브러리의 make_classification 함수를 이용하여, 다수집단은 임의의 이변량 정규분포를 기준으로 990

개의 표본을 생성하고, 소수집단은 또 다른 이변량 정규분포에서 110개를 랜덤으로 생성하였다. 그래서 생성된 총 1,100개의 표본 중에 100개의 표본은 훈련표본으로 사용하고 나머지 1,000개는 추정 결과를 평가하는 시험표본으로 활용하였다. 모의실험을 위한 소프트웨어는 Python Version3.9의 Sklearn, Tensorflow 라이브러리를 활용하였으며 로지스틱 회귀모형, 서포트벡터 머신, 딥러닝을 실험하였다.

5.2. 표본수와 독립변수 증가에 따른 추정 방법별 성능의 차이

추정 방법별 성능의 차이는 2.2절에서 설명한 것처럼 불균형 데이터의 분류 성능 비교에 적절한 FPR과 TPR을 기준으로 한 ROC AUC를 기준으로 비교하였다. 모의실험 결과 다양한 조건별 차이를 명확히 식별하기 위해 분산분석을 실시하였으며, 그 결과 유의한 차이가 있는 경우나 상호비교를 위해 필요한 경우만 발췌하여 Table에 제시하였으며 세부적으로 95% 신뢰수준에서 Tukey의 다중비교 하한값, 상한값, 그리고 p-value를 동시에 제시하였다.

Table 5. Tukey multiple comparisons of means in 95% confidence level

Samples	Pair	Difference	LB	UB	p-value
overall	L-D	-0.025	-0.032	-0.018	<0.01
	S-D	-0.029	-0.036	-0.022	<0.01
	S-L	-0.004	-0.011	0.003	0.39
Small Data Case (100 samples)	L-D	0.044	0.023	0.064	<0.01
	S-D	-0.005	-0.016	0.025	0.84
	S-L	-0.039	-0.059	-0.018	<0.01
20 Variables Case	L-D	-0.012	-0.024	-0.001	0.04
	S-D	-0.034	-0.046	-0.022	<0.01
	S-L	-0.021	-0.034	-0.010	<0.01

모의실험 결과 분류분석 모형별 분류 성능의 차이는 Table 5에 제시한 바와 같이 전반적으로 딥러닝 방법이 서포트벡터 머신이나 로지스틱 회귀모형보다 유의수준 5%를 기준으로 우수하였다. 그러나 표본수가 100개인 소규모 표본의 경우는 로지스틱 회귀모형이 다른 방법에 비하여 우수한 성능을 보였다. 또한 독립변수가 20개인 경우에는 딥러닝이 분류성능이 가장 우수하고, 그 다음은 로지스틱 회귀모형이며, 서포트벡터 머신의 분류 성능이 가장 저조한 것으로 확인 되었다. 참고로 2.4절의 분류분석 방법별 일반적 특징과 상반되는 부분은 서포트벡터 머신이 독립변수의 수가 증가할 때 유용한 방법이라는 주장이 있었으나 실제 실험 결과는 그렇지 않다는 점을 확인할 수 있었다.

기본적으로 딥러닝으로 모형을 구축하는 경우 레이어를 추가할 때마다 추정해야 할 모수의 수는 기하급수적으로 늘어난다. 즉 모형을 구축하기 위한 훈련표본이 충분하게 많은 경우 딥러닝 방법으로 접근하는 것이 좋지만 그렇지 않고 소수의 훈련표본을 가지고 있는 경우는 딥러닝 방법보다 로지스틱 회귀모형으로 접근하는 것이 유리하다는 것을 확인할 수 있었다. 그리고 독립변수의 증가는 추정해야 할 모수의 증가를 의미하기 때문에 어떠한 모형으로 분류분석을 하더라도 추정 성능은 저하된다. 하지만 추정 성능의 저하 속도는 모형별로 다소 상이하여 여기서 확인한 것처럼 딥러닝은 서포트벡터 머신보다 저하되는 속도가 크지 않다는 것을 알 수 있다.

5.3 불균형 데이터에 대한 해결방법 간의 성능 비교

여기서는 불균형이 있는 경우 분류분석 모형별로 실험 결과를 종합적으로 분석하고, 불균형으로 발생할 수 있는 문제점을 해소하는 방법들의 성능을 비교하도록 한다. 그래서 전반적인 데이터를 이용하여 분산분석을 실시하고, 훈련표본이 적은 경우와 많은 경우로 구분하여 분산분석을 하였다. 전반적인 경우와 훈련표본이 많은 경우는 분류분석 모형의 선택과 불균형에 대한 추정방법의 교호효과는 유의하지 않았지만, 훈련표본이 적은 경우는 교호효과가 유의하여 그 내용을 Table 6에 포함하였다.

Table 6. Tukey multiple comparisons of means in 95% confidence level

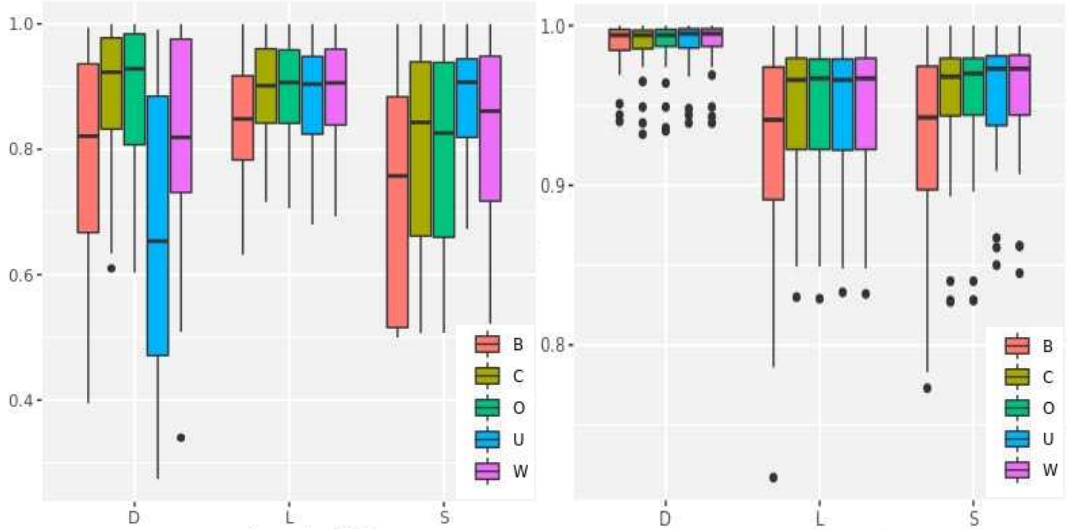
Data	Pair	Difference	LB	UB	p-value
overall	L-D	-0.019	-0.032	-0.007	<0.01
	S-D	-0.035	-0.047	-0.023	<0.01
	S-L	-0.015	-0.027	-0.003	<0.01
	W-B	0.036	0.019	0.055	<0.01
	U-B	0.026	0.009	0.044	<0.01
	O-B	0.038	0.020	0.056	<0.01
	C-B	0.038	0.020	0.056	<0.01
Small Data Case (100 Samples)	L-D	0.072	0.040	0.103	<0.01
	S-D	-0.002	-0.034	0.029	0.99
	S-L	-0.074	-0.105	-0.042	<0.01
	W-B	0.057	-0.019	0.104	0.25
	U-B	0.019	-0.028	0.067	0.80
	O-B	0.072	0.025	0.120	<0.01
	C-B	0.071	0.024	0.119	<0.01
	DU-LB	-0.207	-0.335	-0.040	<0.01
	DU-LO	-0.244	-0.412	-0.076	<0.01
	DU-LC	-0.243	-0.411	-0.075	<0.01
	LW-DU	0.235	0.017	0.403	<0.01
	LU-DU	0.215	0.047	0.383	<0.01
	LO-SB	0.173	0.005	0.341	0.03
	LC-SB	0.172	0.005	0.340	0.04
Big Data Case (100,000 Samples)	L-D	-0.047	-0.063	-0.031	<0.01
	S-D	-0.038	-0.054	-0.022	<0.01
	S-L	0.009	-0.006	0.025	0.35
	W-B	0.025	0.001	0.048	0.04
	U-B	0.024	0.001	0.049	0.04
	O-B	0.023	-0.001	0.047	0.06
	C-B	0.023	-0.001	0.046	0.06

불균형을 포함하고 있는 모든 데이터를 분석했을 때 분류분석 모형별 성능은 딥러닝이 서포트 벡터 머신이나 로지스틱 회귀모형에 비하여 우수하다. 그리고 불균형 해소방법은 가중치 방법, 과소추출, 과대추출, 혼합방법 모두가 기본방법에 비하여 우수하였다.

훈련표본이 100개인 소규모 데이터의 경우 분류분석 모형은 로지스틱 회귀모형이 가장 우수하며, 불균형을 해소하는 방법으로는 과대추출과 혼합방법이 우수하나 혼합방법은 과대추출의 일부

로 볼 수 있으므로 과대추출이 가장 성능이 우수한 것임을 알 수 있다. 특히 교호효과에 대한 비교에서 딥러닝과 과소추출이 결합된 경우 최악의 분류 성능을 보이고 있는데 이는 훈련표본이 적을 때 불균형 해소를 위해 다수표본의 일부 데이터를 버리고 분석하여 정보의 손실이 발생하는 데에서 기인하는 문제점이며, 이 경우 과소추출 방법은 적절하지 않다는 것을 알 수 있다.

반면 훈련표본이 많은 100,000개의 경우는 딥러닝이 다른 모형에 비하여 성능이 우수하고, 불균형 해소방법은 가중치 방법과 과소추출 방법이 우수함을 알 수 있다. 이는 10만 개의 훈련표본 중에 소수집단 표본이 10%라고 해도 1만 개의 훈련표본이 가능하므로 이것으로 충분히 학습시킬 수 있기 때문일 것이다. 즉 표본이 많으면 많을수록 과소추출을 해도 충분한 학습 데이터가 가용하기 때문에 이러한 결과가 나타났다고 할 수 있다.



(a) Small data case (b) Big data case

Figure 1. ROC AUC Performance comparison by classification analysis method

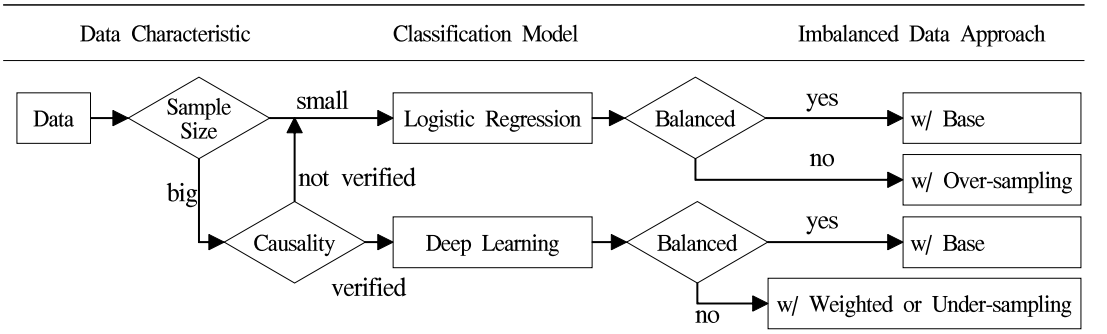
Figure 1은 소규모 표본과 대규모 표본의 경우 추정 방법별 차이를 시각적으로 보여주기 위해 정리한 그림이다. 앞에서 설명한 것처럼 소규모 데이터에서는 로지스틱 회귀모형의 과대추출과 혼합 방법이 우수하고, 대규모 데이터에서는 딥러닝의 분류 성능이 우수함을 알 수 있다.

5.4 자료 유형별 접근 전략

위에서 분석한 내용을 기초로 훈련표본의 수량과 불균형 여부에 따라서 분류분석 모형을 적용하기 위한 접근 전략을 정리하면 Table 7과 같다. 분류분석을 시도하고자 할 때 우선적으로 고려할 사항은 획득한 훈련표본의 규모이다. 획득한 훈련표본이 적으면 로지스틱 회귀모형으로 접근하고 많은 경우 딥러닝으로 접근하는 것이 좋다.

그러나 훈련표본이 많은 경우에도 모형이 검증되지 않았거나 검증이 필요할 때는 로지스틱 회귀모형으로 접근하여 모형을 검증할 필요가 있으며, 검증된 모형일 때 딥러닝을 적용하는 것이 좋다. 그리고 소규모 표본에서의 불균형 문제는 과대추출 방법으로 보완하고, 대규모 표본의 불균형 문제는 가중치 방법이나 과소추출 방법으로 해결하는 것이 선호된다.

Table 7. Approach Strategy for Applying Classification Method



6. 결론

지금까지 이진 분류분석에 있어서 자료의 불균형으로 유발되는 분류 오류를 최소화하기 위해 대표적인 분석 방법인 로지스틱 회귀모형, 서포트벡터 머신, 딥러닝을 기준으로 알아보았다. 이러한 모형들을 비교하기 위해 훈련표본의 수량과 불균형 여부를 기준으로 모의실험을 한 결과 소규모 훈련표본의 불균형 문제는 로지스틱 회귀모형으로 접근하여 과대추출 방법을 활용하여 해결하는 것이 좋으며, 대규모 훈련표본은 딥러닝으로 접근하여 가중치 방법이나 과소추출 방법으로 해결하는 접근 전략을 제시하였다.

물론 자료의 또 다른 특징에 의해 이러한 방법별 성능은 다소 차이가 있을 수 있다. 그러나 일반적으로 로지스틱 회귀모형은 독립변수와 종속변수 간의 인과관계를 식별하고 설명하기 쉬운 점과 추정하고자 하는 모수의 수가 상대적으로 적어 소량의 훈련표본을 이용해도 충분히 모형을 학습시킬 수 있다는 장점이 있다. 딥러닝의 경우는 레이어의 증가에 따라 추정해야 할 모수들이 기하급수적으로 증가하기 때문에 많은 훈련표본이 필요하고, 독립변수와 종속변수 간의 관계를 확인하기 어려우므로 가능하면 검증된 모형에 적용하는 것이 좋다.

그래서 이러한 일반적인 특징을 고려하더라도 초기 모형을 구축할 때는 소규모 훈련표본으로 로지스틱 회귀모형을 구축하여 모형을 검증하고, 이후 많은 자료를 획득하여 딥러닝 방법으로 모형을 순차적으로 확장해나가는 방법도 선호된다.

References

Bang, S. W., Kim, J. O. (2020). Sampling method using Gaussian mixture clustering for classification analysis of imbalanced data, *Journal of the Korean Data Analysis Society*, 22.2, 565-574.

Batista, G. E., Bazzan, A. L., Monard, M. C. (2003). Balancing training data for automated annotation of keywords: a case study, In *Brazilian Workshop on Bioinformatics*, 10 – 18.

Batista, G. E., Bazzan, A. L., Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data, *ACM SIGKDD explorations newsletter*, 6(1), 20-29.

Buda, M., Maki, A., Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks, *Neural Networks*, 106, 249-259. DOI: https://doi.org/10.1016/j.neunet.2018.07.011

Chatterjee, S., Ali S. H. (2012). *Regression analysis by example, 5th Edition*, John Wiley & Sons.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. (2002). SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, 16, 321-357.

Choi, H. (2019). Classification of bloodstream infection microbiome data using group information in taxonomy, *Journal of the Korean Data Analysis Society*, 21(2), 651-659. (in Korean).

Cover, T., P. Hart. (1967). Nearest neighbor pattern recognition, *IEEE Transaction on Information Theory*, 13, 21-27.

Davis, J., Goadrich, M. (2006). *The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on machine learning*, ICML '06. ACM, New York, NY, USA. 233-40. DOI: https://doi.org/10.1145/1143844.1143874

Drummond, C., Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling?, *Proceedings of the ICML*, 3.

Goodfellow, L., Y. Bengio, A. Courville. (2016). *Deep Learning (Adaptive Computation and Machine Learning series)*, Illustrated Edition.

He, H., Garcia, E. A. (2009). Learning from imbalanced data, *IEEE Trans Knowl Data Eng.* 21(9), 1263-1284. DOI: https://doi.org/10.1109/TKDE.2008.239

Japkowicz, N. (2000). The class imbalance problem: Signifcance and strategies, In *Proceedings of the 2000 International Conference on Artifcial Intelligence(ICAI).* 111 - 7.

Jeong, H., Kang, C., Kim, K. (2008). The effect of oversampling method for imbalanced data, *Journal of the Korean Data Analysis Society*, 10, 2089-2098. (in Korean).

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions, *Prog Artif Intell.*, 5(4), 221-32. DOI: https://doi.org/10.1007/s13748-016-0094-0

Kubat M., Holte R. C., Matwin S. (1998). Machine learning for the detection of oil spills in satellite radar images, *Mach Learn*, 30(2), 195 - 215. DOI: https://doi.org/10.1023/A:1007452223027

Ling, C. X., Sheng, V. S. (2008). Cost-sensitive learning and the class imbalance problem, *Encyclopedia of Machine Learning*, 2011, 231-235.

Park, J., 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).

Provost, F., Fawcett, T. (1999). Analysis and visualization of classifer performance: comparison under imprecise class and cost distributions. In *Proceedings of the third international conference on knowledge discovery and data mining*, 43-48

Seifert, C., Khoshgoftaar, T. M., Van Hulse, J., Napolitano, A. (2007). Mining data with rare events: a case study, In *Proceedings of the 19th IEEE international conference on tools with artifcial intelligence - Vol. 02. ICTAI '07, IEEE Computer Society*, Washington, DC., 132-139. DOI: https://doi.org/10.1109/ICTAI.2007.130

Seliya, N., Khoshgoftaar, T. M., Van Hulse J. A. (2009). Study on the relationships of classifier performance metrics, In *2009 21st IEEE international conference on tools with artificial intelligence*, 59-66. DOI: https://doi.org/10.1109/ICTAI.2009.25

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P. J. (2016). Training deep neural networks on imbalanced data sets, In *2016 international joint conference on neural networks (IJCNN)*, 4368-4374. DOI: https://doi.org/10.1109/IJCNN.2016.7727770

Weiss, G. M. (2004). Mining with rarity: a unifying framework, *SIGKDD Explor Newsl.*, 6(1), 7-19. DOI: https://doi.org/10.1145/1007730.1007734

Weng, C. G., Poon, J. (2008). A new evaluation measure for imbalanced datasets, In *Proceedings of the 7th Australasian data mining conference - Vol. 87. AusDM '08. Australian Computer Society*, Inc., Darlinghurst, Australia. 27-32. DOI: http://dl.acm.org/citation.cfm?id=2449288.2449295.

Wilson, D.L. (1972). Asymptotic properties of nearest neighbor rules using edited data, *IEEE Transaction on Systems, Man, and Cybernetics*, 2(3), 431-433.

A Study on the Performance Comparison and Approach Strategy by Classification Methods of Imbalanced Data

Byung Joo Yoo[1]

Abstract

In order to perform a classification analysis on imbalanced data, we are faced with two choices. One is the selection of a model for classification analysis, and the other is the selection of a method to solve the imbalance problem. Therefore, in this paper, I dealt with the problem of sequential approach to imbalanced data, taking into account the characteristics of the data such as the size of the training sample, the number of independent variables, and the degree of imbalance. A simulation is conducted to compare the logistic regression model, support vector machine, and deep learning, which are representative models used for binary classification analysis, to compare the classification performance according to the characteristics of the data. In addition, a simulation was performed to identify the approach strategy for obtaining the optimal result with good classification performance through Tukey's multiple comparison when combined with the methods to resolve the imbalance problem. As a result of the simulation, it was confirmed that the number of acquired samples and the presence of imbalance among the characteristics of the data operate as the dominant factors. In the case of small data, the logistic regression model is the best when combine with the over-sampling method to solve the data imbalance problem. In the case of big data, it was confirmed that the deep learning is the best when combine with the weighed estimation or the under sampling method to resolve the data imbalance problem.

Keywords : imbalanced data, logistics regression model, support vector machine, deep learning, Tukey multiple comparison.

[1]Operations Analysis Branch Chief, ROK Army Ground Operations Command, PO BOX 505-1-3, Seongsan-ro 57, Cheoin-gu, Youngin-si, Gyeonggi-do 17018, Republic of Korea.
E-mail : enjoystat@gmail.com
[Received 4 January 2021; Revised 14 January 2021, 7 February 2021; Accepted 9 February 2021]