인공지능 분류 및 인식 모델 학습용 데이터의 균일성 측정 방법

허준호, 이상복

한국정보통신기술협회

jhher@tta.or.kr, jangpo@tta.or.kr

Method for Measuring the Uniformity of Data for Training Artificial Intelligence Classification and Recognition Models

Jun-Ho Her, Sang Bok Lee

Telecommunications Technology Association

요 약

데이터 불균형은 인공지능 분류 및 인식 모델의 성능 저하 원인 중 하나로 알려져 있다. 이를 극복하기 위해 표본화 기법을 활용하여 데이터의 구성을 균등하게 함으로써 성능을 향상하는 방법이 활발히 연구되고 있다. 또한, 불균일한 데이터로 학습한 모델의 성능을 높은 신뢰도로 측정할 수 있는 지표에 관한 연구도 진행 중이다. 한편, 질병이나 사건/사고, 재해 등과 관련된 데이터는 수집이 어려워 정상 상황에 비해 양이 부족할 수밖에 없다. 이런 경우, 주어진 여건에서 데이터 불균형을 최소화하는 데이터 구성 계획을 수립하게 되며, 이러한 계획은 데이터 균일성의 품질 목표로 볼 수 있다. 본 논문은 명확한 데이터 구성 계획하에 구축하는 인공지능 학습용 데이터의 균일성을 하나의 수치로 측정할 수 있는 품질 지표, 구성비 중첩률을 제안한다. 이 지표는 데이터 구축 공정 모니터링이나 완성된 데이터에 대한 제3자 품질검증에 활용될 수 있다. 구성비 중첩률의 효과를 입증하기 위해, 2023년도의 정부 지원 공모사업을 통해 구축된 인공지능 학습용 데이터 세트 19종에 대한 구성비 중첩률과 모델 성능과의 상관관계를 분석한다.

Ⅰ. 서론

데이터 불균형은 인공지능의 분류 및 인식 모델의 성능 저하를 일으키는 주요 원인 중 하나로 알려져 있다. 이 문제를 해결하기 위해 다양한 표본화(sampling) 기법을 활용하여 데이터 구성을 균등하게 조정하고 성능을 개선하려는 연구가 활발히 이루어지고 있다[1, 2]. 또한 특히, 불균등한 데이터로 학습된 모델의 성능을 신뢰도 높게 측정할 수 있는 지표에 관한 연구도 중요한 주제가 되고 있다[3]. 한편, 질병, 사건/사고, 재해 등 비정상적 상황에서 수집된 데이터는 그 발생 빈도가 낮고 수집이 어려워, 데이터 불균형 문제를 해결하기가 더욱 어렵다. 이러한 상황에서, 주어진 여건에 따라 데이터 불균형을 최소화하기 위한 구체적인 데이터 구성 계획을 수립하는 것이 필수적이며, 이 계획은 데이터의 균일성을 확보하는 품질 목표로 설정될 수 있다.

본 논문에서는 명확한 데이터 구성 목표를 바탕으로, 인공지능 분류 및 인식 모델 학습용 데이터의 균일성을 하나의 수치로 측정할 수 있는 지표를 제안한다. 이 지표는 바운딩박스 등과 같은 라벨링의 정확도를 측정하는 데 흔히 사용하는 지표인 중첩률(IoU; Intersection over Union)에 착안하여 개발했으며 구축 공정 중 데이터 균일성에 대한 모니터링이나 구축 완료된 데이터의 균일성을 제3자가 검증하는 데 활용할 수 있다. 아울러, 2023년도 인공지능 학습용 데이터 구축 사업을 통해 마련된 데이터 세트 19종에 해당 지표를 적용한 결과와 모델 성능 간의 상관관계를 피어슨 상관계수(Pearson correlation coefficient[4])로 분석하고 제안하는 지표의 효과에 대해 살펴본다.

Ⅱ. 본론

1. 균일성 측정 지표: 구성비 중첩률

제안하는 균일성 측정 지표는 '구성비 중첩률'로 명명하고, 목표 데이터 구성비와 결과 데이터 구성비를 막대그래프로 표현했을 때, 두 막대그래프 간의 중첩률로 정의하고 수식은 다음과 같다:

$$\text{구성비 중첩률}(\%) = \frac{\sum_{k=1}^{K} \text{중첩막대길이}_k}{\sum_{k=1}^{K} \text{최대막대길이}_k} \times 100$$

여기서, K는 데이터 범주의 개수이다.

한편, 목표 데이터 구성비 매우 균일할 필요는 없으며, 현실적인 수집 여건에서 소수(minority) 데이터를 최대한 설정한다.

엑스레이(x-ray) 이미지를 통해 반려동물 질병 여부를 판별하는 지도학습 기반의 인공지능 모델을 개발한다고 가정하고 구성비 중첩률을 계산하는 예를 들고자 한다. 흔히 질병 이미지는 정상 이미지에 비해 수집이 어려워 질환: 20%, 정상: 80%의 수집 구성비 목표를 수립했다고 가정하고 구축 결과 질환: 15%, 정상: 85%의 구성비로 측정되었다고 가정하면(표 1), 아래의 계산식에 따라 구성비 중첩률은 90.48%로 계산된다.

· 구성비 중첩률(%) = (15 + 80) / (20 + 85) * 100 = 90.48

표 1. 목표·결과 구성비 및 구성비 중첩률 예시

종류	목표 구성비	결과 구성비	구성비 중첩률
질환	20%	15%	90.48%
정상	80%	85%	

2. 구성비 중첩률과 모델 성능 측정 결과 및 고찰

과학기술정보통신부의 인공지능 학습용 데이터 구축 사업을 통해 2023년에 구축한 142종의 데이터 세트 중 분류 및 인식과 관련된 19종의 데이터 세트를 대상으로 구성비 중첩률과 모델 성능의 상관관계를 분석하고자 한다. 그중 12종의 데이터 세트에 관한 구성비 중첩률과 모델 성능 결과만 나타내면 표 2와 같다(특정 데이터명 대신 데이터 분야와 순번을 조합해서 표기). 모델 성능은 과업 특성상 다양한 지표(mAP, 정확도, F1-점수 등)를 사용하였기 때문에, 결괏값을 백분위(%)로 통일하여 제시한다.

3. 구성비 중첩률과 모델 성능의 상관관계

구성비 중첩률과 모델 성능의 선형 관계를 수치로 평가하기 위해 피어슨 상관계수를 사용했는데, 알려진 대로 이 계수는 두 변수 간의 선형 상관관계의 강도와 방향을 -1에서 1 사이의 값으로 표현하고 수식은 다음과 같다:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

여기서, x, y는 각 점수이고 n은 점수 쌍의 개수이다.

19종의 데이터 세트에 대한 상관관계를 산점도 도표로 나타내면 그림 1과 같으며 피어슨 상관계수 수식을 적용한 결과 상관관계는 약 0.4421로, 이는 둘 사이에 중간 정도의 양의 상관관계가 있음을 나타낸다. 이러한 결과는 제안된 균일성 측정 지표(구성비 중첩률)가 모델의 성능과 완전히 독립적이지 않고 어느 정도 관련되어 있음을 보여준다. 이는 데이터 균일성이 분류 및 인식 모델 성능에 영향을 미치는 여러 요인 중 하나임을 시사하며, 제안된 구성비 중첩률이 데이터 균일성을 효과적으로 측정하는 지표로 활용될 수 있음을 의미한다. 또한, 구성비 중첩률이 높다고 해서 모델의 성능이 반드시 향상되는 것은 아니므로, 일반적으로 구성비 중첩률의 기준치를 50% 수준으로 설정하는 것이 적절하다고 볼 수 있다. 이는 바운딩박스와 같은 라벨링의 정확도를 나타내는 IoU 문턱값으로 주로 사용되는 0.5(50%)와도 일맥상통한다.

표 2. 12종 데이터 세트의 구성비 중첩률 및 모델 성능 결과

데이터 종류	균일성 항목 (구성비 중첩률[%])	모델 성능 항목 (성능지표, 지표별 결과[%])
이미지1	의류 대분류 분포(100)	의류 이미지 검출/분류 성능(mAP, 91)
이미지2	낙상 종류 분포(98.26)	낙상 유형 분류 성능(정확도, 93.89)
헬스케어1	정상/중증도 분포(57.15)	혈관/Plaque 탐지 성능(DSC, 81.69)
헬스케어2	클래스 분포(81.23)	혈관/Plaque 탐지 성능(mIoU, 96.76)
농축수산1	식물 분포(95)	식물분류 성능(F1-점수, 97.6)
농축수산2	말 구분 분포(75.42)	말머리/전신 식별 성능(mIoU, 88.55)
안전환경1	클래스별 분포(100)	객체 탐지(mAP@0.5, 87.6)
안전환경2	토지 클래스별 분포(68.09)	토지 클래스 탐지 성능(FIoU, 78.26)
스포츠1	씨름 동작별 분포(88.73)	씨름 동작 분류 성능(Top-3 정확도, 98.97)
스포츠2	판정별 분포(99.59)	판정 성능(정확도, 90.48)
지역특화1	불량 데이터 분포(84.25)	불량 탐지 성능(mAP@0.5, 97.8)
지역특화2	조업 종류별 분포(99.72)	조업 종류 분류 성능(정확도, 96.55)

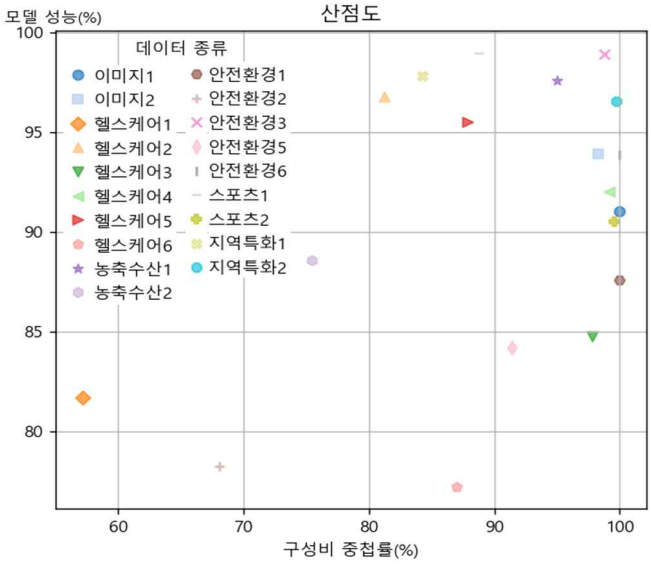

그림 1. 구성비 중첩률과 모델 성능에 대한 산점도

Ⅲ. 결론

본 논문에서는 데이터 불균형에 따른 모델의 성능 저하 문제와 관련하여, 데이터 불균형의 정도를 측정하기 위한 새로운 지표인 '구성비 중첩률'을 제안하였다. 이 지표는 명확한 데이터 구성 계획하에 구축하는 인공지능 분류 및 인식 모델 학습용 데이터의 균일성을 수치화하여 평가함으로써, 구축 공정 중 데이터 균일성에 대한 모니터링이나 구축 완료된 데이터의 균일성을 제3자가 검증하는 데 활용할 수 있다. 19종의 데이터를 대상으로 분석한 결과, 구성비 중첩률과 모델 성능 사이에는 중간 정도의 양의 상관관계가 존재하는 것으로 나타났으며, 이는 제안하는 지표의 효용성을 방증하는 결과다.

그러나 구성비 중첩률과 모델 성능의 관계를 더 깊이 있게 분석하기 위해서는 더 많은 데이터와 다양한 유형의 인공지능 모델을 통한 검증이 필요하다. 따라서 향후 연구에서는 다양한 분야에서 구성비 중첩률의 적용 가능성을 탐색하고, 더욱 정교화할 방안을 모색할 계획이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 초거대AI 확산 생태계 조성 사업(2100-2131-305, 2024년도 초거대AI 확산 생태계 조성 사업)에 의해서 수행되었습니다.

참 고 문 헌

[1] Ye, Han-Jia et al. "Procrustean Training for Imbalanced Deep Learning." 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021): 92-102.

[2] Park, Seulki et al. "The Majority Can Help the Minority: Context-rich Minority Oversampling for Long-tailed Classification." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021): 6877-6886.

[3] Paula Branco, Luís Torgo, and Rita P. Ribeiro. 2016. A Survey of Predictive Modeling on Imbalanced Domains. ACM Comput. Surv. 49, 2, Article 31 (June 2017), 50 pages. https://doi.org/10.1145/2907070

[4] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient