데이터 라벨링을 통한 신뢰성 확보

이형주 빅데이터 프로젝트그룹(PG1004) 위원, 주식회사 크라우드웍스 이사

1. 머리말

AI가 산업 전반에 영향을 주기 시작하면서, 대부분의 관심은 “어떻게 모델 구조를 설계하고 어떤 알고리즘을 적용해야 좋은 AI 모델을 만들 수 있을까?”라는 질문에 대한 답을 찾는 데 쏠렸다. 이에 따라, 데이터 사이언티스트나 데이터 모델 엔지니어 역할을 할 수 있는 인재를 뽑기 위해 회사 간 치열한 경쟁이 벌어졌으며, 경쟁의 열기는 아직도 식지 않은 모양새다.

이에 반해, AI 모델 개발 업무의 80%를 차지하는 데이터 수집/가공/정제에 대한 관심과 투자는 상대적으로 적었고, AI 모델 개발의 주요 업무로 다루어지기보다는 보조적이고 선택적인 과정으로 다루어지는 경향이 있었다.

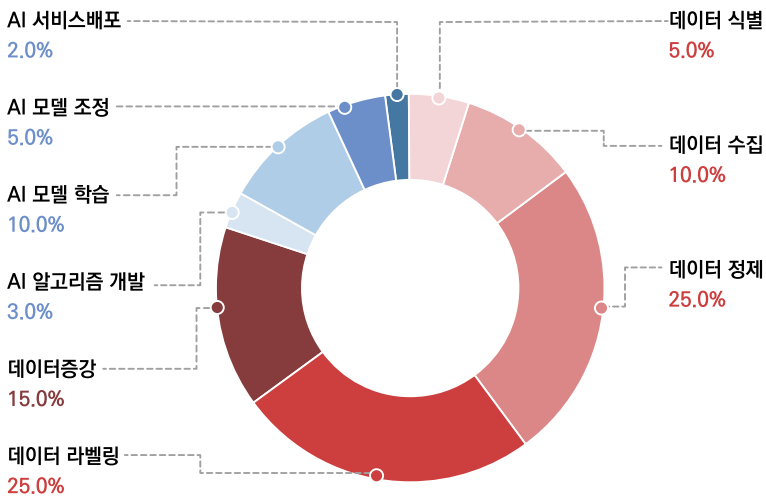
하지만, 최근 “Data-centric AI”라는 키워드가 대두되면서 이와 같은 인식이 변하기 시작했다. 특히 AI 관련해 학계나 산업계에 큰 영향력을 행사하고 있는 앤드류 응(Andrew Ng)이 화두를 던지면서 데이터, 특히 좋은 데이터(Quality Data) 확보의 중요성에 대한 공감대가 형성되었다.

‘좋은 데이터’의 조건은 일반적으로 ‘신뢰성’이라는 용어로 표현된다. 데이터의 신뢰성이라는 용어는 데이터의 형태, 산업의 특성, 데이터를 다루는 과정 등 여러 입장에서 다른 해석을 할 수 있어 하나의 구체적인 정의로 표현하기는 힘들다. 하지만, 최근 연구 기관을 중심으로 데이터의 품질 및 신뢰도에 대한 일반적 기준을 만드는 시도가 계속 되고 있다.[2]

우선 이 글에서 다룰 데이터 신뢰성의 정의와 이를 판별할 수 있는 요소들을 분류해 보겠다. 이후 실제 데이터 라벨링 과정에서 신뢰성의 요소들을 향상할 수 있는 방안에 대해서 논하도록 한다.

2. 데이터의 신뢰성

흔히 이야기하는 ‘데이터의 신뢰도’는 사용 목적 및 집합 단위에 따라 다르게 정의될 수 있는데, 이는 데이터 활용 목적이 다양하고, 데이터가 수집되고 가공되는 각 과정에서 필요하고 측



[그림 1] AI 모델 개발에 필요한 업무 및 구성 [1]

<표 1> 가공 여부에 따른 데이터 신뢰성 정의

	원천 데이터의 신뢰성	가공 데이터의 신뢰성
정의	데이터의 성공적인 가공을 위해 원천 데이터가 기본적인 조건을 얼마나 갖추었는지에 대한 정도	모델 학습을 위해 가공 데이터가 얼마나 적합한 구성과 일관성의 정도
측정 기준의 예시	기술적 품질, 데이터 세트의 크기, 각 각 데이터의 포맷 및 속성의 기준 부합	가공의 일관성, 클래스 간 크기의 유사성, 기준 일치도

<표 2> 개별데이터의 신뢰성 요건 구분

신뢰성 요건 구분	내용	신뢰성 평가 기준
객체 요건	가공 할 대상 객체의 정의와 상세 조건 (법적으로 소형차로 구분되는 자동차)	대상 객체가 맞는가?
컨텍스트 요건	대상 객체의 상황에 대한 조건 (차선을 위반한)	대상 객체가 해당되는 상황에 놓여 있는가?
가공 방식 요건	상세한 가공 방식 및 도구에 대한 사용 방법 (예 : 폴리곤)	지정된 가공 방식으로 작업되었으며, 그 결과가 주어진 조건을 만족시키는가?

정 가능한 신뢰성 요소가 다르기 때문이다.

2.1 목적에 따른 데이터 신뢰성의 정의

데이터는 다양한 목적을 가지지만 일단 이 글에서의 목적은 'AI 모델 구축을 위해 필요한 학습의 목적'으로 한정하도록 한다. 이러한 목적을 배경으로 '모델 학습 데이터 신뢰성'의 정의를 내리면 다음과 같이 정의할 수 있다.

AI 모델 학습과 성능 개선 목적에 부합하는 정도

이렇게 데이터 신뢰성이 정의되는 경우, 신뢰성의 평가는 모델의 성능과 거의 일치한다고 볼 수 있다. 이러한 데이터 신뢰성의 정의는 데이터의 목적과 그 목적의 달성 여부에 따라 정확하게 정량적인 평가가 내려질 수 있고, 결과 자체로 이러한 신뢰성 평가에 이견을 달 수는 없을 것이다.

하지만, 이렇게 정의를 하는 경우에는 다음과 같은 문제가 있다.

어디까지나 결과적인 신뢰도 평가밖에 되지 못하며, 사전에 파악하기가 어렵다.

보통 AI 모델이 설명 불가능한 (unexplainable model) 구조를 갖고 있다는 걸 고려했을 때, 신뢰성의 높고 낮음의 이유에 대한 설명이 불가능한 경우가 많다.

2.2 가공 여부에 따른 데이터 신뢰성의 정의

데이터라는 표현은 보통 상황에 따라 의미가 달라진다. 즉, 데이터라는 단어가 상황에 따라 가공 전 원천 데이터를 지칭하는 경우도 있고, 가공 후 가공 결과 데이터를 의미하기도 한다.

원천 데이터의 신뢰성은 정해진 기술 및 구성 측면의 요구 조건에 얼마나 부합하는지를 평가하는 측면이 강하다. 따라서 데이터 세트를 준비하는 과정에서 해당 신뢰도를 높이기 위한 활동을 실행할 수 있다. 원천 데이터의 신뢰도가 높다는 의미는 데이터 가공 시에 원천 데이터로 인해 문제가 발생할 소지가 매우 낮다는 의미이다. 하지만 원천 데이터의 신뢰성을 확보했다고 해서 가공 데이터의 신뢰성이 확보되거나 모델 학습에 완벽히 부합하는 데이터를 확보했다고 말하

기는 어렵다.

오히려 모델 학습에 부합하는지에 대한 부분은 가공 데이터의 신뢰성으로 일부 측정할 수 있다. 하지만 가공 데이터가 아무리 신뢰도가 높다고 해도 모델 학습에 최적인지는 알 수 없다. 클래스의 사이즈 같은 항목을 예로 들면, 각각의 클래스를 정확하게 학습하기 위해서 필요한 데이터의 수를 미리 알 수 없기 때문이다. 가공 데이터의 신뢰성은 가공 과정에서도 일부 확보할 수 있으나, 일부는 가공 후 파악되는 항목도 있다.

2.3 개별 데이터 신뢰성의 정의

2.2에서 가공 데이터 신뢰성은 데이터 세트 기준에서의 신뢰성이다. 하지만 개별 가공 데이터의 신뢰성도 정의 내릴 수 있다.

데이터의 가공 내용이 가공 요건에 부합하는 정도

개별 데이터의 가공 요건은 크게 객체 요건과 컨텍스트 요건, 가공 방식 요건 세 가지로 구분할 수 있으며, 개별 가공 데이터의 신뢰성은 이 세 가지 측면에서 평가될 수 있다.

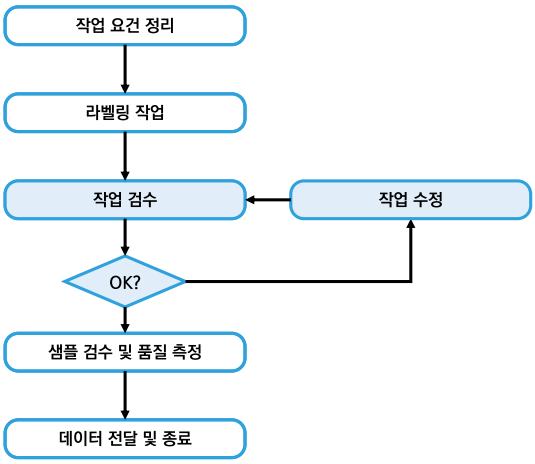
개별 데이터의 신뢰성은 사전 준비와 관리를 통해 데이터 라벨링 과정에서 측정 및 관리가 가능하다. 따라서, 이하 3장부터는 개별 데이터의 신뢰성 확보를 중심으로 필요한 사항에 대해 논의하도록 한다.

2.4 데이터 신뢰성 확보의 순서

위에서 언급한 데이터 신뢰성들은 일반적인 작업 순서에 따라 다음과 같은 순서로 측정되고 확보할 수 있다.

다만 가장 마지막 단계인 모델 학습 데이터 신뢰성은 앞서 언급한 대로 모델의 속성이나 과업의 특징에 따라 결과적으로 확보되는 지표의 성격이 강해 확보라는 표현 대신 측정이라는 표현을 사용하였다.

[그림 2] 데이터 신뢰성 확보 순서



작업	참여자
작업 요건 정리	데이터 의뢰자, 데이터 서비스 제공자
라벨링 작업	라벨링 작업자
작업 검수	라벨링 감수자, 데이터 서비스 제공자
작업 수정	라벨링 작업자
샘플 검수 및 품질 측정	데이터 서비스 제공자

[그림 3] 데이터 라벨링의 과정과 과정 별 참여자[3]

<표 3> 용어 정의의 세 가지 요소

구분	설명
용어의 일관성	해당 프로젝트 내에서 사용되는 언어는 항상 동일한 의미를 가져야 한다는 의미
용어의 일반성	가능한 일반적으로 사용되는 용어를 활용해야 한다는 의미
용어의 명확성	불명확하거나 이중적인 의미를 가지지 않고 그 뜻이 명확해야 한다는 의미

3. 라벨링 과정에서의 데이터 신뢰성 확보 방안

일반적으로 라벨링은 [그림 3]과 같은 순서로 진행된다. 그런데 위의 작업은 여러 참여자가 함께 진행해야 하는 과정으로 이 부분에 대한 이해가 신뢰성 확보 방안을 고려할 때 반드시 감안되어야 한다.

즉 각 작업 단계의 주체가 다르고, 각자 가지고 있는 배경 지식의 이해도나 범위가 다르기 때문에 소통 오류를 최소화하고 주관적 판단을 최대한 배제할 수 있는 방안을 확보하는 것이 신뢰성 확보에 가장 핵심이라 볼 수 있다.

3.1 정확한 용어의 정의

라벨링 과정에서 신뢰성을 확보하기 위해 가장 기본적인 작업이 용어의 정의이다. 용어의 정의는 일관성, 일반성, 명확성 세 측면에서 고려되어야 한다. 용어의 정의가 중요한 이유는 앞서 설명한 라벨링에 다양한 주체가 참여하기 때문이다. 적절하게 정의된 용어들은 배경 지식과 경험이 다른 각각의 작업 주체들이 서로 명확하게 커뮤니케이션을 하는 데 도움이 된다. 실제 최초 데이터 의뢰자와 서비스 제공자의 협의는 특정 비즈니스 용어와 기술용어가 함께 사용되지만, 실제 작업자들은 용어를 모르거나 혹은 동일한 용어를 다양한 범위와 깊이로 이해하는 경우가 많다. 따라서 적절한 용어 정의가 없으면 이후 커뮤니케이션에 혼란이 발생할 수 있다.

3.2 요건 정의 방식

요건 정의를 하는 방식은 Black list 방식과 White list 방식으로 나눌 수 있다.

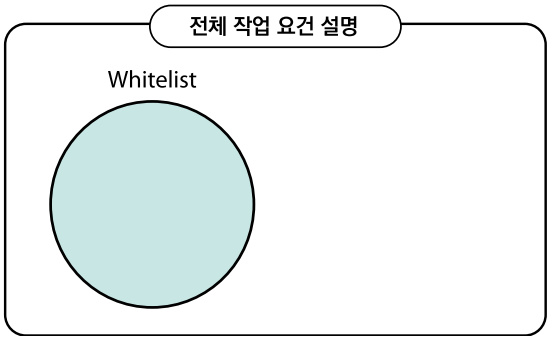
3.2.1 Black list 방식

Black list 방식은 작업 요건을 정의할 때, 제외해야 하거나 하지 말아야 할 내용들을 정리하는 방식이다. 즉, 아래 그림처럼 Blacklist 형태로 언급된 내용을 제외한 내용이라면 모두 인정한다는 의미이다. 예를 들어 자동차 바운딩 작업을 진행하는 프로젝트에서 '검은색 자동차는 제외'라는 식으로 대상, 컨텍스트 가운데 제외할 대상들만 정의하는 방식이다.

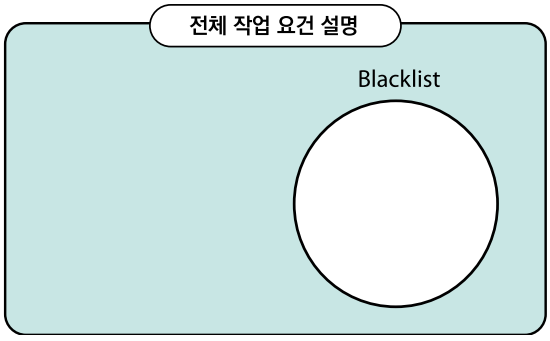
이 방식은 일부의 예외 사항만 제외하면 작업 요건이 성립되는 경우 사용한다. 다만, 작업 과정 중 예외 사항이 추가되는 경우에는 이전 작업 데이터의 신뢰도에 문제가 생길 수 있어 이 방식을 활용하기 어렵고, 예외 사항이 많아 작업자가 모두 숙지하기 어려운 경우에도 문제가 생길 수 있다.

3.2.2 White list 방식

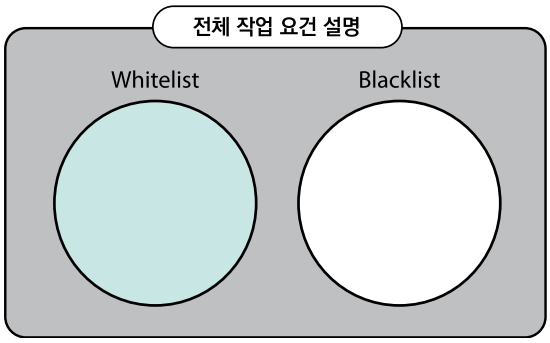
White list 방식은 작업 요건을 반드시 포함하거나 실행해야 할 항목들로 구성하고, 나머지는 인정하지 않는 방식이다. 앞서 언급한 예시에서



[그림 4] Blacklist 방식에서의 작업요건 범위



[그림 5] Whitelist 방식에서의 작업요건 범위



[그림 6] 혼합 사용 방식에서의 Gray area

'흰색 자동차만 포함' 같은 형태로 작업 대상을 규정하고, 여기에 해당하지 않는 대상은 작업 대상에서 빠지는 방식이다.

White list 방식으로 작업을 하는 경우 추가적인 요건이 발생하게 되면, 이전 작업이나 데이터 중 요건을 충족했는데 제외되는 경우가 생기는 등 일부 내용의 유실이 발생한다.

3.2.3 Black list/White list 방식 혼합 사용의 위험성

작업 요건을 쉽게 정의하기 위해서 두 가지 방식을 혼합하여 사용하는 경우가 있다. 해야 할 작업 요건을 정의하는 동시에 예외 사항이나 금지 사항을 동시에 작성하는 형태이다.

이는 많은 문제의 원인이 된다. 이러한 요건 정의는 Gray area를 만들어낼 수 있어, 라벨러들의 혼란이나 작위적인 판단을 불러일으킬 수 있다. 예를 들어 Whitelist로 '흰색 자동차' Blacklist로 '검은색 자동차'를 지정하면 나머지 색깔의 자동차에 대한 판단이 모호해진다.

따라서, 작업 요건의 혼란을 주지 않기 위해서는 Blacklist 방식과 Whitelist 방식 중 하나를 택해야 하며, 이러 선택이 불가한 경우에는 이후 설명할 작업 분리와 같이 작업의 형태를 바꿔주는 등의 조치가 필요하다.

3.3 상세 요건의 정량적 표현

요건을 정의할 때 부사 및 형용사의 사용은 최대한 피해야 한다.

예를 들어 객체 바운딩 작업 요건으로 '최대한 타이트하게'라는 식의 표현을 사용하면, 해당 작업에 참여하는 모든 구성원들이 이 문구를 각각 해석하고 판단하게 된다. 그리고, 작업 자체에 문제가 없다고 하더라도 데이터 세트의 일관성은 보장하기 힘들어진다.

따라서 가능한 정량적 표현을 써서 요건을 정의해야 한다. 예를 들어 '최대한 타이트하게'라는 표현 대신 '마진 폭이 1픽셀 이내로' 등의 표현이 될 것이다. 다만, 정량적인 표현을 사용하기 위해서는 작업 도구에서 정량적인 요건들을 측정하거나 판단할 수 있는 기능을 반드시 제공해 주어야 하며, 그렇지 않은 경우에는 작업과 검수 과정에서 부담을 발생시킬 수 있다.

3.4 작업 방식의 설계

작업의 내용이 복잡해지면, 작업의 요건이 길어지고 한 번에 해야 할 작업의 종류나 양이 많아진다. 이러한 작업 설계는 오류 발생처럼 신뢰성 유지에 어려움을 초래한다.

이러한 경우 작업 방식의 변경을 통해 작업자가 해야 할 일을 바꾸거나, 작업을 분해해야 한다.

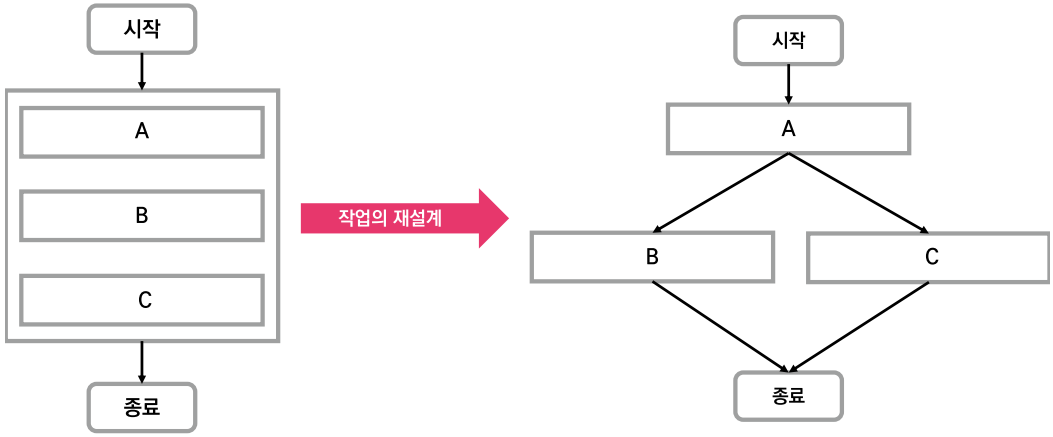
3.4.1 작업자 역할의 변경 - AI 모델의 활용

AI 모델을 활용하여 작업 요건을 변경시킬 수 있는데, 특히 작업 자체의 난이도가 높거나 시간이 많이 걸리는 경우 고려할 수 있다.

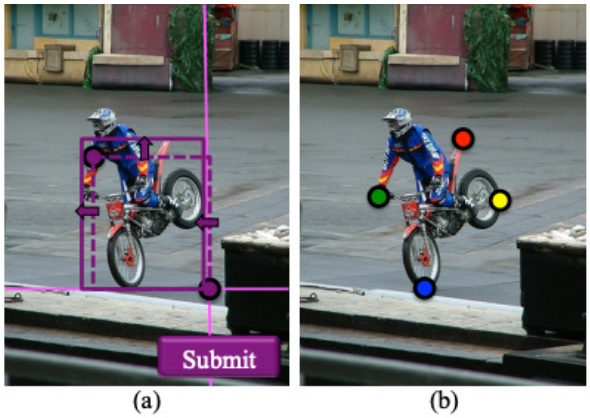
예를 들어 작업과 검수로 이루어진 작업에서 AI 모델이 작업의 전부 혹은 일부를 진행하게 되면, 작업, 수정/보완, 검수의 과정으로 바뀌게 되며, 작업자의 역할이 직접적인 작업에서 수정/보완으로 바뀌게 된다.

3.4.2 작업의 재설계

작업을 분리하고 다시 재조합하는 형태로 작업을 재설계하면, 복잡성에서 기인하는 여러 가지 문제들을 해결할 수 있다. 아래 그림처럼 A, B, C라는 복수의 작업 요건이 필요하고, 이를 동시에 진행하기에는 문제가 있다고 판단되는 경



[그림 7] 작업 재설계의 예시



[그림 8] Motorbike 바운딩 박스를 그리는 일반적인 방법(좌)과 extreme clicking 방식의 예

우 우측과 같이 각 작업을 직렬 혹은 병렬로 이어 붙이는 형태로 작업을 재설계할 수 있다.

이러한 방식은 작업자들이 한정된 작업에 집중하게 하여 효율을 높일 수 있고 병렬 작업이 가능한 경우에는 시간을 줄일 수 있는 장점이 있지만, 작업이 너무 많이 분리되는 경우 각 작업 간 데이터 전달 등의 과정에서 과도한 비용이 발생할 수 있다는 점은 유의해야 한다.

3.5 적합한 작업 도구의 제공

가공 방식 요건은 많은 경우 작업 도구의 기능과 밀접한 관련이 있다. Click 후 Drag 하는 방식으로 작업되는 바운딩 박스 작업 과정은 일견 더 이상의 개선이 어려운 것으로 보인다. 하지만, 일련의 연구에 따르면, 몇 개의 점을 찍는 것으로 해당 작업을 진행할 수 있으며, 이 경우 품질과 속도에서 개선이 있다는 연구 결과도 있다. [4]

따라서 적합한 작업 도구에 대한 지속적인 연구 개발이 필요하며, 이러한 개선은 작업자의 실수를 줄이고 효율을 높여 데이터의 퀄리티를 향상시키는 데 결정적인 역할을 할 수 있다.

4. 맺음말

데이터의 신뢰성을 높이기 위해 라벨링 작업에서 고려해야 하는 사항들에 대해서 알아보았다. 데이터 라벨링은 AI 기술과 비즈니스 양쪽을 이해하고 있어야 제대로 된 요건 설계가 가능하다. 최적화된 요건을 찾는 과정 또한 기술 및 인지적 요소의 고려와 지속적 테스트가 병행되어야 하는 상당히 어려운 작업이다. 하지만 이러한 요소들이 데이터의 신뢰성을 높여주며, 데이터 라벨링의 효과적 진행에 필수적이라는 것도 엄연한 사실이다.

따라서 데이터 라벨링의 과정에 대해 다양한 학문적 방법으로 접근하여 고도화를 추진하는 한편 일련의 표준화 과정을 통해 산업계 전반에 걸친 know-how의 공유가 필요하다고 생각한다.

참고문헌

[1] 한국지능정보사회진흥원(2020)

[2] 인공지능 학습용 데이터 품질관리 가이드라인 v2.0 (https://aihub.or.kr/node/71634)

[3] 정보통신 단체 표준 TTAK.KO-10.1336-Part1

[4] Extreme clicking for efficient object annotation, Dim P. Papadopoulos , https://homepages.inf.ed.ac.uk/keller/publications/iccv17.pdf