Journal of Korea Multimedia Society Vol. 24, No. 10, October 2021(pp. 1403-1413)
https://doi.org/10.9717/kmms.2021.24.10.1403

인공지능 데이터 품질검증 기술 및 오픈소스 프레임워크 분석 연구

윤창희†, 신호경††, 추승연†††, 김재일††††

An Evaluation Study on Artificial Intelligence Data Validation Methods and Open-source Frameworks

Changhee Yun†, Hokyung Shin††, Seung-Yeon Choo†††, Jaeil Kim††††

ABSTRACT

In this paper, we investigate automated data validation techniques for artificial intelligence training, and also disclose open-source frameworks, such as Google's TensorFlow Data Validation (TFDV), that support automated data validation in the AI model development process. We also introduce an experimental study using public data sets to demonstrate the effectiveness of the open-source data validation framework. In particular, we presents experimental results of the data validation functions for schema testing and discuss the limitations of the current open-source frameworks for semantic data. Last, we introduce the latest studies for the semantic data validation using machine learning techniques.

Key words: Artificial Intelligence, Data Validation, Data Quality Management, Open-source Framework, Review Study

1. 서 론

최근 빅데이터 및 인공지능 기술 발전에 의한 산업구조의 변화와 IT 기반 서비스 확산 및 고도화가 빠르게 이루어지고 있으며, 인공지능 모델 개발(Development)뿐 아니라 확산(Deployment) 과정에서 요구되는 소프트웨어 공학, 데이터 관리, 시스템 최적화, 인공지능 모델의 유지·관리 등에 대한 관심이 높아지고 있다. 특히, 인공지능 기술은 대규모 데이터로부터 주어진 작업을 달성하기 위한 다양한 패턴을 기계가 스스로 학습하는 데이터 주도적 접근을 따르기 때문에, 인공지능 모델의 실제 환경 적용을 위해서는 인공지능 학습용 데이터의 품질 관리와 지속적 규모 확장, 학습 데이터 구축 프로세스 효율화 등이 중요하다[1-2].

컴퓨터비전, 자연어처리, 음성인식 등의 인공지능, 특히 딥러닝 분야 기술들은 데이터 양에 비례하여 성능이 향상되는 것으로 알려져 있으며, 학습 가능한 형태로 어노테이션 및 레이블링이 포함된 양질 학습 데이터를 수집하고, 데이터의 품질관리를 위한 노력

※ Corresponding Author : Jaeil Kim, Address: (41566) 80 Daehak-ro, Buk-gu, Daegu, Korea, TEL : +82-53-950-7283, FAX : +82-53-950-7283, E-mail : jaeilkim@knu.ac.kr
Receipt date : Sep. 23, 2021, Revision date : Oct. 14, 2021
Approval date : Oct. 19, 2021
† AI Future Strategy Center, National Information-Society Agency (E-mail : yunch@nia.or.kr)
†† School of Computer Science and Engineering, Kyungpook National University (E-mail : parkland106e@naver.com)
††† School of Architecture, Kyungpook National University (E-mail : choo@knu.ac.kr)
†††† School of Computer Science and Engineering, Kyungpook National University
※ This work is supported by the Korea Agency for Infrastructure Technology Advancement(KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 21AATD-C163269-01).

이 계속되고 있다. A. Paleyes 등은 데이터 관리, 모델학습·검증·배포 단계에서 발생하는 문제가 전체 파이프라인에 영향을 미치는 요소들에 대한 분석을 통해 머신러닝을 배포할 때에 필요한 고려사항을 도출하였고[3], T. Rukat 등의 논문에 따르면, 데이터 품질 모니터링을 위한 데이터 단위 테스트, 임계값 정의 등을 위해서는 많은 양의 도메인 지식이 필요하고, 입력 데이터 검사를 통한 품질측정, 문제 해결, 모델 예측 성능에 대한 품질 문제 정량화를 강조한다[4]. S.E. Whang 등은 모델학습에서 머신러닝 알고리즘의 핵심은 바이어스 없는 양질의 데이터로서, 데이터 수집, 데이터 검증 및 클랜징, 모델학습 등 단계를 체계적으로 구분하고, 단계별 데이터 검증 방법을 소개한다[5]. 이와 같이 기존 데이터 품질 관리 기법들에 대한 분석을 통해 일반화된 형태의 품질관리 프로세스를 정립하려는 분석 연구가 활발히 이루어지고 있으나, 데이터 품질관리 프로세스를 구현한 오픈소스 프레임워크에 대한 분석 연구는 찾기 어렵다.

본 논문은 인공지능 학습용 데이터에 대한 자동화된 품질관리 기술을 분석하고, Google의 TensorFlow Data Validation(TFDV)와 같이 인공지능 모델 개발 과정에서 데이터 유효성 검증을 지원하는 오픈소스 프레임워크에 대한 공개 데이터를 이용한 분석 연구를 소개한다. 특히, 공개 데이터를 이용하여 현재 오픈소스 프레임워크에서 구현된 데이터 유효성 검증 기능을 검토하고, 의미론적 데이터(Semantic Data)에 대한 오픈소스 프레임워크들의 한계점과 이를 극복하기 위한 최신 연구들을 분석 소개한다. 본 논문의 2장은 데이터 관리 플랫폼, 품질검증 기법과 관련 연구를 검토하고, 공통된 인공지능 데이터를 위한 품질관리 프로세스를 설명한다. 3장에서는 오픈소스 기반 데이터 품질관리 프레임워크들의 장단점을 분석하고, 공개 영상데이터와 메타데이터를 이용한 사례 분석을 통해 현재 오픈소스 데이터 관리 프레임워크의 한계점을 도출한다. 4장에서는 이에 대응하는 최신 연구들을 소개하며, 5장에서 결론을 맺는다. 이를 통해 대규모 학습 데이터 관리 및 모델학습·검증·구현 등을 중심으로 오류를 최소화하여 양질의 데이터를 효율적으로 도출할 수 있는 방안을 모색하였다.

2. 인공지능 데이터 품질검증

2.1 인공지능 데이터 품질검증 필요성

인공지능 분야에서 양질의 데이터가 우수한 성능을 보장할 수 있기 때문에, 엄격한 품질검증을 통한 학습 데이터 품질 확보가 매우 중요하다. 특히, 모델 설계, 배포 이후 환경 변화, 데이터 분포 변화 등에 의해 성능이 악화되는 인공지능 기술의 특성에 대응하기 위해 인공지능 모델을 재학습시키고 성능을 유지하기 위하여 학습 데이터를 지속적으로 수집하는데, 이 과정에서 데이터 품질 유지를 위한 자동화된 프레임워크의 개발이 필수적으로 요구되고 있다.

구체적으로 A. Saria와 A. Subbaswamy는 기계학습 모델 오류 및 성능을 저하시키는 요인을 크게 데이터 세트 오류, 데이터 편향성, 모델 가정의 오류, 부정확한 보고라고 구분한다. 데이터 세트 오류는 데이터 세트 내의 특정 클래스 또는 하위 집단에 오류가 생기면 분류기 성능의 정확도가 높더라도 전체 성능을 저하시킬 수 있고, 데이터 편향성은 데이터 훈련환경이나 수집 방식, 수집 데이터의 분포에 따라 발생할 수 있으며, 모델의 일반화 성능 저하의 주요 요인이 된다고 설명한다[6]. 예를 들어, 특정 병원에서의 흉부 X-레이 데이터로 모델을 훈련하였을 때는 성능이 좋았으나 다른 병원에서는 성능이 저하된 경우가 발생할 수 있는데, 이는 데이터 수집 환경 변화, 데이터 수집 방법 차이 등으로 데이터 분포 차이가 발생하기 때문이다.

모델 가정의 오류는 모델의 잘못된 사양으로 발생하는 경우가 많은데, 선형모델과 같이 피처 간에 많은 상호작용이 발생하는 복잡한 설정에는 부적절하게 적용될 수 있으며, 종속 데이터의 경우에도 샘플에 독립적인 오류가 있다고 가정하지만, 결과값이 연결된 지리 데이터나 소셜 네트워크 데이터는 모델 가정이 적용되지 않는 경우가 많기 때문이다. 부정확한 보고는 잘못된 보고 실행으로 인해 원래 의도한 모델의 목적과 실제 사용된 방법이 불일치하는 경우를 의미한다. 특히, 데이터뿐 아니라 메타데이터에 대한 스키마가 부정확하게 정의되고, 데이터 수집 과정 중 지켜지지 않는다면 인공지능 모델의 적절한 동작을 보장할 수 없다. 이에 대응하여 인공지능 모델에서 처리 가능한 데이터임을 확인할 수 있는 스키마 기반 데이터 검증이 필수적으로 활용되고 있다.

인공지능 모델의 활용에서 어려운 점 중 하나는 신규 학습데이터를 이용한 재학습 과정에서 발생하

는 문제로써 인공지능 모델이 지속적으로 수집되는 새로운 데이터에 과적합되면서 과거 학습지식을 잊어버리는 망각현상이다. 기존 데이터와 다른 패턴, 분포를 따르는 신규 데이터로 인해 인공지능 모델 성능이 하락될 수 있고, 특히 의미론적 데이터에서는 데이터 분포 분석과 이상 데이터 검출이 어렵기 때문에 재학습 과정의 구현이 어렵다. 이를 보완하기 위한 다중 데이터 세트를 이용하는 전이학습(Transfer Learning) 기법과 망각현상을 보완하기 위한 점진학습(Incremental Learning) 기법, 데이터 레이블링 및 모델 업데이트 과정에서 학습된 패턴의 일관성을 유지하고자 하는 반지도 학습(Semi-supervised Learning)등이 활용되고 있다. 하지만 지속적으로 증가하는 학습 데이터에 대한 의미론적, 통계적, 스키마적 품질검증 없이는 근본적 해결이 어려우며, 이를 효율적으로 해결하기 위한 품질관리 자동화 도구에 대한 연구가 필요하다.

2.2 인공지능 데이터 플랫폼: Data Lakes

대규모 데이터를 효율적으로 관리하고 활용할 수 있는 방법론으로써 Data lakes 및 Data warehouse 개념이 연구되어 왔으며, 특히 Data warehouse가 새로운 아키텍처 패턴인 Data lake로 대체될 것이라 전망하고 있다[7]. Data warehouse는 기업의 다양한 시스템에서 생성되는 데이터를 저장·분석하기 위한 비즈니스 인텔리전스(BI)를 구현할 목적으로 개발되어 성장하였지만, 인공지능 모델 개발을 위한 학습 데이터 종류가 SNS, Sensor Data, Image, Video 등으로 다양해지고, 그 규모 또한 방대해지고 있기 때문에 기존 Data warehouse 구성만으로는 해결이 어려운 문제가 많아졌기 때문이다[8].

Data lake는 Data storage에 저장된 데이터를 전통적인 Data warehouse의 ETL/ELT(Extract, Transform, Load)와 같은 데이터 처리 방식과 더불어 정형·비정형·반정형 데이터를 실시간으로 수집·정제·관리할 수 있는 데이터 가상화(Data virtualization)와 데이터 검증(Data Validation) 기능으로 확장된 플랫폼 아키텍처로 볼 수 있다. Fig. 1은 Data Lake 플랫폼에서 데이터 수집부터 분석, 기계학습 모델 활용까지 이르는 데이터 흐름을 도식화한 것이다. Andrei Paleyes는 기계학습을 적용할 때 발생할 수 있는 데이터 관리, 모델학습·검증·구현, 윤리, 보안

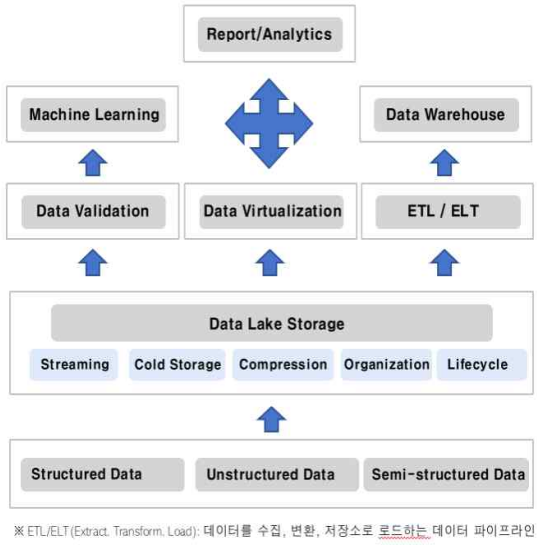


Fig. 1. Data lake as data platform architecture.

등의 이슈를 중심으로 문제점을 추출하고 해결방안에 대해 체계적으로 공유하는 것이 중요하다고 강조하고 있으며, Data Lake 플랫폼 안에 데이터 관리, 검증 프로세스를 구체화하여 제안하였다[3].

2.3 인공지능 데이터 품질검증 방법론

Data Lake 플랫폼을 통한 대량의 데이터 수집 및 관리, 인공지능 모델의 개발 방법론[3]이 개발되는 동시에 인공지능 데이터 품질 관리 기술에 대한 중요성도 함께 강조되고 있다. 특히, 데이터가 광범위하게 생성되고 데이터 패턴 형태도 다양해지는 추세에 반해 이를 관리하는 데이터 플랫폼이나 기관, 기업에서 구체적인 데이터 품질검증에 대한 이해가 부족할 경우, 모델 배포 과정에서 사용 방법에 대한 시행착오와 지속적인 데이터 수집 과정에서 노이즈 데이터가 포함되어 모델 성능 하락을 야기하는 등 여러 가지 부작용이 발생할 수 있다[9].

S.E. Whang은 심층 학습 응용 프로그램에서 자주 발생하는 데이터 수집 및 품질 문제를 해결하기 위해 최첨단 데이터 수집 기술과 데이터 유효성 검사 기술을 소개한다[5]. 특히, 인공지능 모델학습 과정에서 학습 데이터의 단순한 오류라도 모델의 최종 학습 결과에 영향을 미칠 수 있기 때문에, 사소한 데이터 오류가 인공지능 학습 과정에 전파되어 더 많은 파이프라인 상태를 오염시키지 못하도록, 학습 이전 혹은 학습 과정 초기에 데이터 오류를 포착하는 것이 중요

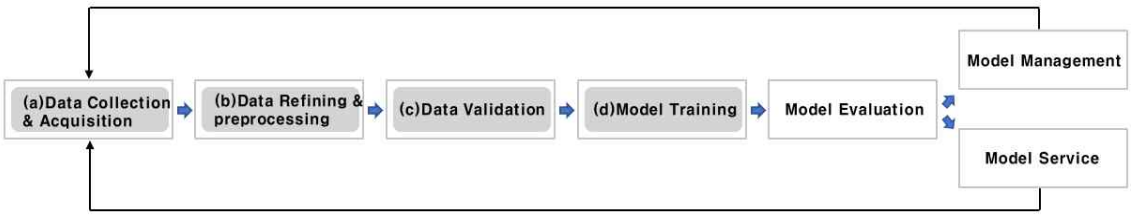


Fig. 2. End-to-end process of AI development.

하다고 강조하고 있다[11]. Fig. 2는 데이터 수집부터 마지막 인공지능 서비스까지 데이터 흐름을 도식화한 것이다. 데이터 수집·획득(a) 및 데이터 전처리를 통한 정제작업(b) 이후 데이터 오류를 발견하기 위한 품질검증 과정(c)을 거쳐 실제 인공지능 모델을 학습하는 네 번째 단계(d)가 데이터 검증을 통해 오류를 발견하고 보완할 수 있는 단계이다.

데이터 오류 검출 및 개선은 데이터 수집 과정에서부터 시작할 수 있으며, Y. Roh 등은 데이터 수집 과정을 데이터 수집에서 모델 제공에 이르는 종단간 프로세스로 체계화하여, 프로세스 단계별 데이터 검증 및 품질 향상 방법들을 분석하였다[10]. 데이터 수집은 센싱 모달리티로부터의 데이터 획득(a), 레이블링을 통한 학습용 데이터 구축(b), 인공지능 모델 학습 과정에서의 데이터 처리(c)의 3단계로 구분한다. 먼저, 데이터 획득(a)은 인공지능 모델이 해결해야 할 Task에 맞추어 훈련에 적합한 데이터 세트를 센싱 모달리티로부터 획득하는 단계로 정의하며, 데이터 규모 확대와 데이터 오류 검증을 위한 크라우드 소싱이나 기존 데이터를 보완, 통합하여 재가공하는 과정을 통해 인공지능 모델학습에 유효한 데이터를 수집할 수 있다.

두 번째로 수집 데이터를 기반으로 인공지능 모델이 학습할 Task에 따라 데이터 레이블을 지정하는 단계에서는 숙련된 작업자가 수동으로 레이블을 지정하여 오류를 줄이고, 레이블링의 일관성 및 규모 확대를 위한 크라우드 소싱을 수행하기도 한다. 또한, 대규모 데이터에서 레이블링에 소요되는 비용과 시간을 줄이기 위하여, 반지도 학습 혹은 약지도 학습 기법을 포함하기도 하는데, 데이터의 오류를 줄이고, 인공지능 모델의 성능을 향상시키기 위해, 일부 데이터를 능동적으로 레이블링하는 Human-in-a-Loop 데이터 레이블링 방식도 활용된다[2].

마지막으로 인공지능 모델학습 과정에서 데이터 처리 과정(c)에서는 데이터 편향성 및 노이즈 데이터 검출과 제거, 라벨링 수정 등을 통해 반복적인 학습 과정에서 발견 가능한 데이터 오류를 정제하고, 학습된 인공지능 모델을 재학습 시키는 방식으로 모델 성능을 개선할 수 있다. 또한 적은 수의 데이터로부터 발생 가능한 인공지능 모델 편향성을 개선하기

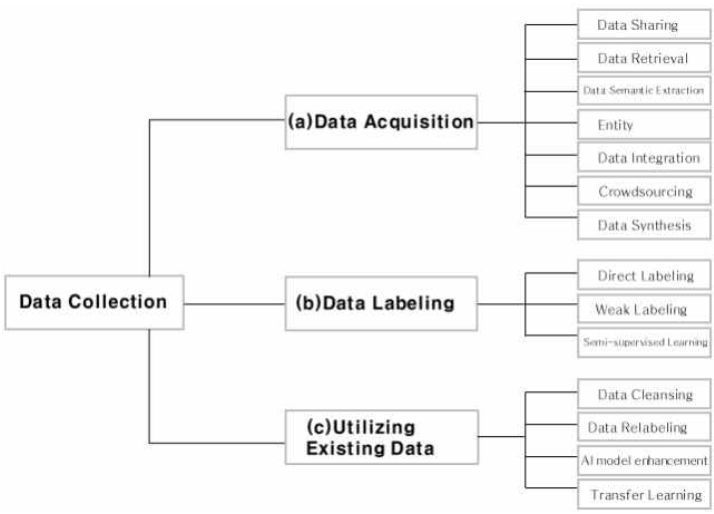


Fig. 3. Data collection approaches for artificial intelligence.

위하여, 대규모 데이터로 사전 학습된 모델을 활용하여 일반화 성능을 높이는 전이 학습 기법을 도입하기도 한다[10]. T. Rukat 등은 대규모 데이터에서의 자동화된 데이터 검증을 위한 단위 테스트 기반 프레임워크인 Deequ를 제안하였다[4]. Deequ는 지속 증가하는 데이터에 대한 인공지능 모델 재학습 과정 중 새로운 데이터에 대한 효율적인 검증을 위하여 기존 데이터에 대한 통계적 정보와 스키마 정보를 유지하여 이상데이터 검출을 수행하고, 학습 과정에서 배제할 수 있도록 한다.

Hynes 등은 Data Linter라는 데이터 검증 도구를 제안하였다[11]. Data Linter의 구조는 Lint Detector, LintExplorer로 구성되며, Fig. 4는 Linter의 데이터 검증 및 인공지능 모델링 프로세스를 소개한다.

Data Linter에서 LintDetector(a)는 데이터의 이상치를 감지하고, LintExplorer(b)가 데이터 이상을 사용자에게 알리고, 데이터의 수정 방향을 제안하는 방식으로 동작한다. 이때 Data Linter는 문제를 감지한 원인과 해당 데이터 인스턴스를 디버깅 코드와 함께 제공한다. 감지 가능한 이상치 항목은 Lint Detector에서 정의할 수 있으며, 사용자의 요구에 따라 확장할 수 있다.

2.4 인공지능 데이터 품질검증 프로세스

Fig. 5는 인공지능 데이터 수집 이후, 데이터 검증

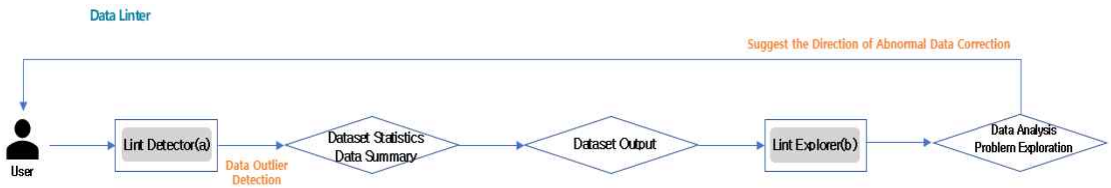


Fig. 4. Data linter for data validation in AI development process.

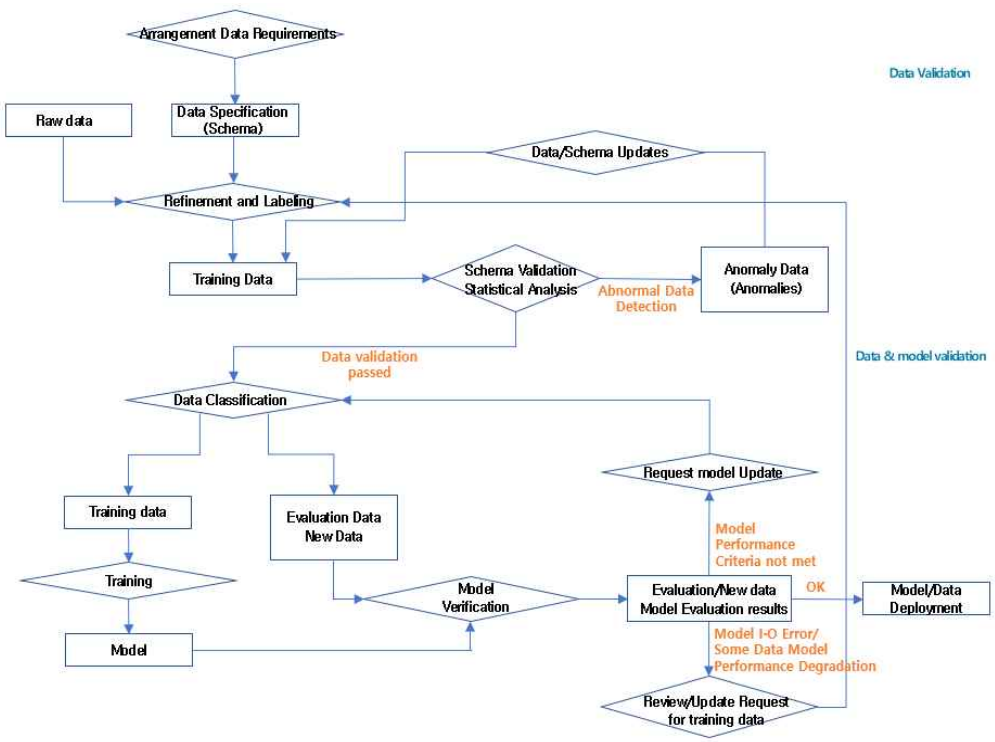


Fig. 5. The AI training data quality validation.

과 데이터 및 모델 검증의 2개 영역으로 나누어진 인공지능 데이터 검증 프로세스를 보여준다. 먼저, 학습 데이터 도출을 위해 데이터의 요구사항을 정리하여 데이터 명세(스키마)를 원시 데이터와 함께 데이터 정제 및 라벨링 과정을 선행한다. 이후 스키마 검증 및 오류통계 분석을 통해 데이터 검증 과정을 거친 데이터는 모델학습 및 평가용 데이터로 구분하고, 문제가 있는 데이터는 이상 데이터(Anomalies)로 검출되어 데이터·스키마 업데이트 과정을 통해 다시 학습데이터 검증을 받는다. 데이터 검증을 통과한 학습 데이터는 특정 모델에 적용함과 동시에 신규 데이터를 포함한 평가용 데이터에도 입력하여 모델 검증에 활용한다. 모델평가를 통해 오류가 없는 학습 데이터는 최종적으로 서비스 데이터로 반출되겠지만, 모델 성능의 기준에 일부 충족되지 못한 데이터는 학습 데이터로 재입력되어 학습을 시키고 모델 입출력 오류나 일부 데이터 모델 성능을 하락시켜 갱신이 필요한 데이터는 정제 및 라벨링 단계로 각각 재입력되어 품질검증 과정을 반복한다.

3. 오픈소스 데이터 품질검증 프레임워크

데이터 품질 관리 프로세스를 반영한 오픈소스 품질검증 프레임워크에는 Cerberus, Voluptuous, Pandera, TFDV 등이 있다[12]. Table 1은 현재 오픈소스 품질검증 프레임워크의 특징을 사용환경, 스키마 검증 지원여부, 데이터 통계 분석, 비정형 데이터 지원, 딥러닝 라이브러리 연계 측면에서 구분하여 설명한다. 오픈소스 데이터 품질검증 프레임워크는 모두 Python 환경에서 동작이 가능하며, 데이터 품질검증 과정 중 스키마 검증 통계 분석과 이상치 검출 기능을 제공하고 있다. 하지만 TFDV를 제외한 나머지 오픈소스 프레임워크는 영상, 텍스트 등 비정형 데이터 검증 기능은 제공하지 않고, TensorFlow나 PyTorch와 같은 딥러닝 라이브러리와 연계되어 인공지능 모델의 성능 평가와 데이터 정제를 수행하는 기능을 포함하지 않는다.

TFDV는 TensorFlow Extended(TFX)에 확장 가능한 데이터 검증 시스템으로써 E. Breck 등이 제안하였다[17]. 해당 시스템은 Data Analyzer, Data Validator 그리고 Model Unit Tester로 구성되어있다. Data Analyzer는 데이터 검증을 위해 미리 정의된 데이터 세트의 통계를 계산하고, Data Validator는 데이터 세트가 정의된 스키마에 일치하는지 검사한다. 마지막으로 Model Unit Tester는 스키마를 통해 생성된 데이터를 사용해 학습 코드에 오류가 있는지 검사한다. 시스템은 싱글 배치 데이터 검증과 학습데이터와 실제 입력데이터 간 큰 차이가 있는지 검사하는 기능을 제공한다.

본 논문에서는 오픈소스 데이터 검증 프레임워크 중 모델 검증과 연계될 수 있는 TFDV를 대상으로 데이터 검증 프로세스 유효성 확인을 위해 의료 기록 메타 데이터와 비정형 영상 데이터로써 복강경 영상 공개 데이터를 사용한 실험을 진행하였다. Table 2는 사용 데이터 세트의 스키마 정보이다. Table 2에서 나타나 있듯이 비정형 데이터 검증 유효성에 사용되는 복강경 영상 공개 데이터 세트에는 비정형 데이터로 내시경 영상과 세그멘테이션이 포함되어 있다.

우선 정형 데이터 검증 유효성 확인을 위해 의료 기록 메타 데이터를 학습용 데이터 세트와 검증용

Table 1. Characteristics of open source data validation libraries [13–16].

Division	TFDV	Cerberus	Voluptuous	Pandera
Use Environment	Python	Python	Python	Python
Schema Validation Support	○	○	○	○
Support for Unstructured Data Validation	○	×	×	×
Data Statistical Visualization	○	×	×	×
Supported Data Format	CSV, Pandas Data Frame, TFRecord	YAML, JSON	YAML, JSON, XML	YAML, Pandas Data Frame
Deep Learning Library Linkage	○ (TensorFlow)	×	×	×
Model Validation Linkage	○	×	×	×

Table 2. Mammography meta-dataset schema [18].

Medical Record Metadata Set	
Characteristic	Type
Patient ID	Integer
File Name	String
Project Group	String
Age	Integer
Sex	String

Laparoscopic Endoscopy Open Dataset	
Characteristic	Type
Image	Byte String
File Formet	String
File Size	String
Image Size	Integer
Bounding Box Size	Real Number
Segmentation Image	Bite String

데이터 세트로 나눈 뒤, 검증 데이터 세트에서는 데이터 명, 데이터 타입, 값 범위 변경, 데이터 일부 삭제 등 데이터 수집과 레이블링 과정에서 발생할 수 있는 데이터 오류을 재현하여, TFDV가 해당 오류를 검출할 수 있는지 실험하였다. 학습 및 검증 데이터 세트는 TFDV에 data frame 형식으로 구현하여 입력하였고, TFDV는 학습용 데이터 세트에서 자동으로 추론된 스키마 구조와 속성, 데이터 값의 통계적 파라미터를 기준으로 검증 데이터 세트의 유효성을 검증한다. Table 3은 학습 데이터를 기반으로 TFDV가 자동으로 추론한 스키마 구조를 보여준다. TFDV는 데이터의 특징 이름, 타입, 수치형 데이터 값 범위, 범주형 데이터 값 종류에 대한 추론이 가능하다.

Fig. 6는 추론한 스키마 기반 유효성 검증 결과이다. Fig. 6(a), 6(b)는 수치형(Numerical) 데이터와 범주형(Categorical) 데이터에 대해 추론한 스키마와 다른 데이터 타입이 있는 경우를 검증한 결과이고, 다른 데이터 타입이 입력된 특징과 입력된 데이터 타입을 출력한다. Fig. 6(c)는 검증용 데이터 세트에서 특징 하나를 삭제하고 검증한 결과로, 추론한 스키마와 비교해 삭제된 특징을 출력한다. Fig. 6(d)는 추론한 데이터 값 범위를 벗어난 값을 입력한 경우로, 이상치가 있는 특징과 벗어난 값의 개수를 정상적으로 출력하는 것을 확인하였다.

Table 3. Schema of training dataset, inferred by TFDV.

Feature name	Type	Presence	Valency	Domain
'ID'	INT	required		-
'File_name'	BYTES	required		-
'group'	STRING	required		'group'
'AGE'	INT	required		min: 1; max: 100
'SEX'	STRING	required		'SEX'

Domain	Values
'group'	'HCPD', 'HCPYA'
'SEX'	'F', 'M'

다음으로 비정형 데이터 검증 유효성을 확인하기 위하여, TFDV에 복강경 영상 데이터 세트를 학습과 검증 세트로 나눈 후 TensorFlow에서 제공하는 이진 데이터 포맷인 TFrecord로 변환하였다. 검증용 데이터 세트는 Fig. 7과 같이 학습용 데이터와 다른

Feature name	Anomaly short description	Anomaly long description
'AGE'	Expected data of type: INT but got FLOAT	

a) Type Anomaly (Numerical Data)

Feature name	Anomaly short description	Anomaly long description
'SEX'	Expected data of type: BYTES but got INT	

b) Type Anomaly (Categorical Data)

Feature name	Anomaly short description	Anomaly long description
'group'	Column dropped	The feature was not present in any examples.

c) Missing Value

Feature name	Anomaly short description	Anomaly long description
'AGE'	Out-of-range values	Unexpectedly large value: 457.

d) Outlier

Fig. 6. Structured data schema validation results.

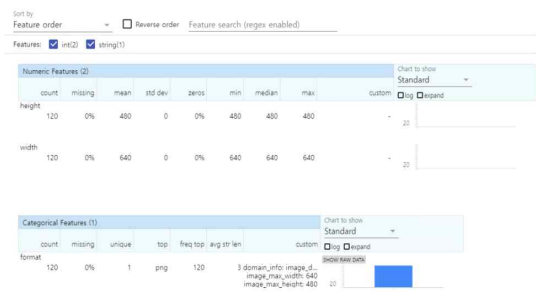
Fig. 7. Training data (left) and abnormal data (right).

영상 모달리티(MRI)에서 획득한 비정상 영상 데이터를 포함하여, TFDV의 비정형 데이터 검증 기능 유효성을 확인하였고, 이와 함께 영상 메타데이터에서의 명칭 변경을 강제하여, 비정형 데이터가 포함된 스키마 검증에 대한 유효성을 함께 확인하였다.

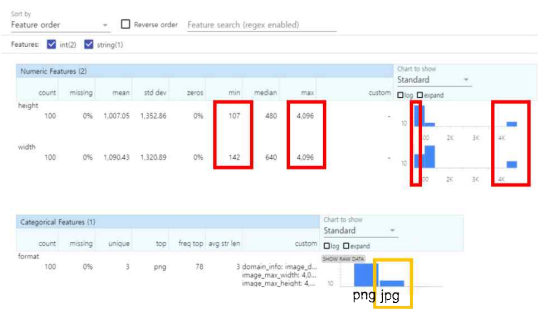
TFDV에서 영상 데이터는 TFrecord 생성 시 바이트 스트링 형식으로 처리되기 때문에, 실험에 사용한 내시경 영상과 세그멘테이션 영상을 바이트 스트링으로 변환하여 저장하였다. 앞선 정형 데이터 실험과 같이 TFDV는 학습용 데이터 세트를 기반으로 먼저 스키마를 추론하고, 검증용 데이터 세트에 대한 스키마 검증을 수행한다. Fig. 8은 임의로 변경된 검증용 데이터 세트에 대해 두 스키마의 차이를 보여준다. 검증 영상의 메타 데이터 명칭 변경으로 인해, 스키마 검증 시 메타 데이터 열을 확인할 수 없다는 메시지와 함께 오류가 정상적으로 검출되었다.

Feature name	Anomaly short description	Anomaly long description
'image/object/mask'	Column dropped	Column is completely missing
'image/object/seg'	New column	New column (column in data but not in schema)

Fig. 8. Schema validation result of unstructured data.



a) Image Statistics of Train Data



b) Image Statistics of Test Data

Fig. 9. Image statistics in TensorFlow data validation.

하지만 스키마 검증는 메타 데이터에 한정되며, TFDV는 ImageStatsGenerator 클래스를 통해 영상 데이터의 높이와 넓이, 파일 포맷을 추론하여, 학습 및 검증 데이터의 통계를 추론하는 기능을 제공한다. Fig. 9는 PNG 포맷으로 저장된 복강경 영상 학습데이터(Fig. 9(a))와 비정상 영상(JPEG 포맷)이 포함된 검증데이터(Fig. 9(b))의 추론된 통계를 보여준다. 학습데이터와 다른 영상 크기 혹은 영상 포맷에 의한 데이터 오류(Fig. 9의 상자표시)를 확인할 수 있었다. 하지만 영상 데이터의 화소값 분석이나 나아가 영상 내용(Context)을 이해하고, 이상 데이터를 검증하는 수준의 의미론적 데이터 검증 기법을 제공하지 않고 있다.

4. 의미론적 데이터 검증 기법

데이터 검증 및 인공지능 학습 프로세스를 반영한 오픈소스 프레임워크는 현재 이미지, 언어 등 비정형 데이터 품질검증을 제공하지 않는다는 한계가 있다. 최근 영상과 문자열 등 비정형 데이터 검증에 관한 연구가 활발히 진행되고 있다. 먼저, 영상 데이터 검증 기법으로써 L. Ruff 등은 이상치 감지 모델로써 딥러닝을 이용한 One-class Deep SVDD라는 모델을 제안하였다[19]. 제안 모델은 데이터를 Feature Space로 맵핑하는 딥러닝 모델학습 과정에서 정상 데이터를 포함하는 Hyper-sphere를 강제하고, 이후 검증 데이터가 정상 데이터의 잠재 벡터가 존재하는 Hyper-sphere 내에 위치하는지에 따라 이상 데이터를 검출한다. One-class Deep SVDD는 MNIST, CIFAR-10 데이터에서 기존 Support Vector Machine 등을 이용한 이상치 탐지 기법에 비해 우수한 성능을 보였다[19]. S. Márquez-Neila 등은 의료 영상에 대한 데이터 검증 딥러닝 모델 Deep Data Validation(DDV)을 제안하였다[19]. DDV는 고해상도인 의료 영상의 특징에 따라 영상을 딥러닝 네트워크를 통해 저차원의 잠재공간으로 투영한 다음 분포를 학습한다. 학습한 분포와 분포가 다른 데이터를 제외하는 방식으로 동작한다. DDV는 망막 OCT, 흉부 X-ray, 뇌 MRI 데이터에 대해 One-clss Deep SVDD 모델보다 우수한 성능을 보였다[20].

문자열 데이터 검증 방법으로써 M. Hulsebos 등은 메타 데이터에서 입력 문자열 데이터가 속하는 컬럼을 자동 분류하는 모델을 제안하였다[21]. 제안

모델인 Sherlock은 문자 분포, 단어 임베딩, 문자열의 통계 파라미터 요소를 통합해 입력 문자열이 속하는 칼럼을 찾는다. Sherlock의 모델 구조는 각 요소를 고정된 크기로 압축하는 부분 신경망과 압축된 요소별 특징 벡터를 통합해 데이터가 포함되는 칼럼을 찾는 분류 네트워크로 구성된다. Sherlock은 머신러닝 모델, 사전 기반 분류, 정규 표현식을 이용한 칼럼 분류 기법 성능과 비교하여 가장 높은 성능을 보였다. 특히, 문자열 기반 메타 데이터 타입 중 성적과 같이 표현 문자열의 범위가 제한되는 경우에서 보다 우수한 성능을 보였다.

G. Pang 등은 딥러닝 기반 의미론적 데이터의 이상 탐지 기법을 딥러닝 잠재 특징 벡터(Latent Feature Vector)를 이용한 이상치 연산 기법, 정상성(Normality) 학습 기법, End-to-End 이상치 학습 기법의 세 가지 분류로 나누어 분석한 리뷰 연구를 발표하였다[22]. 먼저, 딥러닝 특징 벡터 기반 이상치 연산 기법은 영상 분류 모델과 같이 Task 수행을 위해 학습된 딥러닝 모델로부터 정상 데이터와 신규 데이터의 잠재 특징 벡터 간 유사도를 기반으로 이상치를 결정한다. 정상성 학습 기법은 정상으로 구분된 데이터 세트의 분포를 학습하는 비지도 학습 기반 딥러닝 모델을 이용하여, 신규 데이터에 대한 복원 오차 혹은 복원 데이터와 입력 데이터 간 차이 등을 이용하여 이상 데이터를 결정하는 기법 등을 포괄한다[23]. 마지막으로 End-to-End 이상치 학습 기법은 수집 데이터 세트에서 데이터 간 거리 등 이상치를 나타낼 수 있는 척도를 먼저 연산하고, 신규 데이터에 대해 이상치 척도를 추정할 수 있는 딥러닝 모델 구축 기법을 나타낸다. G. Pang 등의 이상 탐지 기법 분석 논문[22]으로부터 정적 영상과 스트리밍 비디오 등에 대한 이상 탐지 기법 연구를 확인할 수 있으며, 이러한 이상 탐지 기법들이 의미론적 데이터 검증 프로세스의 자동화에 활용될 수 있을 것이다.

5. 결 론

본 논문에서는 인공지능 학습용 데이터에 대한 데이터 검증 기법을 검토하였고, 오픈소스 프레임워크에 대한 분석을 소개하였다. 인공지능 개발 과정에 통합된 형태의 학습 데이터 스키마 추론 및 검증 기술, 의미론적 데이터 이상치 탐지 기술은 지속적으로 수집되는 학습 데이터 오류를 줄이고, 데이터 품질을 지속적으로 유지할 수 있게 함으로써 인공지능 모델의 성능 향상을 이끌 것으로 기대한다.

REFERENCE

[1] E. Caveness, et al., “Tensorflow Data Validation: Data Analysis and Validation in Continuous ML Pipelines,” *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. pp. 2793-2794, 2020.

[2] A. Jain, et al., “Overview and Importance of Data Quality for Machine Learning Tasks,” *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. pp. 3561-3562, 2020.

[3] A. Paleyes, R.-G. Urma, and N.D. Lawrence, “Challenges in Deploying Machine Learning: A Survey of Case Studies,” *arXiv preprint*, arXiv:2011.09926, pp. 1-3, pp. 15-16, 2020.

[4] T. Rukat, et al., “Towards Automated ML Model Monitoring: Measure, Improve and Quantify Data Quality,” *ML Ops Workshop at the Conference on Machine Learning and Systems (MLSys)*, pp. 1-2, 2019.

[5] S.E. Whang and J.-G. Lee, “Data Collection and Quality Challenges for Deep Learning,” *Proceedings of the VLDB Endowment 13.12,* pp. 3429-3431, 2020.

[6] S. Saria and A. Subbaswamy, “Tutorial: Safe and Reliable Machine Learning,” *ACM Conference on Fairness, Accountability, and Transparency, Atlanta, Ga.* pp. 1-3, 2019.

[7] M. Armbrust, et al., “Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics,” *CIDR,* pp. 1-4, 2021.

[8] K. Kumar, *New Trends in Data Warehousing Techniques*, ResearchGate, 2020.

[9] J.-C. Kim, et al., “A Study on Automatic Missing Value Imputation Replacement Method for Data Processing in Digital Data,” *Journal of Korea Multimedia Society*, Vol. 24. No. 2, pp. 245-246, 2021.

[10] Y. Roh, et al., "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 33, No. 4, pp. 1328-1330. 2021.

[11] N. Hynes, D. Sculley, and M. Terry, "The Data Linter: Lightweight, Automated Sanity Checking for ML Data Sets," *NIPS MLSys Workshop.* pp. 1-3, 2017.

[12] V. Shah, K. Yang, and K. Kumar, "Improving Feature Type Inference Accuracy of TFDV with SortingHat," Corpus ID: 235273771, pp. 1-7, 2020.

[13] TFDV(2021), https://www.tensorflow.org/tfx/guide/tfdv (accessed October 8, 2021).

[14] Cerberus(2021), https://docs.python-cerberus.org/en/stable/ (accessed October 8, 2021).

[15] Voluptuous(2021), https://github.com/alectho-mas/voluptuous (accessed October 8, 2021).

[16] Pandera(2021), https://pandera.readthedocs.io/en/stable/ (accessed October 8, 2021).

[17] E. Breck, et al., "Data Validation for Machine Learning," *MLSys*. pp. 2-4, 2019.

[18] Laparoscopic Endoscopy Open Dataset(2021), https://opencas.webarchiv.kit.edu/?q=node/30 (accessed October 8, 2021).

[19] L. Ruff, et al., "Deep One-class Classification," *International Conference on Machine Learning,* PMLR 80, pp. 3-5, 2018.

[20] P. Márquez-Neila and R. Sznitman, "Image Data Validation for Medical Systems," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 1-2, 2019.

[21] M. Hulsebos, et al., "Sherlock: A Deep Learning Approach to Semantic Data Type Detection," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1500-1504, 2019.

[22] G. Pang, et al., "Deep Learning for Anomaly Detection: A Review," *ACM Computing Surveys*, Vol. 54, Issue 2, pp. 1-8, 2021.

[23] S.-H. Jeong, et al., "A Study on Classification Evaluation Prediction Model by Cluster for Accuracy Measurement of Unsupervised Learning Data," *Journal of Korea Multimedia Society*, Vol. 21. No. 7, pp. 779-780, 2018.

윤 창 희

1999년 2월 단국대학교 정보관리학과(학사)
2017년 7월 University for Peace, Sustainable Peace in the Contemporary World (석사)
2021년 2월 경북대학교 정보과학과 박사수료
2016년 7월~2018년 7월 러시아 SKOLKOVO FOUNDATION ICT 자문관
1999년 7월~현재 한국지능정보사회진흥원 수석연구원
관심분야 : 인공지능, 기계학습, 통신공학

신 호 경

2020년 2월 경북대학교 컴퓨터학부(학사)
2020년~현재 경북대학교 컴퓨터학부 대학원 석사과정
관심분야 : Deep Learning, Medical Imaging, Data Validation

추 승 연

1994년 2월 경북대학교 건축공학과(공학사)
1998년 8월 홍익대학교 건축학과(공학석사)
2004년 2월 독일 뮌헨 공대 건축학과(공학박사)
2004년 2월~2005년 1월 독일 뮌헨공대 건축학과 CAAD 연구소 수석연구원
2005년 2월~현재 경북대학교 건축학부 교수
2014년 12월~2015년 8월 산학연대구지협의회 회장
2018년 6월~2019 5월 경북대 산학협력단 부단장
관심분야 : 건축계획 및 설계, AI 건설자동화, BIM, GIS, AR/VR, FM, DFS

김 재 일

2007년 2월 아주대학교 미디어학(공학사)
2007년 3월~2015년 2월 한국과학기술원 전산학(공학박사)
2014년 11월~2016년 5월 삼성전자 책임연구원
2016년 6월~2018년 2월 University of North Carolina at Chapel Hill 박사후연구원
2018년 3월~현재 경북대학교 컴퓨터학부 조교수
관심분야 : 영상분석, 기계학습, 인공지능