데이터 분석 및 머신러닝을 위한 데이터 품질 표준화

하수욱 한국전자통신연구원 지능정보표준연구실 책임연구원
정의현 안양대학교 융합소프트웨어학과 교수

1. 머리말

빅데이터 분석이나 인공지능과 같은 데이터 기반의 정보기술 분야, 특히 머신러닝 분야에서 데이터의 품질 관리를 중요하게 고려해야 한다. 데이터 분석 알고리즘이나 머신러닝을 위한 학습 모델이 아무리 우수하더라도 저품질의 데이터가 입력될 경우 해당 결과물은 신뢰할 수 없기 때문이다. 따라서 데이터를 다루는 대부분의 조직은 데이터 수집 단계부터 데이터를 분석, 학습하거나 시각화하기까지의 데이터 생애주기에 따라 데이터 품질에 대한 요구사항을 식별하고 개선하기 위해 노력한다. 본고에서는 빅데이터 분석 및 머신러닝 분야에서 데이터 생애주기에 따른 품질 고려사항과 함께, JTC 1 SC 42 인공지능 데이터 작업그룹(WG2)을 중심으로 진행 중인 데이터 품질 표준화 동향에 대해 살펴본다.

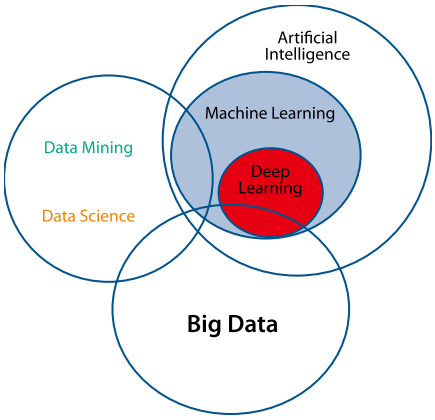
2. 빅데이터 분석과 머신러닝

2.1 데이터 과학

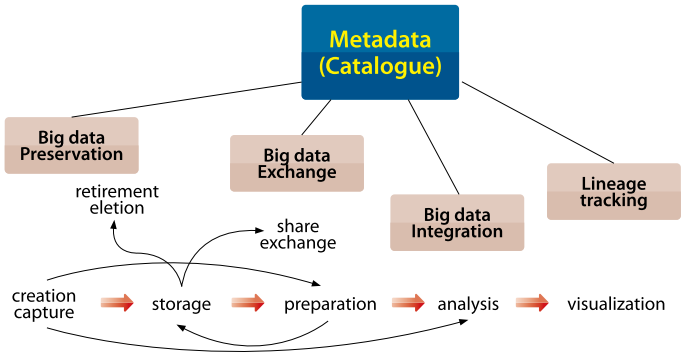
데이터 과학(Data science)이란 데이터에서 지식을 추출하는 절차를 말한다[1]. 전통적인 데이터 과학에서는 통계학에 기초한 데이터 마이닝 방법이 주로 사용됐으며, 이는 데이터 과학자들이 가설을 수립하고, 수집된 데이터를 바탕으로 가설을 테스트함으로써 검증하는 방법으로 진행됐다. 그러나 최근에는 대용량의 데이터, 그리고 이를 처리할 수 있는 컴퓨팅 능력을 기반으로 컴퓨터가 학습을 통해 데이터의 특성을 분석하고 처리하는 방법이 더 각광받고 있다.

2.2 빅데이터 분석

데이터의 수집 및 저장, 관리 관점에서 빅데이터 기술의 등장은 '데이터 범용성'의 기반을 제공했다. 저마다 목적을 가지고 생성된 이질적인 데이터들이 다양한 분야에 광범위하게 사용되기 시작한 것이다. 범용성은 대용량 데이터를 효율적으로 저장하고 필요한 데이터만 자유롭게 뽑아서 사용하게 한 데이터 파이프라인 기술과 자유롭게 확장 가능한 클라우드 컴퓨팅 환경에 힘입어 가능했다. 또한 이러한 변화는 기존의 고비용의 RDBMS 환경과 별도로 업무에 종속적이지 않으면서도 대용량의 축적된 정보 속에서 새로운 가



[그림 1] 빅데이터, 머신러닝과 데이터 과학[2]



[그림 2] 빅데이터 생애주기와 메타데이터 [ITU-T Y.3603][4]

치를 찾아내는 '데이터 기반 기술 생태계(data-driven ecosystem)'의 발전을 가속화했다.

2.3 머신러닝

머신러닝(ML, Machine Learning)이란 시스템이 데이터, 또는 경험으로부터 학습할 수 있도록 컴퓨터 기술을 사용하는 프로세스를 의미한다[3]. 데이터 마이닝에 사용되는 알고리즘 상당수가 머신러닝 분야에서도 똑같이 사용된다. 다만 동일한 알고리즘일 경우 알고리즘을 튜닝하는 데 사용되는 함수의 파라미터를 데이터 마이닝에서는 변수(variable)로, 머신러닝 분야에서는 하이퍼 파라미터(hyperparameter)로 달리 명명한다는 차이가 있다. 이는 데이터 마이닝이 주어진 데이터 속에서 데이터의 패턴과 규칙을 찾는 것에 집중하여 그 분석 결과를 리포팅하는 데 치중한다면, 머신러닝은 주어진 데이터를 바탕으로 예측 모델을 생성하고 이를 새롭게 수집되는 데이터에 적용하여 미래를 예측하는 데 더욱 초점을 맞추기 때문이다.

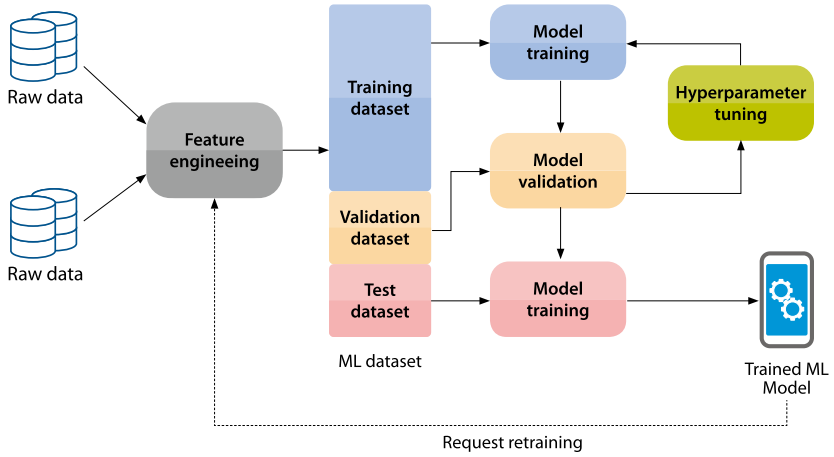
3. 데이터 생애주기

3.1 빅데이터 라이프사이클

빅데이터 라이프사이클은 보편적으로 수집, 저장, 추출 및 전처리, 분석, 시각화의 단계로 구분된다. 엄밀히 말하면 이는 빅데이터 분석의 절차라고 할 수 있으며, 데이터의 흐름과 생애주기 관점으로 한정할 경우 [그림 2]에 나타난 바와 같이 데이터의 생성(creation and capture), 저장(storage), 준비(preparation), 은퇴/삭제(retirement and deletion), 공유(share and exchange)로 구분할 수 있다. 이러한 절차는 시스템 또는 용도에 따라서 자유롭게 구성될 수 있다.

3.2 일반적인 머신러닝 절차

ITU-T Y.3531은 클라우드 환경에서 머신러닝 서비스를 제공하는 데 필요한 기능적 요구사항을 정의하는 국제표준이다. 이 표준에서 기술하는 일반적인 머신러닝 절차를 살펴보면 데이터 수집 및 저장, 데이터 준비(또는 전처리)과정들은 생략하고 머신러닝의 핵심 절차인 테스트를 통한 모델 검증, 그리고 튜닝의 반복 수행을 통한 모델의 생성에 초점이 맞춰졌음을 알 수 있다. 데이터의 흐름 관점에서 살펴보면 원시 데이



[그림 3] 일반적인 머신러닝 절차의 예 [ITU-T Y.3531][5]

터(raw data)는 데이터의 생성과 저장을 포괄하며, 피처 엔지니어링(feature engineering) 절차는 데이터 준비 또는 전처리 과정과 모델 학습(model training) 과정은 데이터 마이닝 또는 빅데이터 분석(analysis) 절차와 매핑된다. 유일한 차이점은 그 결과다. 빅데이터 분석의 결과가 시각화를 통하여 사용자에게 전달되는 것이라면, 머신러닝의 작업 결과는 생성된 모델을 응용에 적용하는 것이다.

3.3 데이터 생애주기와 데이터 품질 요소

데이터 품질 요소를 정의하는 국제표준인 ISO/IEC 25012는 데이터 품질을 '내재된 데이터 품질', '내재적이며 시스템에 의존적인 데이터 품질', 그리고 '시스템에 의존적인 품질' 크게 세 가지로 구분하고, 이를 다시 15개 품질로 세분화하여 제시한다[6]. 이에 비해 Zaveri는 링크드 데이터를 위한 품질 요소로 접근성, 고유성, 상황, 그리고 표현의 4가지 품질 카테고리 아래 18개의 하위 품질 요소들로 정의했다[7]. 이들 품질 요소를 기초로 빅데이터 분석과 머신러닝을 위한 데이터 생애주기를 데이터 수집, 데이터 전처리, 데이터 탐색, 데이터 분석(모델 훈련), 적용(또는 시각화)의 단계로 단순화하여 개별 단계에 따른 품질 요소를 살펴보면 다음과 같다.

3.3.1 데이터 수집

데이터 수집 단계에서는 데이터와 관련된 문제를 파악하고 정의한다. 파일, 데이터베이스, 인터넷, 센서 등 다양한 데이터 소스에서 필요한 데이터를 어떻게 수집할 것인지, 또한 복수의 데이터 소스로부터 수집된 데이터들을 어떻게 결합할 것인가에 대한 정책 또한 수립해야 한다.

3.3.2 데이터 전처리

데이터 전처리는 수집된 원천 데이터 세트를 정제하고 변환하여 유용한 형태로 만드는 과정이다. 원천 데이터 세트에는 결측치, 데이터 중복, 이상 데이터, 잡음 등이 포함되기 마련이므로 이 과정에서는 정제(cleaning), 특정 변수 선택, 특정 형태로의 데이터 변환 등 다음 단계에서 활용하기 쉽도록 구성하는 활동을 진행한다. 이 단계에서는 다양한 데이터 품질 요소가 고려돼야 하며, 향후 빅데이터 분석 및 머신러닝 학습의

예측 품질에 막대한 영향을 미치므로 가장 중요한 과정이라고 할 수 있다. 이 단계에서 적용 가능한 데이터 품질 척도는 <표 2>와 같다.

3.3.3 데이터 탐색

이 단계에서는 다양한 탐색적 데이터 분석 기법과 기술 통계를 이용하여 데이터 분석 및 머신러닝에 적용할 모델을 선택한다. 문제의 성질에 따라서 분류(Classification), 회귀(Regression), 클러스터 분석(Cluster Analysis), 연관(Association) 등 어떤 종류의 머신러닝 기법과 모델을 사용할 것인지 정한다. 라벨링이 필요한 모델을 선택하는 경우에는 라벨링을 진행하며, 학습용 데이터 세트와 테스트용 데이터 세트를 따로 준비한다. 이 단계에서 적용되는 데이터 품질 척도는 다음 <표 3>과 같다.

3.3.4 데이터 분석 및 적용

위의 단계를 통해 도출된 데이터 분석 결과, 또는 머신러닝 모델은 자체적인 평가 기준이나 성능 기준에 따라 정량적으로 측정하고 평가할 수 있다. 때에 따라서는 사용된 데이터의 이력 정보나 편향성 정보 등을 통해 간접적인 품질 정보를 전달하기도 한다. 이 단계에서 고려할 수

<표 1> 데이터 수집 단계에서의 데이터 품질 척도

차원	정의
추적성 (Traceability)	데이터에 대한 액세스 감사 추적 및 특정 사용 콘텍스트에서 데이터에 대한 변경 사항을 제공하는 속성이 데이터에 있는 정도이다.
신뢰성 (Credibility)	특정 사용 상황에서 사용자가 사실로 간주하고 믿을 수 있는 속성이 데이터에 있는 정도이다. 신뢰성에는 진정성(원산지, 귀속, 약속의 진실성)의 개념이 포함된다.
기밀성 (Confidentiality)	특정 사용 콘텍스트에서 권한이 부여된 사용자만 액세스하고 해석할 수 있도록 데이터에 속성이 있는 정도이다. 기밀성은 ISO / IEC 13335-1 : 2004에 정의된 정보 보안(가용성, 무결성과 함께)의 한 측면이다.
복구성 (Recoverability)	특정 사용 상황에서 오류 발생 시에도 지정된 수준의 작업 및 품질을 유지하고 보존할 수 있는 특성이 있는 정도이다.

<표 2> 데이터 전처리 단계에서의 데이터 품질 척도

차원	정의
효율성 (Efficiency)	특정 사용 콘텍스트에서 적절한 양과 유형의 리소스를 사용하여 처리할 수 있고, 예상되는 성능 수준을 제공할 수 있는 속성이 데이터에 있는 정도이다.
준수성 (Compliance)	데이터가 특정 사용 상황에서 시행 중인 표준, 관습 또는 규정 및 데이터 품질과 관련된 유사한 규칙을 준수하는 속성을 갖는 정도이다.
이해도 (Understandability)	데이터를 사용자가 읽고 해석할 수 있는 속성이 있고 특정 사용 콘텍스트에서 적절한 언어, 기호 및 단위로 표현되는 정도이다. 데이터 이해 가능성에 대한 일부 정보는 메타데이터에서 제공된다.

<표 3> 데이터 탐색 단계에서의 데이터 품질 척도

차원	정의
일관성 (Consistency)	데이터가 모순이 없고 특정 사용 콘텍스트에서 다른 데이터와 일관된 속성을 갖는 정도이다. 하나의 엔티티에 관한 데이터와 비교 가능한 엔티티에 대한 유사한 데이터 사이에서 또는 둘 다일 수 있다.
적시성 (Currentness)	데이터가 특정 사용 콘텍스트에서 적절한 시간의 속성을 갖는 정도이다.
정확성 (Accuracy)	개념 또는 이벤트의 특정 문맥에서의 사용이 의도된 속성의 실제 값을 정확하게 나타내는 속성이 데이터에 있는 정도이다.
정밀도 (Precision)	데이터가 정확하거나 특정 사용 상황에서 차별성을 제공하는 속성을 갖는 정도이다.
수고도 (Effort)	어느 정도의 노력으로 인공지능 학습용 데이터 세트를 만들 수 있는지 나타내는 정도이다.

있는 데이터의 품질 척도는 <표 4>와 같다.

4. JTC 1 SC42 데이터 분석 및 머신러닝을 위한 데이터 품질 표준화 현황

4.1 배경 및 경과

JTC 1 SC 42는 인공지능과 관련한 국제표준 개발을 담당하고 있으며, WG2에서는 인공지능 및 빅데이터를 위한 표준 갭분석 및 관련 표준 개발을 추진해 오고 있다. WG2는 2019년 4월 일본 총회에서 빅데이터 품질을 위한 표준 개발 필요성이 제기됨에 따라 빅데이터 품질 임시 그룹(Big data quality Ad-hoc Group)을 결성했으며, 2020년 4월까지 빅데이터 분석 및 머신러닝 분야에서 데이터 품질 표준의 필요성과 함께 수행되어야 할 표준화 아이템을 발굴하는 작업을 추진했다. 이러한 과정을 거쳐 2020년 7월부터 [표 5]의 4개 항목에 대한 표준 개발이 진행되고 있다.

4.2 5259-1 개요, 용어 및 예제

ISO/IEC 5259-1은 빅데이터 분석 및 머신러닝 분야에서의 데이터 품질에 대한 개요를 제공할 목표로 개발 중이다. 데이터 품질의 개념 및 품질 척도에 대한 정의를 포함하는 품질 프레임워크와 함께, 공통으로 쓰이는 용어를 정의하고 전체 시리즈 표준의 구성과 연관성에 대한 전반적인 범위를 포함한다. 본 프로젝트는 우리나라 주도로 개발이 진행되고 있으며, 2023년까지 개발을 완료할 계획이다.

4.3 5259-2 데이터 품질 측정

이 표준은 빅데이터 분석 및 머신러닝 모델에 대한 품질을 측정하고 이를 보고하는 절차에 대한 지침을 제공하는 것을 목적으로 한다. 데이터 품질 관련 ISO 8000 시리즈를 기반으로 소프트웨어 및 데이터 품질에 대한 ISO/IEC 25012와 ISO/IEC 25024의 사상을 수용하여 조직이 데이터 품질 목표를 수립하고 달성하기 위한 가

<표 4> 모델 훈련 단계에서의 데이터 품질 척도

차원	정의
완성도 (Completeness)	엔티티와 연관된 주제 데이터가 특정 사용 콘텍스트에서 예상되는 모든 속성 및 관련 엔티티 인스턴스에 대한 값을 갖는 정도이다.
공정도 (Fairness)	입력데이터의 편향이 어느 정도 분포를 지니는지 나타내는 정도이다.
라벨링 편향 (Labeling Bias)	라벨링의 편향이 어느 정도 분포를 지니는지 나타내는 정도이다.
변화도 (Variance)	원본 데이터에서 어느 정도의 변화가 있었는지 나타내는 정도이다.

<표 5> 데이터 품질 표준화 추진 현황 (JTC 1 SC 42 WG 2)

담당 (WG)	프로젝트명	제목	Target schedule	프로젝트 에디터
WG 2	ISO/IEC 5259-1	Artificial intelligence - Data quality for analytics and ML - Part 1: Overview, terminology, and example	CD 10/21· DIS 10/22 FDIS 4/23· IS 10/23	하수욱 (한국)
WG 2	ISO/IEC 5259-2	Artificial intelligence - Data quality for analytics and ML - Part 2: Data quality measure	CD 4/22· DIS 4/23 IS 4/24	김경숙 (일본)
WG 2	ISO/IEC 5259-3	Artificial intelligence - Data quality for analytics and ML - Part 3: Data quality management requirements and guidelines	CD 10/21· DIS 10/22 FDIS 4/23· IS 10/23	Martin Saerbeck (독일)
WG 2	ISO/IEC 5259-4	Artificial intelligence - Data quality for analytics and ML - Part 4: Data quality process framework	CD 10/21· DIS 10/22 FDIS 4/23· IS 10/23	Wanzang Ma (중국)

이드를 제공하는 데 중점을 둔다. 2020년 11월 현재 NWIP(New Work Item Proposal)에 대한 투표가 진행 중이며, 큰 문제 없이 승인될 것으로 예상된다.

4.4 5259-3 데이터 품질 관리 요구사항 및 가이드라인

이 표준은 빅데이터 분석 및 머신러닝에 사용되는 데이터의 품질을 설정, 구현, 유지하고 지속적으로 개선하기 위한 요구사항과 지침을 제공하는 것을 목적으로 한다. 이를 위해 자세한 수준의 절차와 방법을 제공하기보다, 조직의 규모나 성격과 관계없이 데이터 품질 관리를 위해 적용할 수 있는 상위 수준의 권장사항을 개발할 계획이다.

4.5 5259-4 데이터 품질 절차 프레임워크

이 표준은 빅데이터 분석 및 머신러닝에서 사용되는 데이터 품질을 보장하기 위해 적용 조직의 유형, 규모 또는 특성에 무관하게 적용 가능한 일반적인 사항 관련 지침을 제공하는 데 목적을 둔다. 이 표준은 데이터 라이프사이클에 따라 서로 다른 소스로부터 수집된 데이터에 대한 학습과 평가에 전반적으로 활용할 수 있도록 개발할 계획이다.

5. 맺음말

본고에서는 데이터 사이언스 분야, 특히 빅데이터 분석 및 머신러닝 분야에서의 데이터 품질 이슈와 품질 요소들을 확인하고, 국제표준화기구인 JTC 1 SC 42 WG 2에서 개발 중인 데이터 품질 관련 표준화 항목들에 대하여 살펴보았다. 이전에는 데이터의 용도가 제한적이라 비교적 용이하게 품질을 식별할 수 있었다. 그러나 빅데이터 분석 및 인공지능의 시대에 접어들어 수많은 데이터가 다양한 용도로 활용되면서 더 이상 전통적인 방법으로는 품질을 식별하기 어려워졌다. 이러한 상황에서 현재 추진 중인 데이터 품질 표준은 데이터의 신뢰성과 품질을 향상시키는 기폭제로 사용될 수 있을 것으로 기대된다.

※ 본 연구는 '지능정보기술 확산을 위한 인공지능 데이터 표준 개발(2020-0-00895)' 과제의 일환으로 수행되었다.

참고문헌

[1] ISO/IEC 20546 Information technology - Big data - Overview and vocabulary, 2019.

[2] Visually Linking AI, Machine Learning, Deep Learning, Big Data and Data Science, https://whatsthebigdata.com/2016/10/17/visually-linking-ai-machine-learning-deep-learning-big-data-and-data-science/

[3] ISO/IEC CD 23053.2 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), 2020

[4] ITU-T Y.3604, Big data - Requirements and conceptual model of metadata for data catalogue. 2019. https://www.itu.int/rec/T-REC-Y.3603-201912-I/en

[5] ITU-T Y.3531, Cloud computing - Functional requirements for machine learning as a service, 2020, https://www.itu.int/rec/T-REC-Y.3531-202009-I

[6] IS/IEC 25012, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model, 2008

[7] Amrapali Zaveri et al. 'Quality assessment for Linked Data: A Survey. Semantic Web', vol. 7, no. 1, pp. 63-93, 2015. URL: https://dx.doi.org/10.3233/SW-150175