Journal of Digital Convergence
Vol. 16. No. 10, pp. 205-212, 2018

ISSN 1738-1916
https://doi.org/10.14400/JDC.2018.16.10.205

기계학습 활용을 위한 학습 데이터세트 구축 표준화 방안에 관한 연구

최정열
성결대학교 컴퓨터공학부 부교수

A study on the standardization strategy for building of learning data set for machine learning applications

JungYul Choi
Division of Computer Engineering, Sungkyul University, Associate professor

요 약 고성능 CPU/GPU의 개발과 심층신경망 등의 인공지능 알고리즘, 그리고 다량의 데이터 확보를 통해 기계학습이 다양한 응용 분야로 확대 적용되고 있다. 특히, 사물인터넷, 사회관계망서비스, 웹페이지, 공공데이터로부터 수집된 다량의 데이터들이 기계학습의 활용에 가속화를 가하고 있다. 기계학습을 위한 학습 데이터세트는 응용 분야와 데이터 종류에 따라 다양한 형식으로 존재하고 있어 효과적으로 데이터를 처리하고 기계학습에 적용하기에 어려움이 따른다. 이에 본 논문은 표준화된 절차에 따라 기계학습을 위한 학습 데이터세트를 구축하기 위한 방안을 연구하였다. 먼저 학습 데이터세트가 갖추어야할 요구사항을 문제 유형과 데이터 유형별로 분석하였다. 이를 토대로 기계학습 활용을 위한 학습 데이터세트 구축에 관한 참조모델을 제안하였다. 또한 학습 데이터세트 구축 참조모델을 국제 표준으로 개발하기 위해 대상 표준화 기구의 선정 및 표준화 전략을 제시하였다.

주제어 : 기계학습, 인공지능, 참조모델, 표준화, 학습 데이터세트

Abstract With the development of high performance CPU / GPU, artificial intelligence algorithms such as deep neural networks, and a large amount of data, machine learning has been extended to various applications. In particular, a large amount of data collected from the Internet of Things, social network services, web pages, and public data is accelerating the use of machine learning. Learning data sets for machine learning exist in various formats according to application fields and data types, and thus it is difficult to effectively process data and apply them to machine learning. Therefore, this paper studied a method for building a learning data set for machine learning in accordance with standardized procedures. This paper first analyzes the requirement of learning data set according to problem types and data types. Based on the analysis, this paper presents the reference model to build learning data set for machine learning applications. This paper presents the target standardization organization and a standard development strategy for building learning data set.

Key Words : Machine learning, Artificial Intelligence, Reference model, Standardization, Learning data set

1. 서론

2016년 3월 이세돌 9단과 구글의 알파고의 바둑 대결 이후로 국내에 인공지능에 대한 열기가 뜨겁게 달아오르고 있다. 기존에는 이미지 처리, 음성 인식, 패턴 인식, 자연어 처리, 로보틱스 등 전문가 그룹에서 수행되던 인공

* Corresponding Author : JungYul Choi (passjay@gmail.com)
Received July 25, 2018 Revised September 18, 2018
Accepted October 20, 2018 Published October 28, 2018

지능 기술이 기계학습과 딥러닝 기술을 중심으로 빅데이터와 데이터사이언스 분야에서도 활발히 적용되고 있다 [1-3]. 또한 사용자 성향을 반영하여 음악을 재생해주는 인공지능 스피커, 사용자 정보에 기반한 광고 및 검색 등 지식서비스, 무인 자율주행자동차, 무인항공기 등 인공지능을 활용한 응용 서비스들이 우리의 일상에 보편화되기 시작했다.

인공지능 기술을 활용하기 위해서는 무엇보다도 강력한 컴퓨팅 성능, 효율적인 인공지능 알고리즘, 그리고 인공지능을 학습시킬 위한 충분한 양의 데이터가 요구된다 [4]. 클라우드 컴퓨팅 인프라와 고성능 GPU의 활용을 통해서 오랜 시간 처리해야 하는 인공지능 알고리즘을 보다 빠르고 신속하게 처리할 수 있게 되었다. 예를 들어 구글의 알파고는 약 1200개의 CPU를 이용해서 구현되었다. 인공지능 알고리즘도 심층신경망, 강화학습 등 지속적으로 발전하고 있어 기존에 해결하지 못하는 문제들에 적용하여 다양한 분야로 응용되고 있다 [5].

사물인터넷 기술의 등장과 사회관계망서비스 등을 통해서 생산되는 방대한 양의 데이터가 수집됨으로 인해서 기존에는 분석하기 어려웠던 정보들이 발굴되고 있다. 따라서 전통적인 인공지능 분야뿐 만 아니라 빅데이터와 데이터사이언스 분야에서도 데이터의 활용과 처리에 대한 중요성이 더 부각되고 있다. 이를 위해서 구글, 페이스북, 네이버, KT 등과 같은 포털, 사회관계망서비스업체 및 통신회사 등은 데이터 수집을 위해서 부단히 노력하고 있다. 또한 빅데이터 활용을 위해서 공공데이터포탈 등도 구축되어 여러 응용분야에 적용되고 있다 [6-9]. 이러한 상황에서 보다 다양한 서비스 및 응용이 개발되고 이들을 쉽게 활용하기 위해서는 무엇보다도 잘 정리된 데이터의 수집 및 관리가 요구된다 [10,11].

이를 위해서 본 논문은 기계학습을 위한 학습 데이터세트 구축 방안을 연구하고, 이러한 데이터세트가 실제로 활용되기 위해서 선행되어야하는 표준화 방안에 대해서 살펴본다. 이를 위해서 2장에서는 국내외에서 진행되고 있는 데이터세트와 관련된 표준화 동향을 살펴본다. 그리고 현재 구축되어 활용되고 있는 데이터세트 현황을 3장에서 살펴본다. 4장에서는 기계학습 활용을 위한 학습용 데이터세트의 요구사항을 문제 유형과 데이터 유형에 따라 살펴보고, 학습용 데이터세트 구축 참조 모델을 제시한다. 5장에서는 이러한 데이터세트 구축 방안에 따라서 국제표준으로 개발하기 위한 절차와 전략을 제시한다.

2. 표준화 동향

2.1 국내 표준화 동향

한국정보통신기술협회(TTA)가 발간한 ICT표준화전략맵 ver.2018은 인공지능 분야의 주요 표준화 항목으로 기계학습·추론 및 플랫폼 기술, 상황·감성 이해, 음성·언어 이해, 시각·영상·동작 이해, 사용자 맞춤형 지능정보 체계로 제시하고 있다 [12]. 그 중에서도 기계학습·추론 및 플랫폼 기술 분야는 세부적으로 머신러닝 데이터 구축, 클라우드 기반 지능형 인프라 관리, 빅데이터 연동 지능형 서비스로 구분하여 표준화 전략을 제시하고 있다. 특히, 머신러닝 데이터 구축 분야의 주요 표준화 내용으로는 “머신러닝 기법을 적용하기 위해 입력으로 사용될 다양한 형태의 원시 데이터를 수집하고 정제하여 데이터세트를 구축하는 표준 기술“로 제시하고 있다 [12]. 다만, 해당 분야와 직접적으로 관련된 국내 표준화 활동은 없으며 국제 표준화 기구인 ITU-T SG13에서 해당 표준을 다룰 것을 제시하고 있다.

데이터세트가 활용될 주요한 분야 중에서 빅데이터 관련 표준화 활동은 TTA의 빅데이터 특별기술위원회(SPG22)[1]에서 진행 중이다. 현재 개발된 주요 표준안으로 빅데이터 제공 서비스 개념, 정의 및 유스케이스(TTAK.KO-10.0974) [13]과 빅데이터 제공 서비스 요구사항 및 기능 구조 (TTAK.KO-10.0975) [14]가 있다. 위 표준안 모두 데이터세트와 구축과 직접적인 관련은 없으나 향후 데이터세트 표준안이 마련된다면 기계학습은 물론 빅데이터 분야에서도 쉽게 활용될 것으로 기대된다.

2.2 국제 표준화 동향

인공지능 분야의 국제표준화는 분야별로 ITU-T, JTC 1, ISO 등 다양한 표준화기구에서 진행되고 있다. ITU-T SG13는 미래 네트워크와 클라우드 분야 표준화를 담당하고 있으며, 클라우드 컴퓨팅 환경과 빅데이터 환경에서 인공지능을 활용하기 위한 논의를 하고 있다[2]. 사물인터넷과 스마트시티 분야 표준화를 담당하고 있는

1) 한국정보통신기술협회 (TTA), 정보통신표준화위원회 산하 빅데이터특별기술위원회 (SPG22), http://committee.tta.or.kr/standard/general.jsp?commit_code=STC2

2) ITU-T SG13: Future networks, with focus on IMT-2020, cloud computing and trusted network infrastructures, www.itu.int/en/ITU-T/studygroups/2017-2020/13/

SG20[3]에서는 인공지능을 활용한 사물인터넷 기술에 대한 논의를 하고 있다 [15]. ISO TC37는 언어 자원 분야의 표준화를 담당하고 있으며, 자연어 질의응답 서비스 프레임워크와 메터데이터에 대한 표준을 개발하고 있다[4]. JTC 1 SC 29는 오디오, 이미지, 멀티미디어 코딩 등의 표준을 담당하고 있으며, 미디어 정보 교환을 위한 프레임워크 및 메타데이터 등을 개발하고 있다[5].

미래네트워크, 클라우드 컴퓨팅, 사물인터넷, 멀티미디어 등의 기존의 표준 작업반에서 인공지능 기술을 활용하는 표준화 노력 외에 ITU-T와 JTC 1은 인공지능 분야에 특화된 표준 작업 그룹을 신설하고 표준화를 진행하고 있다. 먼저 ITU-T는 2017년 11월에 5G를 포함한 미래 네트워크를 위한 기계학습 분야의 포커스 그룹을 SG13 산하에 신설하였다[6]. 해당 포커스그룹은 3개의 작업반으로 구성되는데 WG1은 유즈케이스, 서비스 및 요구사항을, WG2는 데이터 포맷 및 기계학습 기술을, WG3은 기계학습 인지형 네트워크 구조에 대한 연구를 수행한다.

ISO/IEC JTC 1은 2017년 10월 인공지능의 용어, 프레임워크, 핵심 기술과 서비스 분야의 표준화를 담당하기 위해서 SC 42 신설을 승인하였으며, 2018년 4월에 1차 총회가 개최되었다[7]. SC 42는 인공지능 기반 표준을 개발할 1개의 작업반과 신규 표준화 항목 발굴을 위한 3개의 연구반 (SG1: 인공지능 시스템, SG2: 인공지능 신뢰성, SG3: 사례 및 응용)으로 구성된다. 현재는 JTC 1/WG 9으로부터 이관받은 빅데이터 참조구조 [16]를 표준화하고 있으며, 새롭게 인공지능 개념 및 용어 [17]와 기계학습을 위한 인공지능 시스템 프레임워크 [18]에 대한 표준 개발이 승인되었다. 향후 학습 데이터세트 및 유즈케이스 등 다양한 인공지능 분야의 표준이 개발될 것으로 기대된다.

이와 같이 인공지능 분야에서 다양한 표준화 노력이 진행되고 있지만, 기계학습 및 빅데이터 처리를 위한 데이터세트에 대한 중요성에도 불구하고 표준화는 이제 시작 단계에 머무르고 있다. 다음 절에서는 먼저 데이터세트 구축 현황을 살펴보고, 이어서 학습 데이터세트 표준화 구축 방안과 표준화 전략을 논의한다.

3. 데이터세트 구축 현황

사물인터넷의 등장과 다양한 사회관계망서비스의 활성화, 그리고 웹서비스 등으로 인해서 폭발적으로 발생한 빅데이터를 효과적으로 처리하기 위한 노력이 수년간 진행되어 왔다. 특히, 정부에서는 공공데이터 활성화를 위해서 정부와 지자체를 중심으로 공공데이터세트를 구축하고 이를 활성화하기 위한 노력을 진행하고 있다. 다음 표는 국내에서 구축되어 활용되는 대표적인 공공데이터세트를 제시한다.

Table 1. Example of public data set

Data set name	Data set	Dat area	Data format
Open Data Portal	file: 22,486 open api: 2,535	education, financial, security, health, weather, science, welfare, food, culture, law etc	map, chart, xls, json, cvs, rdf, lod, open api,
Seoul Open Data Plaza	file: 4952, OpenAPI: 4469	health, culture, administration, industrial, traffic, education, security, population, etc	xls, csv, hwp, chart, openapi, map, link lod
Gyeonggi Data Dream	file: 1470	traffic, construction, family, health, tour, fire, industrial, education, housing	xls, csv, json, xml, txt, rdf, chart, map, api
Healthcare Bigdata Hub	data: 38, OpenAPI: 17	medicine, hospital, patient	xls, csv, docx, hwp, json

제공되는 공공데이터는 각종 통계 처리와 기계학습을 위한 데이터로 활용할 수는 있으나, 기관마다 제공되는 데이터의 형식이 달라 서비스 개발 및 알고리즘에 적용하기 위해서는 개발자가 포맷을 다시 수정해야 하는 어려움이 있다.

한편, 기계학습을 위한 데이터세트는 응용별로 이미지 데이터, 텍스트 데이터, 음성 및 오디오 데이터 등이 존재한다 [19,20]. 다음 표는 연구용으로 사용되고 있는 데이터세트들을 제시한다.

3) ITU-T SG20: Internet of things (IoT) and smart cities and communities (SC&C), www.itu.int/en/ITU-T/studygroups/2017-2020/20/

4) ISO/TC 37 Language and terminology, https://www.iso.org/committee/48104.html

5) ISO/IEC JTC 1/SC 29 Coding of audio, picture, multimedia and hypermdeia information, https://www.iso.org/committee/45316.html

6) ITU-T Focus Group on Machine Learning for Future Networks including 5G (FG-ML5G), https://www.itu.int/en/ITU-T/focusgroups/ml5g/

7) ISO/IEC JTC 1/SC 42 Artifical Intelligence, https://www.iso.org/committee/6794475.html

Table 2. Example of data sets for machine learning

Data type	Data set	Amount	Feature	Reference
image	ImageNet	14million	image object recognition	www.image-net.org
	MNIST	60 thousand	classification (handwrighting digits)	http://yann.lecun.com/exdb/mnist/
	CIFAR- 100	60 thousands	classification (100 different images)	http://www.cs.toronto.edu/~kriz/cifar.html
text	IMDB Reviews	50 thousand	sentiment analysis	http://ai.stanford.edu/~amaas/data/sentiment/
	Yelp Review	5 million	reviews on restaurant, hospital, shops, etc	https://www.yelp.com/dataset
audio	Free Spoken Digit Dataset	1.5thousand	classification (10 voice recoded digits)	https://github.com/Jakobovski/free-spoken-digit-dataset
	Million Song Dataset	1million	feature and meta data	https://labrosa.ee.columbia.edu/millionsong/

기계학습을 위한 데이터세트들은 응용 서비스의 종류와 그 대상에 되는 데이터에 따라서 다양한 종류가 있다. 연구자들이나 서비스 개발자들은 기존에 구축된 데이터세트들을 이용하기 위해서는 개별 데이터세트들의 특징과 사용법들을 익혀야 한다. 하지만 기존에 구축되어 사용 중인 데이터세트들이 상당히 많이 있기 때문에 매번 다양한 데이터세트들을 활용하기가 어렵다. 특히 학습용 데이터세트를 새롭게 구축할 경우에는 데이터세트에 관한 요구사항이나 구축 절차가 명시적으로 존재하지 않기 때문에 연구자나 개발자들이 참조할 수 있는 표준화된 참조모델이 요구된다.

4. 학습 데이터세트 구축 방안

4.1 학습 데이터세트 요구사항 분석

3장에서 살펴본 바와 같이 기존에 구축되어 활용되는 기계학습용 데이터세트들은 도메인 (교육, 금융, 범죄, 인구통계, 건강, 음식 등), 문제 유형(회귀, 분류, 군집화 등), 그리고 데이터 유형(텍스트, 이미지, 음성, 음악, 영상 등)에 따라 다양하게 구축되었다. 또한 제공되는 자료의 형식(파일 포맷)도 다양하게 존재한다. 특정 문제를 해결하기 위해서 학습 데이터세트를 구축하기 위해서는 일반적으로 다음과 같은 사항들을 고려해야 한다[2,21].

- 학습 데이터세트가 적용하고자 하는 문제와 밀접한 연관성을 가지고 있는가
- 학습 데이터세트가 기계학습 모델을 수립하기 위해서 충분한 규모로 제공되는가
- 학습 데이터세트가 전체 데이터를 대변하고 있으며, 특정 부분에 치우쳐 있지는 않은가
- 데이터들이 기계학습 모델에 적용하기에 용이한 파일 형식을 가지고 있는가
- 데이터에 누락된 자료가 있거나 이상치를 포함하고 있는가
- 데이터에 식별자가 잘 설정되어 있는가
- 데이터가 특정 문제를 해결하기 위해서 충분한 특징(속성)을 포함하고 있는가

위 요구사항의 핵심은 과연 '구축된 학습 데이터세트가 주어진 문제를 해결할 수 있는가'이다. 만약 구축된 학습 데이터세트가 주어진 문제를 해결하기에 충분하지 않은 규모이거나 적절한 특징을 포함하고 있지 않다면 데이터세트를 재구축해야 한다.

기계학습 모델을 수립하기에 충분한 학습 데이터세트의 규모와 관련해서는 다음의 예를 고려할 필요가 있다. 분류 문제의 경우에는 적절한 규모의 학습 데이터세트를 선택하기 위해서 다음과 같은 사항을 고려해야 한다. N을 데이터세트의 크기라 하고, d를 데이터의 특징(또는 차원, 속성)이라 하고, k를 분류 문제에 있어서 분류의 수라고 하고, m을 데이터세트에서 소수 샘플의 비율이라고 가정하자. 먼저 $d > k$를 만족해야 하는데, 이는 데이터세트들의 속성이 분류의 수보다는 커야함을 의미한다. $N > kd$를 만족해야 하는데, 이는 데이터세트의 수가 분류의 수보다 커야 함을 의미한다. 분류 항목별로 고르게 데이터세트가 구성되기 위해서는 $m \sim N/k$가 되는 것이 바람직하다.

추가로 문제 유형별, 데이터 유형별 학습 데이터세트를 구축하기 위해서 표 3과 표 4와 같은 사항을 고려할 수 있다 [2,3].

Table 3. Requirement of data sets for problem types

Problem types	Requirements
regression	inputs and outputs should be paired for modeling data should be in form of numerical data
classification	data should be labeled data should be provided uniformly for each classification items, and not biased to specific items
clustering	data should contain specific feature for clustering the definition of similarity taking into account a given problem should be defined in advance

Table 4. Requirement of data sets for data types

Data types	Requirement	Data format
image	image standard and size should be unified for the data set data should contain specific feature for the given problem	JPG, BMP, TIFF, etc
text	data should contain specific feature for the given problem data set should be stored in a specific file format for the processing software	CSV, JSON, XLS, TXT, etc
audio	noise in audio data should be minimized to extract feature extract the feature from data set by preprocessing tone, height of sound	WAV, MP3, etc
numerical data	variable types should be selected considering the range of numerical value should use encoding method that the processing system can handle	String, Integer, Double, Boolean, etc

4.2 학습 데이터세트 구축 참조모델

기계학습에 활용하기 위한 학습 데이터세트는 다음과 같은 과정을 거쳐서 구축될 수 있다.

1 단계: 학습 데이터세트의 요구사항 도출

4.1절에서 살펴본 바와 같이 구축하고자 하는 학습 데이터세트가 어떤 문제를 해결하고자 하는지, 문제의 유형은 무엇인지, 그리고 어떤 유형의 데이터를 대상으로 하는지를 고려하여 고려해야할 사항들을 검토한다.

2 단계: 요구사항을 고려한 데이터 수집

주어진 문제를 해결하기 위한 모델을 수립하기 위해서 충분한 양의 데이터를 수집해야 한다. 앞서 지적하였듯이 기계학습을 위해서 가장 중요한 부분이 문제와 높은 연관성을 가지고 문제를 해결하기에 충분한 데이터를 확보하는 것이다. 텍스트 데이터의 경우는 웹페이지, 트위터 등과 같은 사회관계망서비스, 신문 기사 등으로부터 크롤링해서 수집할 수 있다 [22]. 따라서 데이터 수집을 위한 크롤링 프로그램을 구축해야 한다. 이때 모델을 수립하기에 적합한 키워드를 사전에 선정해야 한다. 수집한 데이터의 형식(포맷)은 상이할 수는 있으나 학습 데이터세트로 구축하기 위해서는 동일 파일 형식으로 변환해야 한다 [23].

3 단계: 수집한 데이터의 전처리 수행

데이터 전처리 과정은 수집한 데이터를 문제 해결을 위한 모델에 활용하기에 적합한 데이터로 처리하는 과정이다. 예를 들어, 이상데이터 처리, 누락데이터 처리, 데이터 포맷 변환, 데이터 정규화 과정 등을 포함한다. 데이터의 전처리 과정을 거친 학습 데이터세트는 모델 수립에 활용되기 때문에 적절한 전처리 과정은 문제 해결을 위한 모델 수립 과정에서 매우 중요한 역할을 하게 된다.

데이터 전처리 과정에서 다음과 같은 사항을 고려해야 한다 [2]. 수집한 데이터에 중복 데이터가 존재하는지, 데이터에 누락이 있는지, 데이터 값이 NULL인지, 데이터에 심각한 이상치가 있는지, 오래된 데이터가 있는지, 불연속적인 데이터 등이 있는지를 확인하고, 이를 해결해야 한다. 또한 수집된 데이터의 형식이 각기 다른 경우, 예를 들어 상이한 공백문자 처리, 불규칙한 대소문자, 일정하지 않은 구분 기호, 각기 다른 날짜 표기 등의 경우 이를 일관되게 처리해야 한다. 이렇게 데이터 저장 형식이 일정하지 않은 경우에는 정규표현식을 사용하는 것이 바람직하다 [24].

4 단계: 학습 데이터세트 구축

수집한 데이터에 이상값을 처리하는 등의 전처리 과정을 수행한 다음에 모델 수립에 활용할 수 있는 학습 데이터세트를 구축한다. 이 과정에서는 데이터가 내포하고 있는 특징값을 추출하게 된다 [25]. 예를 들어 텍스트 데이터의 경우 문서에 수많은 단어가 있는데, 단어의 길이, 단어의 등장 빈도수, 단어의 전후 관계 등이 특징값이 될 수 있다. 특징값을 추출하는 과정은 데이터를 가공하여 표 형태로 구성할 수 있다. 표의 행이 데이터의 표본이 되고, 표의 열이 그 데이터의 특징이 될 수 있다. 만약 구축하고자 하는 학습 데이터세트가 이미지라면 이 과정에서 이미지 데이터에 레이블을 추가하게 된다. 이 과정은 도메인 전문가의 영역이라고 볼 수 있으며, 특징값을 어떻게 추출할 지에 따라서 문제의 해결 방법 및 모델 수립에 큰 영향을 미치게 된다.

학습 데이터세트를 모델 수립에 직접 활용하기에 앞서 샘플 데이터세트를 구축할 수도 있다. 간단한 모델을 구성하고 샘플 데이터세트를 적용하여, 구축한 학습 데이터의 적절성 여부를 판단한다. 만약에 모델을 수립하는데 있어서 심각한 오류가 있다고 판단되면, 데이터 수집을 다시 수행할 수도 있으며, 특징값을 다르게 추출할 수도 있다.

5 단계: 학습 데이터세트 구성: 학습용, 검증용, 평가용

구축된 데이터세트를 활용해서 모델을 개발하고, 개발된 모형이 주어진 문제를 해결하기에 적절한지를 판단하기 위해서 데이터세트를 학습 데이터, 검증 데이터, 그리고 평가 데이터로 나눈다 [3]. 학습 데이터는 주어진 문제를 해결하기 위한 모델을 훈련하는데 사용되며, 이를 활용하여 기계학습의 주요 파라미터 결정(예: 신경망 가중치, 바이어스 등)를 결정한다. 검증 데이터는 학습 데이터를 이용해서 수립한 모델의 하이퍼파라미터 (예: 신경망 구조, SVM의 커널 함수, 랜덤 포레스트 트리의 깊이 등) 조정하기 위해서 사용된다. 검증 데이터 없이 학습 데이터만으로 모델을 훈련할 수도 있지만, 하이퍼파라미터를 정교하게 최적화하고 모델의 오버피팅을 방지하기 위해서 별도로 제공할 수 있다. 평가 데이터는 구축된 모델이 학습 데이터 외에 다른 데이터에서도 잘 동작하는지를 판단하기 위해서 활용된다.

구축된 학습 데이터세트를 학습용, 검증용, 그리고 평가용으로 구분하기 위해서 고려해야 할 사항은 다음과 같다.

- 주어진 학습 데이터세트를 어떠한 비율로 나눌 것인가

주어진 학습 데이터세트를 어떤 비율로 학습용, 검증용, 평가용으로 나눌지에 대한 명시적인 연구 결과나 요구사항은 없다. 모델 수립을 위한 학습 데이터를 높은 비율로 선정하고, 검증 및 평가용 데이터는 적은 비율로 선정한다. 예를 들어서 5 : 2.5 : 2.5이거나 6 : 2 : 2 등으로 선정할 수 있다.

- 주어진 학습 데이터세트를 어떠한 방식으로 나눌 것인가

첫 번째 사항보다 더 중요하게 고려해야할 사항으로, 모델이 어떠한 데이터에서도 잘 동작하기 위해서는 학습 데이터가 전제 데이터세트에서 편중되지 않고 고르게 선택되어야 한다. 한 가지 방법은 주어진 비율에 따라서 무작위로 평가용, 검증용, 평가용 데이터를 선택하는 것이다. 만약 시간 정보를 가지고 있는 데이터의 경우는 시간순으로 오래된 데이터를 학습용으로 선택하고, 최신 데이터를 평가용으로 선택할 수 있다. 좀 더 보편적인 방법으로 K 겹 교차 검증 (K fold cross validation) 기법을 활용할 수 있다 [26]. 이 방식은 데이터세트를 K개로 나누고, 그 중 첫 번째 세트를 평가용으로 나머지를 학습용으로 구분하여 모델을 수립하고 평가한다. 그 다음에는 두 번째 세트를 평가용으로 나머지를 학습용으로 구분하여 모델을 수립하고 평가한다. 총 K번 반복하여 최종 모델을 평가하므로 더 안정적이고 정확하나 계산량이 많다는 단점이 있다.

5. 학습 데이터세트 구축 표준화 전략

5.1 대상 표준화기구 선정 및 표준 개발 준비

학습 데이터세트는 기계학습 기반의 여러 응용서비스에 쉽게 적용될 수 있도록 표준화된 절차에 따라 구축될 필요가 있다. 본 절에서는 학습 데이터세트의 구축 참조모델을 국제 표준으로 개발하는 전략에 대해서 기술한다. 먼저 고려해야할 사항은 어떤 표준화 기구에서 표준안을 개발할 것인지를 선택해야 한다. 3장에서 살펴본 바와 같이 인공지능 관련 국제 표준화기구로서 ITU-T FG-ML5G와 JTC 1/SC 42가 있다. 두 기구 중에서 FG-ML5G는 포커스 그룹으로서 표준안을 직접 개발하기 보다는 표준 개발을 위한 선행 작업을 하고, 그 결과물(deliverable)을 실제 표준을 개발하는 연구반에 제공한다. 따라서 학습 데이터세트 참조모델 표준을 개발하기 위해서는 JTC 1 /SC42에서 담당하는 것이 적합하다고 판단된다.

SC 42는 2018년 4월에 1차 총회를 개최하여 향후 개발할 표준 연구 아이템들을 수집하고 있다. 국제 표준은 표준을 제안한 국가에서 주도적으로 개발하게 되기 때문에 신규 연구 아이템 제안 및 선점이 무엇보다 중요하다. 국내 기술을 바탕으로 국제 표준을 개발하기 위해서는 먼저 국내의 기계학습 응용을 위한 학습 데이터세트 관련 학계, 연구소, 산업계 전문가와 사전에 표준 개발에 대한 논의가 필요하다. SC 42 대응 국내 연구반에서는 국내외 연구 동향을 파악하고, 특히 타국가의 표준화 전략을 사전 검토하여 국내 기술 기반의 표준화 개발 방향 및

대응 방안을 결정해야 한다.

5.1 학습 데이터세트 구축 참조모델 개발절차

학습 데이터세트 구축 참조모델의 국제표준 개발 절차는 다음과 같이 진행될 필요가 있다.

1 단계: 학습 데이터세트 요구사항 개발

먼저 국제표준을 개발하기 위해서는 표준 개발에 대한 필요성을 각국에 인지시켜 동의를 얻어야 한다. 이를 위해서 국제표준 개발의 첫 단계로서 인공지능 서비스를 위한 학습 데이터세트의 요구사항을 도출하고 학습 데이터세트의 사례를 분석한다. 다양한 사례 분석은 표준 개발에 대한 필요성을 제시하게 된다. 이를 토대로 문제 및 데이터 유형별 학습 데이터세트 요구사항 및 구축 참조모델에 대한 신규 표준 제안(New Work Item Proposal, NP)을 제안한다.

2 단계: 학습 데이터세트 구축 참조모델 개발

신규 표준 제안(NP)이 승인되면 본격적으로 학습 데이터세트 구축 참조모델을 개발한다. 4장에서 소개한 바에 따라서, 문제 유형 및 데이터 유형에 따라 학습 데이터세트의 샘플을 구축한다. 샘플 구축과정에서 발생한 문제와 이들 해결하기 위한 사항 등을 반영하여 구축 참조모델을 확정하여 작업 문서(Working Draft, WD)로 완성한다. 문제 유형와 데이터 유형이 다양하기 때문에 학습 데이터세트 구축에 관한 절차를 기술하는 일반 참조모델을 개발하고, 문제 유형와 데이터 유형에 따른 여러 참조모델을 후속 표준으로 개발할 수도 있다. 작업 문서가 완성된 이후 각국의 의견을 수렴하여 위원회 안(Committee Draft, CD)으로 작성한다. 위원회 안이 통과되면 마지막으로 DIS(Draft International Standard)로 개발을 한 뒤에 타 위원회의 검토를 거쳐 최종 국제표준으로 승인을 추진하게 된다.

3 단계: 학습 데이터세트 활용 유즈케이스 개발

개발한 학습 데이터세트의 구축 참조모델이 국제표준으로서 성공적으로 활용되기 위해서는 참조모델을 활용한 유즈케이스 개발이 필요하다. 이 과정이 국제표준 개발에 필수 단계는 아니지만, 다양한 서비스와 데이터 유형에 따라 구축 참조모델이 잘 적용될 수 있는 지는 확인하는 단계를 거치는 것이 바람직하다. ITU-T FG-ML5G에서 기계학습의 유즈케이스를 개발하고 있기 때문에, JTC 1 SC 42와 협력하여 학습 데이터세트 구축 참조모델을 활용한 유즈케이스를 개발할 수도 있을 것이다.

국제 표준안은 프로젝트의 제안에서부터 국가 간의 의견수렴을 통한 프로젝트 승인, 그리고 기술적인 내용을 담은 프로젝트 개발, 마지막으로 개발된 권고안을 검토하여 최종 표준으로 완성되기 까지 수년이 소요된다. 따라서 국내 주도의 국제 표준을 개발하기 위해서는 우리나라와 우호적인 협력 관계를 갖는 국가들과의 공조가 요구된다. 그리고 프로젝트 에디터는 문서 개발에 대한 책임이 있으므로 지속적인 표준화 회의 참여가 이루어질 수 있도록 전문가 지원이 필요하다. 또한 국가의 이익을 위해 국내 기술이 국제 표준에 반영될 수 있도록 국내 연구소 및 산업계의 적극적인 참여가 요구된다. 특히, SC 42 대응 국내 연구반이 국내 기술의 국제표준 개발에 주도적인 역할을 해야 하며, 필요할 경우 관련 포럼과의 협력이 요구된다. 표준을 통한 산업화에 기여할 수 있도록 해당 기술에 대한 표준특허의 개발이 요구된다.

6. 결론

기계학습은 인공지능 응용서비스는 물론 빅데이터와 데이터사이언스 분야에서도 널리 사용되고 있다. 기계학습의 효과적인 활용을 위해서는 충분히 많은 데이터가 확보되어야 한다. 사물인터넷, 사회관계망서비스, 인공지능 스피커, 웹페이지, 공공데이터로부터 수집되고 있는 다량의 데이터들이 기계학습의 활용을 가속화하고 있다. 그럼에도 불구하고 데이터의 종류의 응용 분야에 따라 데이터 양식과 내용이 상이하여 개발자들이 쉽게 활용하기 어려운 점이 있다. 따라서 본 논문에서는 기계학습을 위한 학습 데이터세트의 표준화가 필요함을 역설하였다. 2018년 4월에 신설된 JTC 1 SC 42는 인공지능 분야의 표준화 기구로서 학습 데이터세트의 표준화를 위해서 적합한 기구이다. 본 논문에서 제안한 바에 따라서 기계학습을 위한 학습 데이터세트의 구축 참조모델을 구축하고, 기능 검증 및 활용을 위한 유즈케이스 개발을 단계적으로 수행해나간다면 기계학습이 다양한 서비스 분야에서 쉽게 활용될 것으로 기대된다. 향후 국내 표준화 위원회를 바탕으로 국내 기술을 바탕으로 참조모델을 개발하고, 이를 통하여 국제 표준안을 개발한다면 국내 인공지능 기술의 확산에 기여할 것으로 판단된다.

REFERENCES

[1] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis & K. Taha. (2015). Efficient machine learning for big data: A review. *Big Data Research, 2(3),* 87-93.

[2] F. Cady. (2017). *The data science handbook.* Wiley.

[3] S. Kim & Y. Jeong. (2017). Machine Learning. Seoul : Hanbit Media.

[4] A. Moore & M. S. Lee. (1988). Cached sufficient statistics for efficient machine learning with large datasets. *Journal of Artificial Intelligence Research, 8,* 67-91.

[5] S. Lee. (2017) Artificial Intelligence: Recent Artificial Intelligence Development Trend and Future Evolution Direction. LG Economic Research Institute(Online). http://www.lgeri.com/report/view.do?idx=19584

[6] Open Data Portal, https://www.data.go.kr/

[7] Seoul Open Data Plaza, http://data.seoul.go.kr/

[8] Gyeonggi Data Dream, https://data.gg.go.kr/portal/mainPage.do

[9] Healthcare Bigdata Hub, http://opendata.hira.or.kr/

[10] T. Borovicka, M. Jirina Jr, P. Kordik & M. Jirina. (2012). Selecting representative data sets. *Advances in data mining knowledge discovery and applications,* InTech. 43-70.

[11] A. Ferrari, G. O. Spagnolo & S. Gnesi. (2017). Towards a Dataset for Natural Language Requirements Processing, *Proceedings of International Working Conference on Requirements Engineering: Foundation for Software Quality.*

[12] ICT Standardization Strategy Map ver.2018. (2017). Telecommunications Technology Association.

[13] TTAK.KO-10.0974, Big Data - Definition, Concept and Use Cases of Data Providing Service, *TTA Standards.* (2016)

[14] TTAK.KO-10.0975, Big Data - Requriements and Functional Architecture for Data Providing Service. *TTA Standards,* 2016.

[15] ITU-T SG20 TR.AI4IoT, Artificial Intelligence and Internet of Thing (work in progress), 2018

[16] ISO/IEC AWI TR 20547-1, Information technology -- Big data reference architecture -- Part 1: Framework and application process

[17] ISO/IEC AWI 22989, Artificial Intelligence Concepts and Terminology (work in progress),

[18] ISO/IEC AWI 23053, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) (work in progress)

[19] List of datasets for machine learning research, https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research, *Wikipedia,* 2018.

[20] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, & P. Lamere. (2011). The Million Song Dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference, 2(9),* 591-596.

[21] S. J. Raudys & A. K. Jain. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 13(3),* 252-264.

[22] G. Pant, P. Srinivasan & F. Menczer. (2004). Crawling the web. *Web Dynamics,* 153-177.

[23] J. Morcos, Z. Abedjan, I. F. Ilyas, M. Ouzzani, P. Papotti & M. Stonebraker. (2015). Dataxformer: An interactive data transformation tool. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data,* 883-888.

[24] Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan & H. V. Jagadish. (2008). Regular expression learning for information extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 21-30.

[25] I. Guyon & A. Elisseeff. (2006). An introduction to feature extraction. *In Feature extraction.* Springer, 1-25.

[26] T. Fushiki. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing, 21(2),* 137-146.

최 정 열(Choi, Jung Yul) [정회원]

- 2000년 2월 : 인하대학교 전자공학과(공학사)
- 2006년 8월 : 한국과학기술원 정보통신공학과(공학박사)
- 2006년 9월 ~ 2011년 2월 : (주) KT 선임연구원
- 2011년 3월 ~ 현재 : 성결대학교 컴퓨터공학부 부교수
- 관심분야 : 기계학습, 데이터사이언스, 사물인터넷, 그린 통신망
- E-Mail : passjay@gmail.com