기계학습에 유효한 데이터 요건 및 선별: 공공데이터포털 제공 데이터 사례를 통해

오효정[1], 윤보현[2*]
[1]전북대학교 문헌정보학과 부교수, [2]목원대학교 소프트웨어교양학부 교수

Valid Data Conditions and Discrimination for Machine Learning: Case study on Dataset in the Public Data Portal

Hyo-Jung Oh[1], Bo-Hyun Yun[2*]
[1]Professor, Dept. of Library & Information Science, Jeonbuk National University
[2]Professor, Div. of Software Liberal Arts, Mokwon University

요 약 인공지능 기술의 가장 큰 근간은 학습 가능한 데이터이다. 최근 정부나 사기업에서 수집·생산하는 데이터의 종류와 양이 기하급수적으로 증가하고 있지만, 실제 기계학습에 활용 가능한 데이터의 확보로는 아직까지 이어지지 않고 있다. 이에 본 연구에서는 기계학습에 실제 활용 가능한 데이터가 갖추어야 할 조건에 대해 논의하고, 실제 사례연구를 통해 데이터 품질을 저하시키는 요인을 파악한다. 이를 위해 공공빅데이터를 활용해 예측 모델을 개발한 대표사례를 선정, 공공데이터포털로부터 실제 문제 해결을 위한 데이터를 수집 후 데이터 품질을 확인하였다. 이를 통해 유효한 데이터 선별 기준을 적용하고 후처리한 결과와의 차이를 보인다. 본 연구의 궁극적인 목적은 인공지능의 핵심인 기계학습 기술 개발에 앞서 가장 근본적으로 선결되어야 할 데이터 품질을 관리하고 유효한 데이터를 축적하기 위한 기반 마련에 있다.

주제어 : 유효 데이터, 기계학습, 데이터 선별, 데이터 품질, 공공빅데이터

Abstract The fundamental basis of AI technology is learningable data. Recently, the types and amounts of data collected and produced by the government or private companies are increasing exponentially, however, verified data that can be used for actual machine learning has not yet led to it. This study discusses the conditions that data actually can be used for machine learning should meet, and identifies factors that degrade data quality through case studies. To this end, two representative cases of developing a prediction model using public big data was selected, and data for actual problem solving was collected from the public data portal. Through this, there is a difference from the results of applying valid data screening criteria and post-processing. The ultimate purpose of this study is to argue the importance of data quality management that must be most fundamentally preceded before the development of machine learning technology, which is the core of artificial intelligence, and accumulating valid data.

Key Words : Valid Data, Machine Learning, Data Discrimination, Quality of Data, Public Big data

본 논문은 2021년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음.
본 논문은 2021년도 한국연구재단 연구비 지원에 의한 결과의 일부임 ((NRF-2021R1I1A3047435)
*교신저자 : 윤보현(ybh@mokwon.ac.kr)
접수일 2021년 11월 20일 수정일 2022년 1월 11일 심사완료일 2022년 1월 15일

1. 서론

IDC(International Data Corporation)의 '2021 세계 인공지능 시장 전망 예측 보고서'[1]에 의하면 2024년까지 인공지능 시장은 연평균 17.5%의 성장률을 보이고 총 매출은 5,543억 달러에 이를 것이라고 한다. 인공지능 기술은 금융, 교통, 농업, 헬스케어, 법률 등과 같은 다양한 분야의 사업에 적용되고 있으며 여러 플랫폼과 접목하여 서비스되고 있다. 이처럼 인공지능 기술이 여러 분야에서 활용됨에 따라 그 위상이 높아지면서, 인공지능 기술의 기반이 되는 데이터 또한 주목을 받고 있다. 한국의 경우 2020년 디지털뉴딜 정책의 핵심 과제 중 하나로 인공지능 기술의 고도화를 위한 데이터 댐 구축 사업을 선정하였으며[2,3], 미국은 2019년 'OPEN Government Data Act'의 제정을 통해 공공데이터의 기계판독성(machine-readable)을 강조하고, AI R&D 전략에서 인공지능 기술기반 확보를 위해 데이터 구축을 주요 목표로 설정하고 있다[4]. 그 외에도 캐나다, EU, 영국, 중국 등도 데이터를 구축하고 인공지능 연구를 선도하기 위한 전략을 내세우고 있다.

그러나 정부나 사기업에서 수집·생산하는 데이터의 종류와 양이 기하급수적으로 증가하면서 데이터의 품질과 관련된 문제가 발생하고 있다[5]. Viscusi 외[6]는 이탈리아의 지방자치단체가 제공하는 공공데이터의 품질을 평가하였는데, 지방자치단체가 운영하는 포털의 40%가 불완전한 데이터를 제공하고 있었으며, 기계판독이 가능하지 않은 형식의 데이터셋이 55% 이상이라고 보고하였다.

기업 역시 데이터 품질과 관련된 애로를 겪고 있는 것으로 파악되었다. 가트너(Gartner)의 연구에 의하면 기업이 낮은 품질의 데이터 처리에 매년 1,500만 달러의 비용을 소모하고 있으며, 이는 정보환경이 복잡해짐에 따라 점점 더 악화될 것이라고 한다[7]. 낮은 품질의 데이터로 인한 손실을 줄이고 데이터 분석과 인공지능 기술의 효율을 높이기 위해 데이터 품질관리가 필요하지만, 대다수의 기업은 데이터의 관리에 대해 중요하게 생각하지 않고 있으며, 학문적인 연구 또한 미흡한 실정이다[8,9].

이에 본 연구에서는 인공지능의 가장 핵심이 되는 기계학습에 실제 활용 가능한 데이터가 갖추어야 할 조건에 대해 논의하고, 실제 사례연구를 통해 데이터 품질을 저하시키는 요인을 파악한다. 이를 위해 공공빅데이터를 활용해 예측 모델을 개발한 대표사례를 선정, 공개된 공공빅데이터 유형을 분석하고 실제 수집 후 데이터 품질을 확인하였다. 이를 통해 유효한 데이터 선별 기준을 적용하고 후처리한 결과와의 차이를 보인다. 본 연구의 궁극적인 목적은 인공지능의 핵심인 기계학습 기술 개발에 앞서 가장 근본적으로 선결되어야 할 데이터 품질을 관리하고 유효한 데이터를 축적하기 위한 기반 마련에 있다.

2. 기계학습에 활용 가능한 데이터 조건

2.1 밀도 높은 데이터

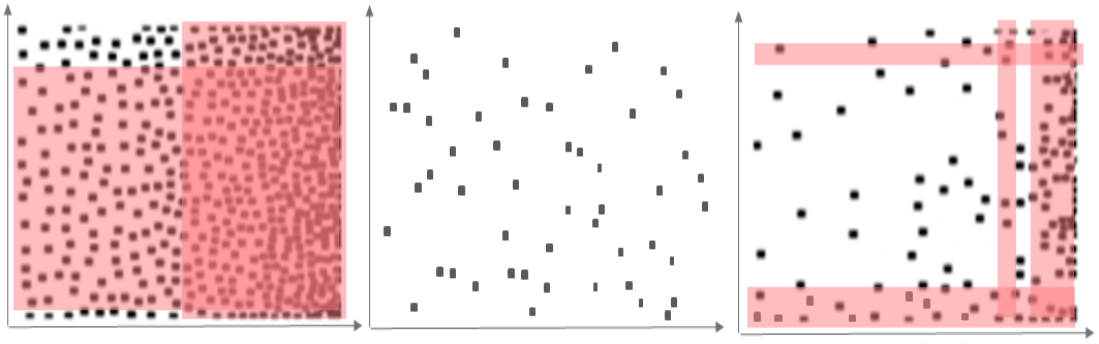
수집된 데이터가 기계학습에 활용되기 위해서는 일관적이고 연속적인 데이터 축적이 필요하다. 단순히 데이터의 건수가 양적으로 많은 것보다, 적더라도 학습에 활용되는 자질들이 골고루 채워진 데이터셋이 더 중요하다.

[그림 1]은 이를 단적으로 보여주는 예로, X축은 시간의 흐름을, Y 축은 수집된 데이터의 자질 즉 종류를 나타내며 [그림 1.a]는 수집된 데이터가 양적, 질적으로 모두 밀도가 높은 경우를 표현한 것이다. 반면 양적으로는 [그림 1.b]와 [그림 1.c]가 거의 유사하게 적은 양이지만 [그림 1.c]의 경우 밀도가 높은 데이터가 다수 포진, 학습에 활용될 수 있는 데이터가 [그림 1.b]에 비해 많다. 즉 실제 기계학습에 활용되기 위해서는 데이터를 구성하는 주요 속성들이 충실이 채워져 있어 결락이 적고 꾸준한 주기와 기간 동안 일관되게 축적하는 것이 중요하다.

2.2 유효한 데이터

축적된 데이터가 아무리 많아도 그 값에 오류가 있거나 서로 다른 용어나 방식으로 축적되었다면 학습에 활용할 수 없거나, 학습시켜도 오류를 일으킨다. IoT 관측 데이터를 예로 들면, 서로 다른 단위나 주기로 측정된 값은 아무리 많이 축적되어 있어도 이를 기계학습에 활용할 수 없다[10]. 혹 학습한다 하더라도 오류를 야기하는 불순한 데이터, 즉 노이즈(noise)로 작용할 수 있다.

이와 같이 고품질의 학습데이터를 구축하기 위해서는 수집된 데이터의 품질을 검증하는 과정이 필요하다. 본 연구에서는 공공데이터포털을 선정, 포털에서 공개하고 있는 데이터를 활용해 예측 모델을 개발한 대표 사례를 중심으로 유효한 데이터의 선별 효과를 살펴보고자 한다.



a. High-density big data b. Low-density small-volume data c. High-density small-volume data

Fig. 1. Density of Data (ex, X: timeline, Y: Temperatures of Air)

3. 사례연구

3.1 공공데이터포털

공공데이터포털(www.data.go.kr)은 명실상부한 우리나라를 대표하는 데이터 창구로, 『공공데이터의 제공 및 이용 활성화에 관한 법률(제11956호)』에 따라 국가에서 보유하고 있는 다양한 데이터를 개방하여 국민들이 보다 쉽고 용이하게 공유·활용할 수 있게 하는 서비스이다. 2022년 1월 현재, 987개 기관으로부터 교육, 국토관리, 환경기상, 농축수산, 보건의료 등 16개 분야의 데이터를 공개하고 있다. 개방 데이터 유형으로는 파일데이터 50,599건을 비롯해 표준데이터셋 8,751건, 오픈API 형태로 제공하고 있는 데이터세트로 8,637건을 제공하고 있다[11].

특히 공공데이터포털에서는 국민과 기업 등 수요 중심으로 개방의 효과성, 시급성 등이 높은 분야를 중심으로 '국가중점데이터'를 선정하여, 민간에서 활용하기 용이한 형태로 정제, 가공하여 개방된 양질의 대용량 데이터를 제공하고 있다. 본 연구에서는 이러한 데이터를 활용해 실제 기계학습에 활용해 실생활에서 발생할 수 있는 다양한 재난을 예측하는 모델을 개발, 사례연구를 수행하였다.

3.2 사례연구1: 국토 교통 분야 예측 모델

첫 번째 사례연구로, 본 연구진은 공공데이터포털에서 제공하고 있는 국토 교통 분야 빅데이터를 활용해 도로상에 블랙아이스 발생 구간을 예측하는 모델을 개발하였다[11,12]. 특히 고속도로 블랙아이스 경우, 치사율이 다른 교통사고에 의해 매우 높아 이를 사전에 예측하여 해당 노선의 운전자에게 미리 주의 경보(alert)를 알린다면 사고율을 매우 낮출 수 있다. 블랙아이스 발생 예측을 위해서는 도로 정보 뿐 아니라 교통정보, 기상정보 등 다양한 분야의 정보를 연계해 분석해야 한다.

[표 1]은 본 연구진이 블랙아이스 발생 예측 모델을 개발하기 위해 공공데이터포털을 통해 수집한 데이터이다. 표에서 보이듯이 4개의 세부 분야의 11종의 데이터셋, 총 12,574,630건의 데이터를 수집하였다. 그러나 오류가 있는 데이터를 제거하고 유효한 데이터를 선별, 보정하는 전처리를 수행하고 실제 학습에 효과가 있는 데이터만을 정제한 결과, 전체 수집 데이터 중 겨우 4.8%인 604,000건만이 모델개발에 활용되었다. 자세한 유효 데이터 선별과정은 4장에서 다루기로 한다.

3.3 사례연구2: 스마트팜 분야

본 논문의 두 번째 사례는 스마트팜 분야로, 공공데이터포털에서 제공하고 있는 농축산 및 환경기상, 국토관리 등의 다양한 분야의 데이터를 활용해 농장별 병해충 발생 예측 모델 연구이다[13,14].

농업은 쌀을 주식으로 하는 우리나라에게 매우 중요한 기간산업이자 생명산업으로, 최근 기후 및 농업환경 변화, 국제 교역 확대로 외래병해충 증가 등 병해충으로 인한 농작물 손실이 날로 증가하고 있다. 이를 위해 기존에 '진단' 위주의 연구가 아닌 '예측'에 방점을 두어, 기상정보 뿐만 아니라 전국 농경지 정보 및 토양, 병해충 정보 등 다양한 7천9백만 데이터를 수집·정제하여 약 370만 데이터를 활용하였다. 특히 수집된 요인들의 상관관계를 분석하여 병해충 예측에 효과적인 자질만 선별하여 병해충 발생 여부 및 발생 병해충을 예측하였다. [표 2]는 공공데이터포털을 통해 수집한 데이터 통계로, [사례연구 1]과 유사하게 수집 대비 4.8% 정도만 실제 학습에 활용되었음을 나타낸다.

〈Table 1〉 Case 1: Transportation Domain Public Data

Data		# of items
domain	Information	
Location	Highway Route Data	45,954
	Elevation Data of Highway Route	1068
Road	Highway Pavement Data	53,846
	Grooving Data	2,295
	Snow Damage Vulnerable Section of Highway	34
Traffic	VDS Installation Data	6,303
	VDS Traffic Data	11,680,000
	VMS Installation Data	1,470
	VMS Message Data	500,000
Weather	Monthly Weather Data	57,381
	VDS Weather Data	226,279
Total Original Collected Data		12,574,630
Used data for Machine Learning after Validation		604,000 (4.80%)

〈Table 2〉 Case 2: SmartFarm Domain Public Data

Data		# of items
domain	Information	
Agricultural Weather	8 special cities, 9 provinces	11,987,461
	synoptic weather	12,294,903
	Hourly	11,059,877
	Extreme_Time	23,559,880
	Meteorology_time unit	7,412,461
	Daily	472,980
	Extreme_day unit	472,979
	Synopsis Weather_Daily	307,819
Soil Test	Farm Map_Soil Test_Relationship	2,900,847
	Soil Test_Chemical Analysis	2,834,354
Pest	Convergence_Pest	2,435,277
	Pests_Investigation Point	8,567
	Pest_Inquiry	3,341,286
Pesticide	Pesticide information	39,983
Crop	crop information	2,558
Total Original Collected Data		79,131,232
Used data for Machine Learning after Validation		3,746,403 (4.73%)

4. 기계학습에 유효한 데이터 선별

이번 장에서는 상기 사례연구들을 통해 실제 수집은 되었으나 기계학습 모델에 활용되지 못한 데이터 즉 '저품질' 데이터 유형을 규명하고 이들 중 유효한 데이터를 선별하는 방안을 제안한다. 사례연구를 통해 데이터 품질을 저하시키는 요인을 살펴본 결과 다음 3가지 유형으로 정의할 수 있었다.

4.1 저품질 데이터 유형

1) 데이터 의미 불일치

사례연구에서 나타나듯이 공공데이터포털 내에는 같은 도메인이라 하더라도 매우 다양한 기관들로부터 제공된, 혹은 수집된 데이터를 공개하고 있다. 그러나 각 기관별로 서로 다른 용어와 분류체계를 활용하고 있어 같은 의미의 데이터임에도 이를 식별하기가 어려운 경우가 다수 발견되었다. 예를 들면 '도로유실' 관련 재난의 경우, A 기관에서는 [사회재난〉국가기반체계 마비] 의 분류가 할당된 반면, B 기관에서는 도로가 유실된 원인이 지진이나 산사태로 발생된 경우 [자연재난〉복합재난]으로, 화재나 싱크홀 등으로 발생된 경우에는 [사회재난]으로 분류되는 등 같은 데이터에 대한 메타데이터 항목이 서로 상이하게 부여되어 있다.

데이터를 지칭하는 명칭 간의 오류도 매우 큰 것으로 파악되었다. 기관별로 서로 다른 용어사전(word dictionary)를 차용하고 있어 같은 의미의 데이터 임에도 DB 내 서로 다른 필드(field)명으로 정의되어 있어, 실제 데이터가 제대로 구축되었음에도 활용되지 못하는 경우가 있었다. 예를 들어 위치정보 중 '경도' 좌표를 의미하는 항목으로 'CORD_SLATITUDE'를 비롯해 'CORD_LON', 'COORD_X' 또한 단순히 'LAT'라는 필드명으로 제공되고 있었다.

이러한 유형은 많은 양의 데이터가 수집되어 있음에도 실제 학습에는 활용하지 못하거나, 서로 다른 자질로 인식되어 결과적으로 혼동을 일으키게 한다. 데이터 의미 불일치를 해소하기 위해서는 사전에 DB 스키마 정의서를 입수, 각 필드명 용어 사전과 분류체계의 의미 등을 파악하여 공통된 체계를 정의한 후, 관련 데이터들을 매핑한 후 활용해야 한다.

2) 데이터 오류

사례연구를 통해 수집된 데이터를 검수한 결과, 다양한 유형의 오류를 발견하였다. 이들 데이터가 실제 학습에 활용되는 경우, 혼돈을 야기하여 성능을 저하시키는 역할을 하는 것으로, 사전에 이러한 유형을 발견, 일관되고 의미있는 값으로 보정하거나 정제해야한다.

예를들면 범주형 변수는 정수형으로, 수치형 변수는 단위를 통일하고자 정규화 방법 등이 있다. 대표적인 오류 유형은 다음과 같다.

① 값의 오류

이 유형은 유효하지 않은 데이터타입(data type)이나 이상치 값이 입력된 경우이다. 예를 들면 '지면온도'라는 데이터 항목에는 수치형(numerical data) 값만 올 수 있는데 "얼음"이라는 문자열(string)이 입력된 경우라던가, "-100"이라는 값이 입력된 경우이다.

② 계상 방식 불일치

기관마다 서로 다른 방식으로 수치를 측정, 계상해서 발생하는 오류로, 예를 들면 같은 날짜에 전국에서 발생한 '교통사고 발생 건수' 항목의 값이 서로 다른 경우가 포착되었다. 원인을 파악한 결과, C라는 기관은 도로에서 발생한 교통사고 건수를 계상한 반면, D라는 기관은 도로 뿐 아니라 해상과 항공을 포함한 전체 교통사고 발생 이력을 축적하고 있었다.

③ 단위 불일치

같은 의미에 같은 방식으로 측정된 값이라도 값 자체의 단위나 의미가 불일치한 경우로, 가장 대표적으로는 온도 관련 항목 값에 섭씨(℃)와 화씨(℉)가 혼용된 경우, 위치 정보 좌표계를 국토지리원정보원 표준(GRS80), 지리원 표준(BESSEL) 등 서로 다른 표준을 기준으로 삼은 경우이다.

수치형 데이터외 문자열 데이터의 경우, 같은 값을 지칭하는데 서로 다른 용어를 사용하는 경우도 있었다. 농업 분야를 예로 들면, "수온이 낮은 물에 의해 농작물에 피해를 가져오는 기상재해"를 의미하는 용어로 E라는 기관에서는 "냉해"로, F라는 기관에서는 "냉수장애"로 입력하고 있었다.

3) 불완전한 데이터

사례연구에서 해결하고자 하는 블랙아이스 발생 이나 병해충 발병 예측과 같은 문제들은 단기간 내 수집된 데이터만으로 학습이 불가하다. 또한 단순히 도로정보나 농지정보 등, 단일 분야 데이터 활용을 넘어 이질(heterogeneous) 및 다종(diverse) 데이터를 연계하여 종합적으로 분석해야한다. 그러나 수집된 데이터를 통합해 분석한 결과, 많은 결락이 발생, 실제 기계학습에는 부족한 것으로 파악되었다. 대표적인 유형으로 값의 결락과 기간 불일치 등이 있었다.

① 값의 결락

만약 어떤 문제를 해결하기 위해 참조해야 할 자질이 50개라면, 학습을 위해서는 이 50개 필드가 모두 채워져 있어야 유효한 데이터이다. 그러나 사례연구를 통해 살펴본 결과, 수집 데이터의 80%가 대부분 이러한 결락에 의해 실제 학습에는 활용되지 못한 것으로 파악되었다.

② 수집 기간 불일치

자연재난과 같이 매년 반복되는 재해의 경우, 다년도의 시계열 분석이 매우 중요하다. 실제 [사례연구 2]의 경우에서도 병해충 발생 학습을 위해 과거 10년의 정보에 기반한 회귀분석이 필요했다. 그러나 기상정보는 8년, 병해충 발생 정보는 과거 3년 정도 밖에 축적되지 않았다. 또한 기간이 겹치는 구간에서도 어떤 정보는 일 단위로 관측된 반면, 어떤 정보는 월 단위 계측이여서 학습에 활용될 수 있는 데이터는 매우 소수에 한했다.

4.2 학습에 유효한 데이터 선별

상기한 저품질 데이터 유형으로 판별된 경우를 비롯해 다양한 전처리 과정[12,13]을 거쳐 정제된 데이터라 하더라도 실제 학습 결과에 도움을 주는 유효한 데이터 선별 과정이 필요하다. 최근에는 학습 모델 자체에 유효한 자질선정(feature selection) 과정이 통합된 딥러닝(deep learning) 기법[15]이 많이 공개되고 있으나, 애초 사전에 불필요한 자질을 제거하여 학습시키는 것은 모델의 정확도를 향상시킬 뿐 아니라 모델 구축에 필요로 한 시간과 자원을 절약할 수 있는 중요한 요소이다. 사전에 학습에 유효한 데이터 선별을 위한 방법으로는 결측값(missing value) 보정, 이산형 변수는 one-hot-encoding을 통해, 연속형 변수는 단위 통일과 수치 보정(반올림 등)을 통한 정규화 등이 있다.

[그림 2]는 〈표 1〉에 기술된 다양한 종류의 정보 중 실제 블랙아이스 발생 예측에 효과적인 자질만을 선별한 결과이다. 블랙아이스 발생 예측을 위한 요인별 상관관계(correlation) 분석 결과로, 전체 42개 자질 중 학습 성능에 긍정적 영향을 미치는 자질은 21개로 파악되었다.

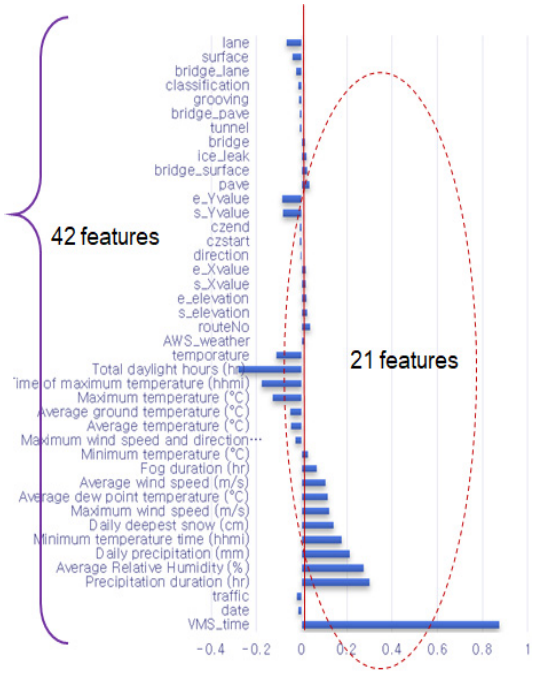


Fig. 2. The Effect of Valid data Discrimination

5. 결론

최근 대한민국의 가장 큰 화두는 '디지털 전환(digital transformation)'으로, 그 근간에는 '데이터'에 있다. 그러나 무작정 많이 모으고, 개방한다고 하여 실제 인공지능 기술에 활용될 수 있는 것은 아니다. 실제 기계학습에 적용할 수 있는 고품질의 데이터를 선별하여 관리하고 용이하게 활용할 수 있는 형식으로 제공하는 것이 중요한 시점이다.

본 연구는 실제 활용 가능한, 즉 유효한 데이터를 선별하는 방안을 제안하고 살제 공공빅데이터을 대상으로 한 학습 모델에 적용함으로써 그 효과를 증명하였다. 사례연구 결과, 공공데이터포털에서 수집한 데이터 중 약 5% 정도만이 실제 예측모델 구축에 유효한 것으로 판명, 데이터 품질 관리의 중요성을 피력한 데에 본 연구의 의의를 두고자 한다.

REFERENCES

[1] IDC. IDC Forecasts Improved Growth for global AI Market in 2021 [Internet], https://www.idc.com/getdoc.jsp?containerId=prUS47482321

[2] T.J.Kim, Data Dam', What Kind of Businesses Are They Made Up [Internet], https://zdnet.co.kr/view/?no=20200902101741

[3] K.V.Cruz, "Moon Jae-In's Strategy Amid Covid-19 Pandemic: Reviving the Green in the Korean New Deal." in *Collection of Essays on Korea's Public Diplomacy*, 2020

[4] D.Fang and L.Deng, "Legal Regulation of Government Data Opening: American Legislation and China's Path: Reflection Based on the US the Open, Public, Electronic, and Necessary (OPEN) Government Data Act," *Information and Documentation Services* Vol.42, No.5, pp.50-57, 2021

[5] D.J.Kim, "Spatial Big Data Plan for Government 3.0 and Creative Economy", *Korea Research Institute For Human Settlements*, No.14, pp.40-47, 2014

[6] G.Viscusi, B.Spahiu, A.Maurino, and C.Batini, "Compliance with open government data policies: An empirical assessment of Italian local public administrations." *Information polity* Vol.19, No.3, pp.263-275, 2014.

[7] Gartner Reserach. Measuring the Business Value of Data Quality [Internet], https://www.gartner.com/en/documents/1819214/measuring-the-business-value-of--data-quality

[8] S.O.Yun and J.W.Hyun, "An Analysis of Open Data Policy in Korea: Focused on National Core Data in Open Data Portal," Korean Public Management Review, Vol.33, No.1, pp.219-247, 2019

[9] W.S.Lim and S.J.Jung, Open Data, Small Amount. Useless Files [Internet], https://www.donga.com/news/article/all/20160517/78152584/1

[10] H.W.Lee, "Intrusion Artifact Acquisition Method based on IoT Botnet Malware," *Journal of KIOTS*, Vol.7, No.3, pp.1-8, 2021

[11] S.H.Yoon, J.H.Na, and H.-J.Oh, "Data Opening Status Analysis and Quality Management Strategies in Land, Infrastructure and Transport Domain,", *Journal of Digital Culture Archives,* Vol.3, No.2, pp.73-85, 2020

[12] J.H.Na, S.H.Yoon, and H.-J.Oh, "Black Ice Formation Prediction Model Based on Public Data in Land, Infrastructure and Transport Domain," *KIPS Transactions on Software and Data Engineering,* Vol.10, No.7, pp.257-262. 2021

[13] S.S.Yu, K.P.Choi, H.Myung, and H.-J.Oh, "Prediction Model of Pest According to Individual Farms Based on Heterogeneous Public Big data." *Journal of KIIT.* Vol.18, No.6, pp.1-9, 2020

[14] K.P.Choi, S.S.Yu, N.H.Yoo, and H.-J.Oh, "Pest Prediction and Prevention Model Visualization using Farm Map for Ecological Smart Farm," *Journal of KIIT.* Vol.19, No.2, pp.105-113, 2021

[15] H.W.Lee and H.S.Lee, "Optimal Machine Learning Model for Detecting Normal and Malicious Android Apps," *Journal of KIOTS*, Vol.6, No.2, pp.1-10, 2020

오 효 정(Hyo-Jung Oh) [정회원]

- 2008년 : 한국과학기술원 컴퓨터공학과(공학박사)
- 2000년 ~ 2015년 : 한국전자 통신연구원 지식마이닝연구실 책임연구원
- 2015년 ~ 현재 : 전북대학교 문헌정보학과 부교수

〈관심분야〉
정보검색, 텍스트마이닝, 빅데이터정보처리

윤 보 현(Bo-Hyun Yun) [정회원]

- 1999년 : 고려대학교 컴퓨터학과 이학박사
- 1999년 ~ 2002년 : 한국전자통신 연구원 선임연구원(팀장)
- 2003년 ~ 현재 : 목원대학교 SW교양학부 교수

〈관심분야〉
빅데이터 분석, 소프트웨어 교육, 정보검색