The Journal of Engineering Geology
pISSN : 1226-5268 eISSN : 2287-7169

Vol. 31, No. 1, March, 2021, pp. 31-42
https://doi.org/10.9720/kseg.2021.1.031

RESEARCH ARTICLE

기계학습을 활용한 도로비탈면관리시스템 데이터 품질강화에 관한 연구

이세혁[1] · 김승현[3] · 우용훈[1] · 문재필[2] · 양인철[3*]

[1]한국건설기술연구원 전임연구원, [2]한국건설기술연구원 수석연구원, [3]한국건설기술연구원 연구위원

The Study for Improvement of Data-Quality of Cut-Slope Management System Using Machine Learning

Se-Hyeok Lee[1] · Seung-Hyun Kim[3] · Yonghoon Woo[1] · Jae-Pil Moon[2] · Inchul Yang[3*]

[1]Research Specialist, Korea Institute of Civil Engineering and Building Technology
[2]Senior Researcher, Korea Institute of Civil Engineering and Building Technology
[3]Research Fellow, Korea Institute of Civil Engineering and Building Technology

Abstract

Database of Cut-slope management system (CSMS) has been constructed based on investigations of all slopes on the roads of the whole country. The investigation data is documented by human, so it is inevitable to avoid human-error such as missing-data and incorrect entering data into computer. The goal of this paper is constructing a prediction model based on several machine-learning algorithms to solve those imperfection problems of the CSMS data. First of all, the character-type data in CSMS data must be transformed to numeric data. After then, two algorithms, i.g., multinomial logistic regression and deep-neural-network (DNN), are performed, and those prediction models from two algorithms are compared. Finally, it is identified that the accuracy of DNN-model is better than logistic model, and the DNN-model will be utilized to improve data-quality.

Keywords: cut-slope management system (CSMS), missing-data, incorrect input, machine-learning, multinomial logistic regression, deep-neural-network

초 록

도로비탈면관리시스템(Cut-Slope Management System, CSMS)은 전국 일반국도 비탈면에 대해 기초·정밀 조사를 바탕으로 데이터베이스를 구축해왔다. 그런데 이러한 데이터는 사람에 의해 기록되기 때문에 데이터 누락 및 오기입 문제가 발생할 수밖에 없다. 본 연구에서는 데이터의 불완전성 문제를 극복하기 위해 여러 머신러닝 기반의 예측모델들을 개발하고 이를 이용한 데이터 품질 강화 가능성을 검토하고자 하였다. 우선 다 범주 문자형 데이터를 수치화하는 과정을 수행하였고, 선정된 데이터 항목들에 대해 다항 로지스틱 회귀분석(Multinomial Logistic Regression)과 심층신경망(Deep-Neural-Network) 기반의 예측모델들을 개발하였다. 그 결과, 심층신경망 모델들의 정확도가 월등히 높은 것으로 나타났다. 향후 개발된 모델들을 활용하여 누락 및 오기입 데이터의 보완이 가능할 것으로 기대된다.

주요어: 도로비탈면관리시스템, 데이터 누락, 데이터 오기입, 머신러닝, 다항 로지스틱 회귀분석, 심층신경망

OPEN ACCESS

***Corresponding author:** Inchul Yang
E-mail: ywinter75@kict.re.kr

Received: 5 January, 2021
Revised: 22 January, 2021
Accepted: 25 January, 2021

Ⓒ 2021 The Korean Society of Engineering Geology

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

서 론

도로비탈면관리시스템(Cut-Slope Management System, CSMS)은 전국 도로 비탈면 현황을 파악하여 위험등급 산정·비탈면 유지대책 도출 등과 같이 사전에 비탈면 붕괴를 방지하기 위해 구축된 시스템으로, 이를 통해 국민의 안전도모를 목표로 한다(KICT, 2019). 이와 같은 CSMS는 국토교통부 위탁고시(제2015-91호)에 따라, 위탁기관들이 지역별 관할도로를 나누어 위탁운영·관리 업무를 수행하고 있으며, 매년 국도 주변 CSMS 데이터베이스를 갱신하여 효율적인 투자 우선순위 및 투명한 예산 집행 의사결정을 지원한다. CSMS는 국토관리사무소 혹은 지자체로부터 받은 기본정보(행정구역, 구간, 연장, 등)를 기반으로 현장에서 취득된 비탈면 정보인 기초조사 자료와, 기초조사 이후에 수행되는 정밀 조사로부터 선정된 위험비탈면에 대해 불연속면 특성, 풍화도, 암반강도, 누수 현황, 배수시설 등의 상세·정밀 자료로 구성되며, 이러한 자료를 통해 대책 공법 수립을 지원한다. 이러한 자료들은 수치형과 문자형 데이터로 구성되며, 기록과정에서 인간오차(Human Error)로 인해 누락 및 오기입이 발생할 수 있으며 전문가에 의해 작성되는 정밀조사 자료는 주관적인 인간오차 발생 가능성이 크다.

최근, CSMS 데이터의 문제 보완을 위해 로지스틱 회귀분석(Logistic Regression)을 적용한 연구가 수행되었다(Woo et al., 2020). 이 연구에서는 비탈면의 기본정보와 기초조사 자료로부터 객관적인 데이터(정보형 데이터)들을 독립변수로 선정하고, 주관적인 '조치', '계측추천'와 같은 판단형 데이터들을 종속변수로 가정하여 로지스틱 예측모델을 구축하였다. 구축된 예측모델은 95% 이상의 매우 높은 정확도로 종속변수 예측이 가능함을 보여주었다. Woo et al.(2020)에서 수행한 예측모델들이 높은 정확도를 보일 수 있었던 이유는 사용된 독립변수들(기초조사 정보)과 종속변수(기초조사를 바탕으로 결정되는 정밀조사 자료) 간에 명확한 인과관계가 있으며, 학습에 사용된 정보들이 일관적으로 기록되었기 때문이다. 그런데, 일반적인 로지스틱 모델은 이항(Binomial)으로 이루어진 종속변수에 대해서만 적용가능하며, 독립변수와 종속변수 간에 인과관계(Causal Relation)가 명확하지 않은 경우 낮은 정확도를 보이는 한계가 있다. 그런데 CSMS 데이터의 상당 부분을 차지하는 기초조사 자료는 다 범주 변수이며 이러한 변수들은 명확한 인과관계를 갖는다고 보기 어렵다. 따라서 세 클래스 이상으로 이루어진 종속변수와 명확한 인과관계가 없는 독립변수들 사용 시에 대해서는 다른 기계학습 방법이 필수적이다.

본 연구에서는 기본정보·기초조사를 바탕으로 전문가에 의해 결정되는 판단형 변수들에 대해서만 예측모델을 개발했던 사전연구와는 달리, 정보형 데이터 항목을 종속변수로 가정하는 경우에 대해서도 기계학습을 수행하고 누락된 정보를 예측할 수 있는 모델을 구축하고자 한다. 이때, 사용되는 정보형 데이터들은 대게 이항이 아니기 때문에, 기존 일반적인 로지스틱 회귀분석이 아닌 다항 로지스틱 회귀분석을 일차적으로 적용하여 한계점과 정확도를 파악하고, 그 후에 정확도를 높이기 위해 심층 인공신경망을 검토 및 적용하였다.

연구 배경

CSMS 데이터 개요

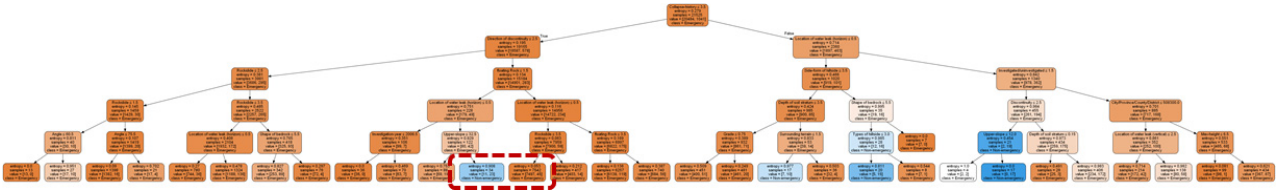
본 연구에서 사용된 데이터는 2006년부터 2019년까지 전국 일반국도 비탈면들에 대하여 기초조사와 정밀조사를 토대로 구축된 데이터이며, 수치형 데이터와 문자형 데이터로 구성되어 있다. 기초사항·기초조사 자료와 관련된 데이터들은 객관적인 '정보형 데이터'들인 반면에, 두 정보를 바탕으로 조사 수행자에 의해 작성되는 정밀조사 자료는 대게 주관적인 '판단형 데이터'들로, 수행자의 경험과 전문성에 큰 영향을 받는다. 정보형 데이터의 예로는 시/도/군/구, 왕복/편도, 차선수, 조사년도, 조사/미조사, 길이, 최대높이, 각도, 구배, 상부경사, 이격거리, 소단개소, 사면종류, 주변지형, 지하수, 누수

위치 세로/가로, 풍화도, 불연속면 방향성, 사면형상, 측면형상, 계곡부, 붕괴이력, 뜬돌, 낙석, 암종, 토층심도, 암반형태, 불연속면 등이 있으며, 판단형 데이터로는 계측추천, 위험도, 피해도, 붕괴유형, 등급재조정, 필요 주공법 종류, 조치 등이 있다(KICT, 2019).

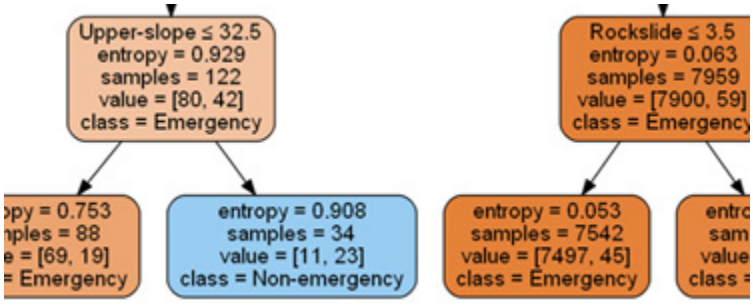
데이터 전처리

기계학습을 수행하기 위해서 데이터 전처리는 필수적이다. 소개된 CSMS 데이터는 30,751개의 조사 수행 건에 대한 73개 항목 정보를 가지고 있다. 본 연구에서는 CSMS 전문가의 의견을 반영하여, 객관적인 정보형 데이터로는 앞서 언급한 29개의 항목을, 판단형 데이터로는 조치, 위험도, 위험등급, 등급재조정의 4개 항목을 연구대상 항목으로 선정하였다. 선정된 항목 중, 문자로 이루어진 정보들은 모두 숫자로 치환되어야 한다. 예를 들어, 주변지형 항목은 구릉, 산악, 준산악, 평지와 같이 4개의 클래스가 있으며, 이는 각각 0, 1, 2, 3으로 치환될 수 있다(Woo et al., 2020). 모든 문자형 클래스들에 대해 0부터 시작하여 각 클래스 수만큼 숫자로 치환한 후, 안정적인 기계학습 모델 피팅(Fitting)을 위해서는, 스케일링(Scaling)이 필수적으로 수행되어야 한다. 다음에 소개되는 기계학습 수행에서는 여러 스케일링 방법 중에서 평균과 표준편차를 이용하여 정규분포로 치환하는 스케일러가 사용되었다.

기계학습 수행 시 많은 경우 데이터 자체 오류로 인해 수치 문제가 발생하게 된다. 이러한 데이터 자체 오류 문제를 본 연구에서는 정보 누락에 의한 문제와 오기에 의한 문제로 분류하였다. 두 문제 모두 인간오차(Human Error)에 의해 발생하게 되는데, 첫 번째 문제는 정보 재취득 수행 혹은 기계학습 기반 확률모델을 구축한 후 누락 정보 예측을 통해 해결이 가능하다. 그러나 정보를 재취득하기 위해서는 현장 재조사가 수행되어야 하는 어려움이 있으며, 확률모델을 이용한 예측은 기계학습 수행을 필요로 한다. 여기서, 기계학습을 수행하기 위해서는 두 번째 오기입 문제를 해결해야만 가능하다. 오기입 사례로는 '절리'가 '졀리'로 기입된 경우, '1'이 '11'과 같이 수치형 자료가 잘못 입력된 경우 등이 있다(Woo et al., 2020). 문자형 오기입은 번거롭지만 여러 프로그램을 활용하여 수동적 수정이 가능하겠지만, 수치 오기입의 경우 파악조차 쉽지 않을 수 있다. 이를 위해 의사결정나무(Kim et al., 2007; Cho et al. 2019), 에이다부스트(Adaboost)와 같은 여러 비지도학습(Unsupervised-Learning)들이 활용될 수 있을 것으로 예상되며, Fig. 1은 의사결정나무를 이용한 특이치 파악



(a) Full-diagram



(b) Part for the rare event

Fig. 1. Example of decision-tree for identification of a rare event: the blue-colored boxes mean a rare event.

사례이다. 분석 결과에서, 주황색으로 표시된 박스들(혹은 데이터 그룹)과는 달리, 몇몇 부분에서 파란 그룹들이 있음을 확인할 수 있고 해당 데이터들은 다른 데이터들과는 다른 특성을 가지는 데이터임을 의미한다. 이와 같이 파악된 특이성을 지닌 그룹을 검토한다면 오기입 문제들을 사전에 발견 할 수 있을 것이다.

로지스틱 회귀분석(Logistic Regression)

Woo et al.(2020)에서 사용된 로지스틱 회귀모델(Logistic Regression)은 간단한 기계학습 혹은 분류 기법으로, 0과 1처럼 이항(Binomial)으로 이루어진 종속변수를 선정하고, 선정된 종속변수와 독립변수들 간의 선형관계를 Logit 변환을 이용하여 확률모델로 구축하는 방법이다(Baek et al., 2016).

$$\mathrm{odds}=\frac{p(y=1|x)}{1-p(y=1|x)} \tag{1}$$

$$\mathrm{logit}(p)=\log\frac{p}{1-p} \tag{2}$$

$$y=\frac{\exp\left(\sum b_0+b_i x_i\right)}{1+\exp\left(\sum b_0+b_i x_i\right)} \tag{3}$$

식 (1)에서 $p(y=1|x)$은 조건부 변수에 대해 1이라는 사건이 발생할 조건부 확률을 의미하며, 로지스틱 모델은 이와 같은 조건부 확률을 기반으로 성공할 확률과 실패할 확률의 비를 나타내는 오즈(Odds) 개념을 이용한다. 식 (2)와 같이 로짓변환($\mathrm{logit}(\bullet)$)은 종속변수(y)와 독립변수(x)들 간에 선형관계가 있음을 정의하며, 최종적으로 산정된 계수 b_i를 통해 식 (3)과 같이 도출된다. 여기서, 오즈·로짓변환은 0과 1사이의 확률을 음의 무한대에서 양의 무한대로 범위를 변환시킴으로써, 이항으로 표현되는 종속변수와 독립변수들의 선형결합을 수식으로 표현할 수 있게 된다. 최종적으로 구축된 확률모델은 독립변수 정보를 가지고 종속변수에 해당하는 사건(Event)을 예측한다.

선행연구에서는 CSMS 데이터 개요에서 언급된 29개의 독립변수들을 이용하여 '조치(Action)'와 '계측 추천(Measurement Recommendation)'이라는 두 종속변수에 대해 로지스틱 모델을 구축하고 정확도를 산정하였다. Table 1은 조치 변수에 대한 로지스틱 수행 결과이며, 97.70%의 높은 정확도가 산정되었다.

하지만, Woo et al.(2020)에서 사용된 로지스틱 회귀모델(Logistic Regression)은 일종의 분류 기법으로, 0과 1로 표현될 수 있는 이항 변수에 대해서만 적용 가능하다는 한계점이 있다. 본 연구의 목적은 0과 1로만 표현되는 이항변수뿐만 아니라 다 범주 변수, 즉 다항확률변수에 대해서도 확률모델을 구축하고 누락/오기입 데이터를 예측 가능성 검토를 목표로 한다. 하지만 대부분의 객관적인 정보형 데이터들은 다중 클래스(예: 주변지형 변수는 구릉, 산악, 준산악, 평지와 같이 4개의 클래스가 있음)로 표현되어 있기 때문에, 이 경우에도 예측모델 구축이 가능한 다항 로지스틱 회귀모델(Multinomial Logistic Regression)을 일차적으로 개발하고 그 정확도를 검토한다.

Table 1. Logistic regression example about the action variable based on 29 random variables (Woo et al., 2020)

	Coef.	Confidence interval		z	$P>\|z\|$
		2.5%	97.5%		
Constant	-3.8290	-3.921	-3.737	-81.432	0.000
Province/City/Town/District	-0.0732	-0.144	-0.002	-2.029	0.042
Round/One-way	0.0147	-0.050	0.079	0.448	0.654
No. of way	-0.0762	-0.161	0.008	-1.768	0.077
Date	-0.1290	-0.214	-0.044	-2.977	0.003
Survey/Non-survey	-0.3116	-0.365	-0.258	-11.44	0.000
Slope length	0.1055	0.040	0.171	3.152	0.002
Max. height	0.1428	0.074	0.212	4.074	0.000
Angle	0.0734	-0.020	0.166	1.545	0.122
Gradient	-0.3170	-0.488	-0.146	-3.640	0.000
Gradient of upper-slope	0.1235	0.070	0.177	4.560	0.000
Distance for rockfall	-0.3674	-0.454	-0.281	-8.341	0.000
Berm	-0.1565	-0.243	-0.070	-3.557	0.000
Slope material	-0.2069	-0.285	-0.129	-5.199	0.000
Topography	-0.0821	-0.152	-0.013	-2.315	0.021
Ground water	0.1473	0.093	0.202	5.320	0.000
Loc. of leaking (top-bottom)	0.3327	0.236	0.430	6.710	0.000
Loc. of leaking (left-right)	0.0595	-0.035	0.155	1.229	0.219
Weathering grade	0.1519	0.062	0.241	3.328	0.001
Discontinuities' type	-0.3712	-0.441	-0.301	-10.405	0.000
Slope-shape	0.0116	-0.044	0.068	0.406	0.684
Slide-shape	-0.0160	-0.077	0.045	-0.517	0.605
Valley in slope	0.0164	-0.037	0.070	0.600	0.549
Collapse record	0.5877	0.537	0.638	22.749	0.000
Floating rock	-0.1125	-0.175	-0.050	-3.524	0.000
Rock fall	0.4388	0.368	0.510	12.124	0.000
Lithology	-0.0388	-0.099	0.022	-1.257	0.209
Soil depth	0.2055	0.147	0.264	6.910	0.000
Bedrock-shape	-0.1985	-0.256	-0.141	-6.818	0.000
Type of discontinuities	-0.0944	-0.173	-0.016	-2.368	0.018

다항 로지스틱 회귀분석을 이용한 예측 모델 구축

다항 로지스틱 회귀분석(Multinomial Logistic Regression)

다항 로지스틱 회귀 분석을 이용하면 K개의 클래스를 가진 종속변수 Y와 이를 설명하는 N개의 독립변수에 대한 예측 모델 구축이 가능하다(Jung et al., 2020). 다중 로지스틱 회귀모형은 다음과 같이 간단한 방법으로 유도할 수 있다. 종속변수 Y가 택하는 범주의 집합을 $K=\{1,\ldots,K\}$로 나타내고, 독립변수 $\mathbf{x}=\{x_1, x_2, \ldots x_N\}$가 택하는 값의 집합은 $\mathbb{R}^N$의 부분집합인 X로 나타내기로 한다. 독립변수의 확률변수를 $\mathbf{X}$라고 한다면, 독립변수와 종속변수는 확률변수 쌍 ($\mathbf{X}$, Y)를

구성하게 된다. 각 부분 집합들이 $x \in X$이고 $k \in K$일 때, $P(Y = K|\mathbf{X} = x)$는 양수이며 이 확률에 대해 종속변수의 다른 클래스를 오즈와 로짓변환 개념을 사용하면 다음과 같이 표현 가능하다(Sim and Kang, 2014).

$$\theta(k|x) = \log \frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y = K|\mathbf{X} = \mathbf{x})} (= \beta_{k0} + \sum_{i=1}^{N} \beta_{ki} x_i) \tag{4}$$

$$\exp \theta(k|x) = \frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y = K|\mathbf{X} = \mathbf{x})} \tag{5}$$

식 (4)에서 양변을 모든 k에 대해 합하면 $P(Y = K|\mathbf{X} = \mathbf{x}) = 1 / \sum_{k=1}^{K} \exp \theta(k|x)$임을 알 수 있으며, 따라서 주어진 독립변수에서 종속변수의 조건부 확률은 최종적으로 다음과 같이 표현된다.

$$P(Y = k|\mathbf{X} = \mathbf{x}) = \frac{\exp \theta(k|x)}{\exp \theta(1|x) + \cdots + \exp \theta(K|x)} \tag{6}$$

식 (6)을 다항 로지스틱 회귀모형이라고 하며, $K = 2$인 경우 식 (3)과 같이 일반 로지스틱 회귀모형이 된다.

다항 로지스틱 회귀분석 적용 및 예측모델 도출

앞서 언급한 29개의 독립변수 중 4개의 클래스를 갖는 주변지형(Topography)을 종속변수로 선정하고 그 외 28개 변수는 그대로 독립변수로 가정하여 다중 로지스틱 회귀분석을 수행하였다. 연구배경에서 설명한대로 데이터 전처리를 수행하였고, 교차 검증(Cross-Validation)을 위해 학습데이터와 검증 데이터의 비율을 19:1 가정, 즉 전체 데이터의 5%를 검증 데이터로 사용하였다. 다항 로지스틱 회귀분석 수행하기 위해 파이썬(Python)의 'sklearn' 모듈을 사용하였고, 다항 로지스틱 모델을 구축하기 위해 'Newton-cg' 알고리즘을 사용하였다. 수행 결과, 도출된 확률모델은 56.30%의 다소 낮은 정확도가 산출되었다. 결과를 더욱 살펴보기 위해 도출된 모델의 계수들과 각 계수들에 대한 신뢰도를 파악할 수 있는 유의확률(*p*-value)을 계산하였고, Table 2에 나타내었다.

Table 2를 보면 상당히 많은 변수를 최적화를 통해 도출해야함을 알 수 있으며, 다소 많은 계수들의 유의확률이 상당히 크게 계산됨을 확인하였다. 유의확률이 0.05보다 크게 계산된 값들과 해당 독립변수들의 계수들을 굵은 글씨와 흐린 배경으로 표시하였다. 예를 들어, '길이' 독립변수의 Class 3과 Class 4의 유의 확률은 각각 0.857과 0.097이다. 여기서 주목해야할 부분은 Class 3과 Class 4의 계수가 Class 1과 Class 2의 계수보다 절대적 크기가 작으며, 이는 종속변수에 대한 영향이 작음을 의미한다. 이 경우, '길이'독립변수를 제거하고 다시 로지스틱 회귀분석을 수행하여도 정확도 개선 측면에서 큰 영향이 없을 것으로 예상할 수 있다. 이를 확인하기 위해, 유의확률이 큰 계수들을 포함하는 조사/미조사, 길이, 각도, 이격거리, 누수위치 세로, 누수위치 가로, 풍화도, 사면형상, 계곡부, 암반형태, 불연속면 등 11개의 변수를 제거하고 다시 다항 로지스틱 회귀분석을 수행하였다. 그 결과, 56.82%의 정확도로 미세하게 개선되었으나 큰 영향이 없음을 확인하였다.

Table 2. Multinomial logistic regression model for topography having 4 classes

	Coef.				z				$P>\|z\|$			
	Class 1	Class 2	Class 3	Class 4	Class 1	Class 2	Class 3	Class 4	Class 1	Class 2	Class 3	Class 4
Constant	0.388	-0.301	1.098	-1.185	59.974	-46.62	169.871	-183.22	0	0	0	0
Province/City/Town/District	-0.174	0.191	0.335	-0.353	-23.401	25.75	45.112	-47.461	0	0	0	0
Round/One-way	0.103	-0.103	-0.020	0.024	15.391	-15.517	-3.456	3.582	0	0	0.001	0
No. of way	0.382	-0.557	-0.066	0.240	49.328	-71.862	-8.498	31.032	0	0	0	0
Date	0.379	-0.847	0.154	0.315	49.013	-109.66	19.922	40.725	0	0	0	0
Survey/Non-survey	0.134	-0.115	**-0.003**	-0.016	19.338	-16.649	-0.389	-2.300	0	0	**0.697**	0.021
Slope length	0.057	-0.068	**-0.001**	0.013	7.372	-8.852	-0.181	1.660	0	0	**0.857**	0.097
Max. height	-0.071	0.215	0.125	-0.269	-8.096	24.45	14.218	-30.572	0	0	0	0
Angle	-0.109	**0.017**	0.046	0.046	-11.19	1.758	4.731	4.700	0	**0.079**	0	0
Gradient	0.033	-0.100	0.031	0.035	3.773	-11.403	3.581	4.048	0	0	0	0
Gradient of upper-slope	-0.071	0.070	-0.014	0.015	-10.065	9.94	-2.014	2.139	0	0	0.044	0.032
Distance for rockfall	**-0.008**	-0.020	-0.055	0.084	-1.251	-2.998	-8.279	12.529	**0.211**	0.003	0	0
Berm	0.048	0.089	0.042	-0.179	5.345	10.019	4.763	-20.127	0	0	0	0
Slope material	-0.018	0.061	0.066	-0.109	-2.234	7.627	8.302	-13.695	0.026	0	0	0
Ground water	-0.031	-0.191	**0.009**	0.212	-3.599	-22.246	1.069	24.777	0	0	**0.285**	0
Loc. of leaking (top-bottom)	**-0.005**	0.452	0.046	-0.493	-0.407	35.752	3.673	-39.017	**0.684**	0	0	0
Loc. of leaking (left-right)	0.242	-0.128	-0.108	**-0.013**	19.601	-9.848	-8.74	-1.013	0	0	0	**0.311**
Weathering grade	**-0.011**	**-0.011**	-0.139	0.161	-1.279	-1.188	-15.547	18.014	**0.201**	**0.235**	0	0
Discontinuities' type	-0.123	0.020	0.139	-0.036	-16.76	2.699	18.896	-4.834	0	0.007	0	0
Slope-shape	-0.010	0.037	-0.075	0.049	-1.517	5.540	-11.398	7.375	0.129	0	0	0
Slide-shape	-0.064	0.053	-0.041	0.052	-8.809	7.278	-5.546	7.077	0	0	0	0
Valley in slope	-0.087	**0.000**	-0.028	0.114	-12.874	0.033	-4.166	17.006	0	**0.973**	0	0
Collapse record	-0.072	0.162	0.141	-0.231	-10.531	23.644	20.549	-33.662	0	0	0	0
Floating rock	0.025	0.030	0.087	-0.141	2.857	3.348	9.811	-16.016	0.004	0.001	0	0
Rock fall	-0.020	0.143	0.084	-0.207	-2.187	15.476	9.105	-22.393	0.029	0	0	0
Lithology	-0.056	0.244	-0.066	-0.122	-8.34	36.161	-9.807	-18.014	0	0	0	0
Soil depth	0.241	-0.348	-0.063	0.169	30.604	-44.182	-7.947	21.525	0	0	0	0
Bedrock-shape	0.063	0.086	**-0.006**	-0.143	9.144	12.513	-0.9	-20.757	0	0	**0.368**	0
Type of discontinuities	**-0.008**	0.067	-0.128	0.069	-0.937	8.179	-15.613	8.371	**0.349**	0	0	0

다른 경우에 대한 다항 로지스틱 예측모델의 정확도를 검토하기 위해, '주변지형' 변수 외에도 5개의 클래스를 갖는 '사면종류(반암, 암반, 자연, 토사, 혼합)', 6개의 클래스를 갖는 '풍화도(신선, 보통풍화, 약간풍화, 심한풍화, 완전풍화, 풍화잔류토)', 9개의 클래스를 갖는 '불연속면(균열, 단층, 암맥, 엽리, 전단대, 절리, 층리, 파쇄대, 없음)'을 선정하여 예측모델을 도출하고 정확도를 계산하였다. 그 결과, 순서대로 78.73%, 59.55%, 70.54%의 정확도가 계산되었다. '사면종류'와 '불연속면' 예측모델의 경우 다른 두 예측모델의 비해 상대적으로 높은 정확도가 계산되었지만, 선행연구에서 보였던 일반 로지스틱 회귀분석 적용 모델들과 비교하여 여전히 낮은 정확도이다. 다항 로지스틱 회귀분석의 정확도가 낮은 이유는 다중 클래스로 이루어진 종속변수의 클래스 수에 따라 로지스틱 모델 내 추정해야하는 계수들이 배로 증가하며, 이는 최적화 관점에서 한정된 정보를 이용하여 많은 변수를 추정해야 하는 고차원(High-Dimension) 문제가 되어 정확도가 떨어질

수밖에 없을 것으로 예상된다.

앞서 도출된 4개의 예측모델들은 독립변수와 종속변수 사이에 인과관계가 없는 경우이다. 추가적으로 다항 로지스틱 예측모델의 정확도 파악을 위해, 인과관계를 갖는 독립변수(기초조사: 정보형 데이터)와 종속변수(기초조사를 기반으로 전문가에 의해 결정되는 정밀조사: 판단형 데이터)에 대해서도 회귀분석을 수행하였다. 앞서 언급된 29개의 독립변수와 3개의 클래스를 갖는 '위험도', 5개의 클래스를 갖는 '위험등급'과 '등급재조정'을 종속변수로 설정하고 다항 로지스틱 모델을 도출하였다. 도출된 모델들의 정확도는 각각 58.39%, 60.40%, 47.07%로 계산되었다. 이는, 인과 관계를 갖는 독립변수, 종속변수 경우에도 다항 로지스틱 회귀분석 적용이 한계가 있음을 시사한다. 다음으로는 다항 로지스틱 예측모델의 낮은 정확도를 극복하기 위해, 심층신경망을 소개하고 같은 예제에 대해 적용하였다.

심층신경망을 이용한 예측 모델 구축

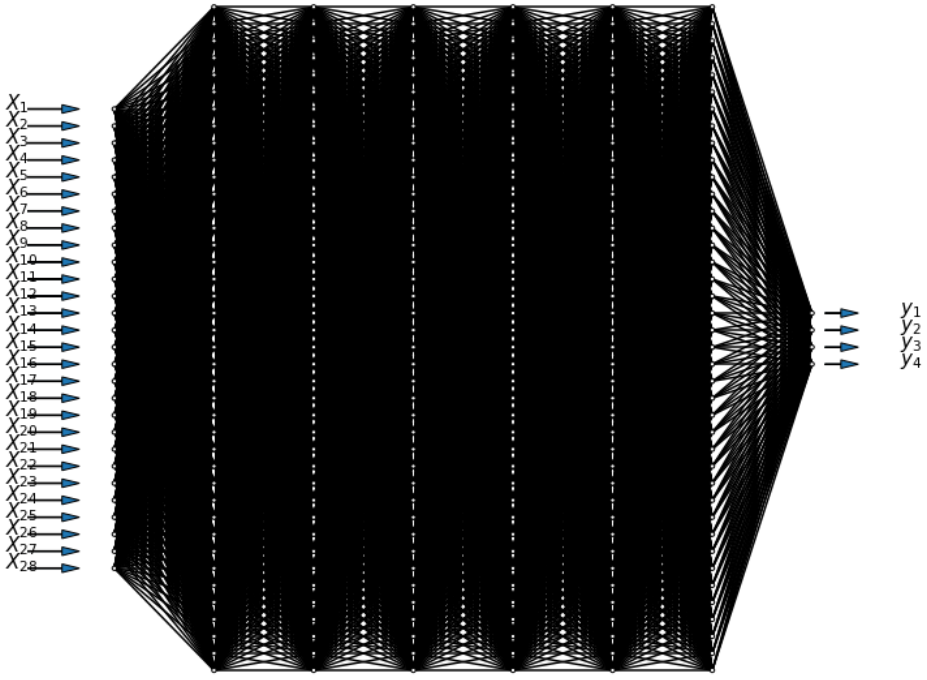
심층신경망(Deep-Neural-Network)

심층학습으로 불리는 딥러닝(Deep-Learning)은 다양한 비선형 변환기법 조합을 통해 방대한 양의 데이터가 내포하는 핵심적인 내용을 추출·추론하는 여러 기계학습 알고리즘의 집합을 의미한다. 기계학습은 어떠한 데이터가 주어졌을 때 컴퓨터가 이해할 수 있는 벡터의 형태로 표현하는 방법(예: 사진의 경우 수치로 나타내기 위해 열벡터로 표현) 개발과, 변환된 수치데이터를 가지고 최적화기법을 기반으로 어떻게 학습 모델을 개발할지로 나뉘게 된다. 대부분의 딥러닝 구조는 인공신경망(Artificial-Neural-Network, ANN)에 기반하며, 여러 은닉층(Hidden Layer)들을 가질 경우 심층신경망(Deep-Neural-Network)이라 불린다(Kim et al., 2010). 초기 심층신경망은 최적화를 위해 오류역전파(Backpropagation) 알고리즘을 사용하였지만(Kim, 2012), 지역 최솟값에 머무는 문제(Vanishing Gradient Problem)와 진동 혹은 발산 문제, 학습데이터에 너무 집중하는 과적합(Overfitting) 모델 도출 등과 같은 단점이 있었다. 또한, 기계학습 모델 구축을 위해 많은 시간이 소요되며 원론적 생물학적 신경망과 다른 문제점 등이 제기되었다.

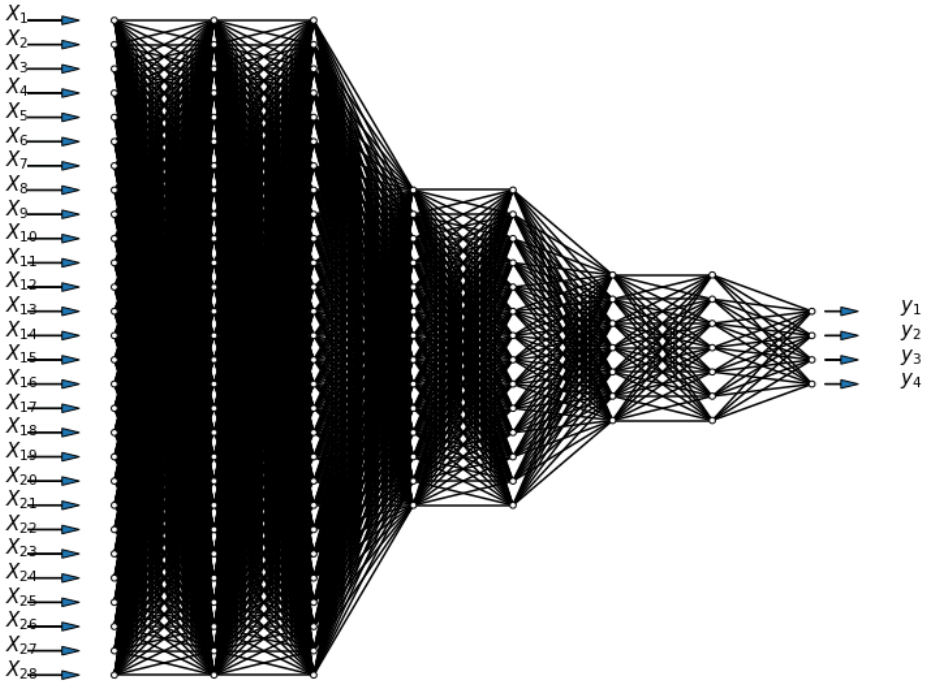
이러한 문제점들은 2000년대에 들어서면서, 비지도학습 기반 제한된 볼츠만 머신(Restricted Boltzmann Machine, RBM)을 기반으로 피드포워드(Feedforward) 신경망의 각 층(Layer)들을 효과적으로 사전훈련(Pre-training)하여 과적합을 방지하고, 지도학습 기반 오류역전파 알고리즘을 사용하는 체계를 통해 비약적으로 개선되었다. 이후에 Drop-out이라는 개념이 도입되면서, 앞선 RBM 기반 사전훈련보다 더욱 효율적으로 과적합을 방지하는 방법론이 일반적으로 사용되고 있다. 이와 같이 심층신경망을 효율적으로 학습시키는 최적화 방안들의 개발과 더불어 추가적으로 두 가지 요인으로 인해 인공신경망이 널리 사용되게 된다. 첫 번째는 하드웨어의 발전이며, 특히 GPU의 사용은 크기가 큰 행렬의 단순 계산을 비약적으로 단축시키는 계기가 된다. 두 번째는, 심층신경망 모델 개발을 위해서는 충분한 양의 '이름이 붙은(Labeled)' 데이터가 기반 되어야 하는데, 통신 기술과 컴퓨터의 발전으로 인해 여러 경로로부터 충분한 양의 데이터베이스가 구축됨에 따라 의미 있는 인공신경망 학습이 가능하게 되었다.

심층신경망은 입력층(Input Layer)과 출력층(Output Layer) 사이에 두 개 이상의 은닉층이 있을 경우를 말하며, 대게 추가된 층들은 점진적으로 적은 수의 노드를 가지는 형태로 구성된다(Kim et al., 2020a, 2020b). 이와 같이 구성된 심층신경망은 비선형 모델링이 가능하며, 최적의 모델 도출을 위해 오류역전파법과 경사하강법이 많이 사용되어 왔지만 이 경우 시간복잡도 및 과적합 문제가 존재한다. 따라서 이와 같은 문제들을 피하기 위해 여러 미니 배치(Mini Batch, 여러 학습 예제들로 나누어 하강법을 동시에 적용)를 구성하고 앞서 언급한 Drop-out 방법을 통해 학습속도 향상과 과적합을 피할 수 있게 된다. 본 연구에서는 파이썬의 텐서플로(Tensorflow) 모듈을 사용하였으며, 앞서 CSMS 데이터 개요에서 소개된

29개의 정보형 변수들 중 '주변지형'을 종속변수로 하고 그 외 28개 변수를 독립변수로 가정하는 예제에 대해 다음 Fig. 2와 같은 초기 모형들을 가정하여 초기학습을 진행하였다.



(a) Model 1



(b) Model 2

Fig. 2. Initial assumed models for deep-neural-network.

모델 1과 2 모두 입력층은 28개의 노드, 출력층은 4개의 노드로 구성되었으며 동일하게 6개의 은닉층이 가정되었다. 모델 1의 경우 모든 은닉층은 64개 노드로 구성되었으며, 모델 2는 점진적으로 노드 수가 적어지는 구조로 처음 2개 은닉층은 28개의 노드, 다음 3개 은닉층은 14개 노드, 그 다음 2개 은닉층은 7개의 노드로 구성되었다. 이와 같이 가정된 두 모델에 대하여 심층신경망 학습을 진행하였고, 그 결과 각각 83.47%와 82.44%의 정확도가 도출되었다. 대게 심층신경망은 점점 적어지는 노드 수를 갖는 은닉층으로 구성하는 것이 일반적이지만, 오히려 모델 1이 다소 더 정확하게 예측하는 결과가 도출되었다. 제안된 모델 1과 2보다 적절한 네트워크(예: 더욱 많은 은닉층을 갖는 신경망) 구성 시, 더욱 정확도 높은 모델을 구축할 수 있지만 그 경우 더욱 많은 학습시간이 요구되기 때문에, 본 연구에서는 모델 1을 선정하여 다른 학습 예제들에 대해서도 사용하였다. 모델 1의 정확도는 앞서 수행되었던 다항 로지스틱 회귀모델이 보인 정확도 56.30% 보다 매우 개선되었음을 알 수 있다. 다음 Table 3은 다른 예제들에 대해 학습을 수행한 결과이다.

Table 3. Results of learning deep-neural-networks about previously mentioned examples

	Number of independent variables (**X**)	Dependent variable (Informative/Decisional, Number of classes)	Accuracy of DNN	Accuracy of multinomial logistic regression
Case 1	28	Topography (Informative, 4)	83.47%	56.30%
Case 2	28	Slope material (Informative, 5)	93.65%	78.73%
Case 3	28	Weathering grade (Informative, 6)	90.62%	59.55%
Case 4	28	Type of discontinuities (Informative, 9)	94.69%	70.54%
Case 5	29	Action (Decisional, 2)	94.81%	96.40%
Case 6	29	Degree of risk (Decisional, 3)	84.74%	58.39%
Case 7	29	Risk-grade (Decisional, 5)	84.70%	60.40%
Case 8	29	Readjustment of grade (Decisional, 5)	84.99%	47.07%

Case 1~4는 29개의 정보형 변수들 중, 순서대로 '주변지형(Topography)', '사면종류(Slope Material)', '풍화도(Weathering Grade)', '불연속면(Types of Discontinuities)'이 종속변수로 설정되고 그 외 나머지 변수들은 독립변수로 가정된 경우들이다. 다항 로지스틱 회귀분석과 심층신경망 모두 학습 시 교차 검증 비율은 5%로 설정하였다. 도출된 정확도 결과들을 보면 심층신경망 모델들이 상대적으로 정확한 것을 확인할 수 있으며, 특히 '풍화도'가 종속변수인 경우를 보면 상당히 높은 정확도로 개선됨을 알 수 있다. Case 5~8은 판단형 변수인 '조치(Action)', '위험도(Degree of Risk)', '위험등급(Risk-Grade)', '등급재조정(Readjustment of Risk-Grade)'이 종속변수로 설정되고 29개의 정보형 변수가 독립변수인 경우들로, Case 5를 제외한 나머지 경우에는 심층신경망이 높은 정확도를 보였다. Case 5의 경우, 로지스틱 회귀모델의 정확도가 심층신경망에 비해 다소 높았으며, 이는 무조건적인 심층신경망 적용이 옳지 않음을 의미한다.

Case 1, 6~8의 경우 다항 로지스틱 회귀모델과 비교해서 상당히 개선되었지만, 80% 대의 다소 부족한 정확도가 계산되었다. 그 이유로는 두 가지로 생각된다. 첫 번째로, 최적의 심층신경망 도출을 위해서는 반복적으로 학습을 수행하고 정확도를 비교해가며 모델을 도출하는 시행착오(Trial and error) 방법이 필수적이다. 또한, 심층신경망 학습 시에 각 노드들의 가중치(Weight) 부여 방식에 따라 정확도가 다르게 도출됨이 알려져 있다(Kim et al., 2020a, 2020b). 따라서 더욱 다양한 심층신경망 형태와 노드마다 적절한 가중치를 부여한다면, 기존의 정확도보다 개선된 모델이 도출될 가능성이 있다.

두 번째로는 데이터 자체의 문제이다. 앞서 데이터 전처리에서도 언급하였지만, 드러나지 않는 오기입은 잘못된 정보를 제공하며 낮은 정확도를 보이는 예측모델 도출의 원인이 될 수 있다. 예를 들어, 주변지형(구릉, 산악, 준산악, 평지) 항목에서 '준산악'이 '산악'으로 잘못 입력된 경우가 있을 수 있다. 이 경우 수치 오기입과 마찬가지로 파악 자체가 쉽지 않다.

앞서 언급한 비지도학습 방법들을 활용하는 방안도 있지만 또 다른 대안으로는, Woo et al.(2020)에서 수행하였던, 기존 몇 개년도 데이터를 학습 데이터로 사용하여 모델을 구축한 후 차년도 데이터를 검증 데이터로 이용하여 정확도를 산출하여 그 값을 검토하는 방안이 있다. 이때 정확도가 높게 산정된다면 기존 데이터와 차년도 데이터는 매우 유사한 것으로 판단 할 수 있으며, 그렇지 않은 경우 기존과 다른 정보가 실제로 발생했거나 혹은 전문가 판단의 실수 혹은 정보 입력의 오류가 발생한 경우로 추측할 수 있다. 이러한 경우 해당 년도의 데이터를 제외하고 심층신경망을 학습시킨다면 정확도가 개선된 모델이 도출될 수 있을 것이다.

결 론

본 연구는 다항 로지스틱 회귀분석과 심층신경망을 이용하여 도로비탈면관리시스템(CSMS) 데이터 내 누락 및 오기입 데이터에 대한 보완 가능성을 검토하고자 수행되었다. CSMS 데이터는 비탈면의 최대 높이, 각도, 사면종류 등과 같은 객관적인 정보형 데이터와, 이를 토대로 현장에서 전문가에 의해 결정되는 조치, 위험도, 위험등급 등과 같은 주관적인 판단형 데이터로 이루어져 있다. 이러한 데이터들은 수치형 데이터와 더불어 범주형(문자형) 데이터가 포함되어 있기 때문에 기계학습 수행을 위해 전처리가 필수적이다. 선행연구에서는 정보형 데이터 항목들을 독립변수로, 판단형 데이터 중 특정 항목들을 종속변수로 가정하여 로지스틱 예측모델을 적용하고 이를 활용하여 데이터 보완뿐만 아니라 도로 비탈면 관리 업무 효율성 증대 가능성을 살펴보았다.

그런데 선행연구에서 사용된 일반 로지스틱 회귀분석은 종속변수가 0 혹은 1과 같은 이항으로 나뉘는 경우에만 적용 가능한 한계가 있다. 하지만, CSMS 데이터 내에 정보형 데이터가 판단형 데이터보다 더 많은 부분을 차지하며, 정보형 데이터 중 상당부분은 다 범주 변수들이다. 본 연구에서는 판단형 변수와 더불어 정보형 변수들의 보완 가능성 검토를 목표로 하였는데, 이 경우 종속변수가 세 개 이상의 범주를 지니게 되어 일반 로지스틱 모델 적용이 불가능하다. 따라서 다 범주 종속변수에 적용 가능한 다항 로지스틱 회귀분석을 검토하고 적용해보았지만, 도출된 예측모델은 상당히 낮은 정확도를 보였다. 그 이유로는 계수 추정 복잡도의 증가로 인해 최적화 관점에서 고차원 문제가 되었기 때문으로 예상된다. 다항 로지스틱 회귀모델의 낮은 정확도 문제로 인해, 심층신경망을 검토하고 동일한 문제들에 대해 기계학습 모델을 도출하였다. 그 결과 대부분 상당히 높은 정확도를 갖는 것으로 나타났다. 하지만, 종속변수가 다 범주가 아닌 이항으로 표현되는 경우는 여전히 기존 로지스틱 회귀모델이 정확함을 확인하였고, 따라서 종속변수의 범주 수에 따른 적절한 머신러닝 방법 선정이 필요함을 알 수 있다. 종속변수에 따른 적절히 선정된 방법을 통해 도출된 예측 모델들은 CSMS 데이터 내 누락 정보들을 보완할 수 있을 것으로 기대되며, 또한 선 수집된 객관적 데이터들만을 이용하여 아직 수집되지 못한 변수들을 사전에 예측함으로써, 의사결정이 중요한 조사 우선순위 결정 혹은 정책 수립 업무들의 효율을 높일 수 있을 것으로 기대된다.

사 사

본 연구는 국토교통부의 ‘2020년 도로비탈면관리시스템 운영 업무 대행’으로부터 지원 받아 수행되었습니다. 이에 감사드립니다.

References

Baek, S.A., Cho, K.H., Hwang, J.S., Jung, D.H., Park, J.W., Choi, B., Cha, D.S., 2016, Assessment of slope failures potential in forest roads using a logistic regression model, Journal of Korean Society of Forest Science, 105(4), 429-434 (in Korean with English abstract).

Cho, K., Lee, B.Y., Kwon, M., Kim, S., 2019, Air quality prediction using a deep neural network model, Journal of Korean Society for Atmospheric Environment, 35(2), 214-225 (in Korean with English abstract).

Jung, M., Kim, J.G., Uranchimeg, Sumiya., Kwon, H.H., 2020, The probabilistic estimation of inundation region using a multiple logistic regression analysis, Journal of Korea Water Resources Association, 53(2), 121-129 (in Korean with English abstract).

Kim, J.H., Shin, J.M., Park, J.J., Ha, T.J., 2010, Development of hazard-level forecasting model using combined method of genetic algorithm and artificial neural network at signalized intersections, Journal of The Korean Society of Civil Engineers, 30(4D), 351-360 (in Korean with English abstract).

Kim, M.J., 2012, Performance comparison of internal accounting control assessment models applying logistic regression and neural networks, Korea International Accounting Review, 46, 1-30 (in Korean with English abstract).

Kim, T., Song, J., Kwon, O.S., 2020a, Pre- and post-earthquake regional loss assessment using deep learning, Earthquake Engineering and Structural Dynamics, 49(7), 657-678.

Kim, T., Song, J., Kwon, O.S., 2020b, Probabilistic evaluation of seismic responses using deep learning method, Structural Safety, 84, 101913.

Kim, T.H., Shin, Y.S., Lee, U.K., Kang, K.I., 2007, Decision support model using a decision tree for formwork selection in tall building construction, Journal of the Architectural Institute of Korea Structure & Construction, 23(11), 177-184 (in Korean with English abstract).

KICT (Korea Institute of Civil Engineering and Building Technology), 2019, Operation of road cut slope management system in 2018, Republic of Korea Ministry of Land, Infrastructure and Transport, 355p.

Sim, S., Kang, H., 2014, A polychotomous regression model with tensor product splines and direct sums, Journal of the Korean Data & Information Science Society, 25(1), 19-26 (in Korean with English abstract).

Woo, Y., Kim, S.H., Yang, I., Lee, S.H., 2020, The Study for Utilizing Data of Cut-Slope Management system by Using Logistic Regression, The Journal of Engineering Geology, 30(4), 649-661 (in Korean with English abstract).