J Korean Soc Disaster Secur 14(3): 17-27, 2021
https://doi.org/10.21729/ksds.2021.14.3.17

Print ISSN: 2466-1147 Online ISSN: 2508-285X
www.dssms.org

KSDS

Original Article

기계학습기법을 이용한 땅밀림 위험등급 분류

이기하[1] · 레수안히엔[2] · 연민호[3] · 서준표[4] · 이창우[4]*

[1]경북대학교 미래과학기술융합학과 부교수
[2]경북대학교 재난대응전략연구소 박사후연구원
[3]경북대학교 미래과학기술융합학과 박사과정
[4]국립산림과학원 산불·산사태연구과 연구원

Classification of Soil Creep Hazard Class Using Machine Learning

Gi Ha Lee[1], Xuan-Hien Le[2], Min Ho Yeon[3], Jun Pyo Seo[4], and Chang Woo Lee[4]*

[1]Associate Professor, Dept. of Advanced Science and Technology Convergence, Kyungpook National University
[2]Post-doctoral Researcher, Disaster Prevention Emergency Management Institute, Kyungpook National University
[3]Ph.D Student, Dept. of Advanced Science and Technology Convergence, Kyungpook National University
[4]Researcher, Division of Forest Fire and Landslide, National Institute of Forest Science

요 약

본 연구에서는 6개의 기계학습 기법들을 활용하여 2019년과 2020년 전국 땅밀림 현장조사 결과를 기반으로 땅밀림 위험지역을 A부터 C까지 3개 등급(A등급: 위험, B등급: 보통, C등급: 양호)으로 구분할 수 있는 분류모형을 구축하고, 분류 정확도를 비교·분석한다. 기계학습 기법으로는 K-Nearest Neighbor, Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting 총 6개를 적용하였다. 분류 정확도 분석결과, 6개의 기법 모두 0.9 이상의 우수한 정확도를 보여주었다. 수치형 자료를 학습에 적용한 경우가, 문자형 자료를 학습한 모형보다 우수한 성능을 나타냈으며, 현장조사 평가점수 자료군(C1~C4) 보다는 전문가의견이 반영된 평가점수 자료군(R1~R4)으로 학습한 모형이 정확도가 높은 것으로 분석되었다. 특히, 직접징후와 간접징후 정보를 학습에 반영한 경우가 예측정확도가 높게 나타났다. 향후 땅밀림 현장조사 자료가 지속적으로 확보될 경우, 본 연구에서 활용한 기계학습기법은 땅밀림 분류를 위한 도구로 활용이 가능할 것으로 판단된다.

핵심용어: 위험등급 분류, 땅밀림, 기계학습, 산지토사재해

ABSTRACT

In this study, classification models were built using machine learning techniques that can classify the soil creep risk into three classes from A to C (A: risk, B: moderate, C: good). A total of six machine learning techniques were used: K-Nearest Neighbor, Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting and then their classification accuracy was analyzed using the nationwide soil creep field survey data in 2019 and 2020. As a result of classification accuracy analysis, all six methods showed excellent accuracy of 0.9 or more. The methods where numerical data were applied for data training showed better performance than the methods based on character data of field survey evaluation table. Moreover, the methods learned with the data group (R1~R4) reflecting the expert opinion had higher accuracy than the field survey evaluation score data group (C1~C4). The machine learning can be used as a tool for prediction of soil creep if high-quality data are continuously secured and updated in the future.

Keywords: Risk classification, Soil creep, Machine Learning, Forest soil sediment disaster

***Corresponding author:** Chang Woo Lee, leecwfr@korea.kr

Received: 15 August 2021, **Revised:** 7 September 2021, **Accepted:** 12 September 2021

1. 서 론

우리나라에서의 지진발생횟수는 1998년 이후 뚜렷한 증가추세를 보이고 있으며, 최근에는 포항, 경주 지역을 중심으로 2016년 252회, 2017년 223회로 예년 평균에 비해 발생횟수가 급격히 증가한 것으로 보고되는 등 지진으로 인한 다양한 형태의 재해가 발생할 수 있을 것으로 예측되고 있다(Choi and Jang, 2017).

특히 최근 지진은 산지 내 대규모 균열 등으로 인한 구조물 피해와 함께 액상화, 산사태 및 토석류 등과 같은 다양한 산지토사재해를 유발하는 것으로 보고되고 있으며, 국내에서도 최근 빈번한 큰 규모의 지진으로 인해 그 어느 때 보다 지진재해에 관한 관심이 증폭되고 있다. 실제로 포항지진 발생 당시 진원지 인근에 설치되었던 땅밀림 무인원격 감시시스템에서 사면변위를 감지한 것을 계기로 지진으로 인한 산지토사재해의 발생 가능성도 제기되고 있다.

땅밀림은 일반적으로 활동면 위의 토층이 일체로 미끄러지는 산사태의 한 형태로 국내에서는 '자연적 또는 인위적 원인에 의해 산지 일부가 중력의 영향을 받아 천천히 밀려 내려오는 것'으로 정의하고 있다(NIFoS, 2018).

이처럼 국내에서도 많은 지역에서 땅밀림이 관측되고 있지만, 땅밀림과 관련된 연구는 땅밀림 발생원인과 조사에 대한 것이 대부분이다(Park, 2015). 땅밀림 현상은 다양한 내부요인(지질, 토양, 지형, 수문, 산림 등) 및 외부요인(강우, 지진, 대규모 공사 등)에 의해 복합적으로 일어나며, 이러한 땅밀림 위험성을 감시, 분석, 예측하기 위해서는 땅밀림 위험지역에 대한 위험등급 분류가 필요하다.

NIFoS(2018)에서는 땅밀림 발생지의 피해 예상 범위를 구획하여 피해를 사전에 방지할 목적으로 '땅밀림 발생지 현장조사 매뉴얼'을 작성하였으며, 여기에는 땅밀림 발생지의 규모, 입지특성 파악 및 체계적인 관리를 위한 현장조사 체계와 항목 등의 내용을 포함하고 있다. 실제로 해당 매뉴얼을 토대로 2018년도부터 전국 땅밀림 발생우려지역에 대한 조사가 수행되었으며 현장조사 결과로부터 위험등급을 구분한 바 있다.

다만, 위험등급을 구분할 경우, 현장조사 야장(field check list)의 평가점수를 기반으로 하여 위험등급 결정에 각 조사항목에 대한 평가점수에 따라 배점이 높은 항목에 대해 가중치가 부여되는 등 객관성이 결여될 수 있다(Table 1 참조). 따라서 현장조사 야장의 주관적 평가점수이외에 땅밀림 발생위험요인이 되는 대상지역의 지질, 토양, 지형학적 특성 등을 직접적으로 활용하여 위험등급을 분석할 수 있는 기술이 요구된다.

땅밀림을 포함한 산지토사재해의 피해를 저감하기 위해서는 재해와 관련된 정보를 조사하고 잠재적인 위험을 예측하는 등 재해예방이 필요하다. 기계학습을 이용한 산지토사재해 분석 및 예측 기술은 최근 들어 전술한 산지토사재해 다양한 요인들에 대한 정보가 수집되고 DB가 구축되면서 주요한 분석도구로 주목받고 있으며, 실제로 해당 분야에서는 매우 활용도가 높은 것으로 알려져 있다(Bergen et al., 2019; Ma et al., 2020).

산지토사재해 분야에서 활용되는 기계학습 사례는 1) 영상자료를 이용한 산지토사재해의 판별 또는 감지(Danneels et al., 2007; Ding et al., 2016), 2) 산지토사재해 인벤토리를 이용한 기계학습기반 위험도 분석(Akgun, 2012; Althuwaynee et al., 2014; Chen et al., 2018), 3) 시계열 정보를 이용한 산지토사재해 예·경보(Segnoi et al., 2015; Kirschbaum and Stanley, 2018) 등으로 구분할 수 있다.

본 연구에서는 산지토사재해 연구에서 비교적 많이 사용되는 6개의 지도학습기반 기계학습기법(supervised machine learning techniques)을 활용하여 2019년부터 2020년까지 2개년 동안 전국을 대상으로 땅밀림으로 의심되는 지역의 현장조사 결과를 기반으로 땅밀림 위험지역을 3개의 위험등급으로 구분하여 분류할 수 있는 모형을 구축하고, 위험등급 분류 정확도를 비교·분석한다.

2. 기계학습기법

2.1 K-Nearest Neighbor(K-NN)

K-NN은 거리기반 분류분석 모형이며, 거리기반으로 분류를 하는 군집화(clustering)와 유사한 개념이지만, 기존 관측값을 활용한다는 측면에서 지도학습 범주에 포함된다. 해당기법은 어떤 자료가 주어지면 그 주변(이웃)의 자료를 분석하여 더 많은 자료가 포함되어있는 범주로 분류하는 방식으로 매우 직관적인 특징을 가지고 있다. 현재 이미지처리, 글자인식, 패턴인식 등 매우 다양한 분야에서 응용되어 사용되고 있다(Sit et al., 2020).

K-NN의 구동은 1) 새로운 자료와 인근의 자료의 거리 분석, 2) 새로운 자료가 포함된 기존 자료의 빈도 분석, 3) 분류수행으로 구분할 수 있다. K-NN 기법은 타 기계학습 분류모형과 다르게 훈련이 따로 필요 없으며, 새로운 자료가 주어지면 주변의 K개 자료에 기반하여 새로운 자료를 분류하므로 사전 모델링이 필요하지 않고, 실시간(real-time) 분석이 가능하다.

2.2 Support Vector Machine(SVM)

SVM은 K-NN 기법과 마찬가지로 분류와 회귀 문제에 모두 활용이 가능하며, 주어진 자료로부터 새로운 자료가 어느 카테고리에 포함될 것인지 판단하는 비확률적 이진선형분류(또는 비선형분류) 모형을 만들게 된다(Sit et al., 2020).

SVM의 기본적인 원리는 학습자료로 주어졌을 때 두 카테고리에서 개별 자료의 거리를 측정하여 두 개의 자료군 사이의 중심을 구한 후에 최적초평면(optimal hyperplane)을 구함으로써 두 개의 자료군으로 분류하는 방법을 학습하게 된다. 여기서 선형으로 자료를 구분할 수 있다면 선형 분류 모형을 적용하고, 그렇지 못할 경우 비선형 분류 모형을 사용하게 된다. 다만, SVM을 이용한 비선형 분류를 위해서는 주어진 자료를 고차원 특징 공간(feature space)으로 사상(projection)하는 작업이 필요하다. 이를 효율적으로 실행하기 위해서 커널(kernel)을 사용한다.

2.3 Logistic Regression(LR)

LR은 독립 변수의 선형 결합을 이용하여 사건의 발생을 예측하는데 사용되는 통계기법이다. LR 기법은 일반적인 회귀 분석기법과 마찬가지로 종속변수와 독립변수 간의 관계를 구체적인 함수로 나타내어 향후 예측 모형에 사용하게 된다(Chae et al., 2004). 이는 독립 변수의 선형결합으로 종속변수를 추정하는 선형회귀(linear regression) 분석과 유사하지만, 선형회귀분석과 다르게 종속변수를 범주형 자료로 하여 입력자료를 줬을 때 해당 자료의 결과가 특정 군(group)으로 나뉘기 때문에 자료의 분류(classification)에 활용될 수 있다.

일반적으로 LR은 종속변수가 이항형 문제(즉, 유효한 범주의 개수가 두 개인 경우)에 사용되나, 두 개 이상의 범주를 갖는 분류 문제인 경우, MLR(multinomial logistic regression) 또는 PLR(polytomous logistic regression) 기법을 이용하여 자료를 분석하게 된다.

2.4 Decision Tree(DT)

DT는 의사결정 규칙에 따른 결과들을 나무(tree) 구조로 도식화한 의사결정지원 기법 중 하나이다(Byeon et al., 2008). 특정 항목에 대해 의사 결정하는데 있어서 질문을 내고 해당 질문에 대한 응답이 나무 구조를 이루고 있는 형태이다. DT는 분류와 회귀 문제에 많이 사용되고 있으며, 최종 의사결정에 도달하기 위해서는 예/아니오의 이진 분류에 대한 질문을 지속하면서 학습을 진행하게 된다.

DT에서는 한번 분기 때마다 변수 영역을 두 개로 구분하는 모형이며, 분류 수행 후 각 영역의 순도(homogeneity)가 증가, 불순도(impurity) 혹은 불확실성(uncertainty)이 최대한 감소하도록 학습을 진행한다. 여기서 불순도는 자료의 복잡성을 의

미하며, 해당 범주 안에 서로 다른 특징을 갖는 자료가 얼마나 섞여있는지를 의미하고, 다양한 자료들이 복잡하게 섞여 있을수록 불순도가 높아진다.

DT는 자료의 전처리(정규화, 결측치, 이상치 등)가 필요하지 않으며, 자료의 형태가 수치형(numeric)과 범주형(categorical) 변수를 모두 다룰 수 있는 장점이 있다.

2.5 Random Forest(RF)

전술한 DT의 경우, 자료의 전처리 생략 및 범주형 자료 처리 등 장점이 있지만, 자료의 규모가 클 경우, 과적합(overfitting)으로 인해 학습의 효율성이 저하될 수 있다.

RF는 DT 기법의 단점을 보완하고자 같은 자료에 대해 여러 개의 의사결정나무를 생성하고 그 결과를 종합하여 예측성능을 향상시키는 기법이다(Woo et al., 2018). 즉, RF는 훈련을 통해 구성해놓은 다수의 DT로부터 분류 결과를 취합해서 결론을 얻게 되며, 몇몇 DT에서는 과적합이 발생할 수 있으나 다수의 DT를 기반으로 예측하기 때문에 그 영향력이 줄어들어 좋은 일반화 성능을 보일 수 있다.

이와 같이 예측에 있어 좋은 성능을 얻기 위해 다수의 학습 알고리즘을 사용하는 것을 앙상블(ensemble) 학습법이라고 부르며, RF 기법은 앙상블 예측기법 중 하나이다.

2.6 Extreme Gradient Boosting(XGB)

부스팅(boosting)이란 여러개의 약한 DT를 조합해서 사용하는 앙상블 기법중 하나이며 약한 예측모형들의 학습에러에 가중치를 두고, 순차적(sequential)으로 다음 학습 모형에 반영하여 강력한 예측모형을 만드는 것을 의미한다(Sit et al., 2020).

여기서, XGB는 DT 기반의 앙상블 기계학습기법으로 경사부스팅(gradient boosting, GB) 프레임워크를 사용한다. 이러한 GB 알고리즘의 병렬학습이 지원되도록 구현한 라이브러리가 XGB이며, 정형자료(structured data)를 대상으로 예측할 경우, 매우 우수한 성능을 보이는 것으로 알려져 있다.

XGB는 GB에 비해 병렬처리가 가능하여 학습 및 분류 속도가 빠르며, 과적합 규제기능을 가지고 있다. 또한, 다양한 분류 기계학습 모형과의 조합(customizing)이 용이하다.

3. 적용 및 분석결과

3.1 자료의 수집 및 처리

전국단위 땅밀림 자료는 현장조사를 통한 지질특성, 토양특성, 지형특성, 수리특성, 산림특성, 기타 땅밀림 징후 등을 종합적으로 검토하여 Table 1의 위험도 판정점수를 근거로 A(위험: 65점 이상, 땅밀림 징후가 있는 경우), B(보통: 35점~65점, 땅밀림 징후가 약간 있는 경우), C(35점 미만, 땅밀림 징후가 없는 경우)등급으로 구분한다.

2019년(2,000개소)부터 2020년(2,010개소)까지 전국 4,010개 땅밀림 현장조사 자료를 수집하였으며, Table 1의 14개 조사항목을 기준으로 구분된 등급별 비율은 A등급(30개소, 0.7%), B등급(1,010개소, 25.2%), C등급(2,970개소, 74.1%)이다. 다만, 현장조사 후 전문가 의견에 의해 재분류된 등급의 경우는 A등급(28개소, 0.7%), B등급(68개소, 1.7%), C등급(3,914개소, 97.6%)으로 대부분의 땅밀림 등급이 C등급으로 나타났다. 즉, 땅밀림 자료는 Table 1의 현장조사 평가점수에 대한 자료군과 현장조사와 전문가의견이 반영된 자료군으로 구분된다.

Table 1. Filed survey check list of soil creep

No	Factor	Score (Level)	Category				
			1	2	3	4	5
	Score	25					
	Level	C					
1	Direct Sign	No 0	Yes 22	No 0			
2	Indirect Sign	No 0	Yes 14	No 0			
3	Parent rock	Plutonic 2	Sedimentory 8	Metamorhphic 5	Igneous 2	Plutonic 2	
4	Rock Weathering	Normal 3	Weathered 7	Soft 4	Normal 3	Hard 2	
5	Slope direction to discontinuous slope	Inverse 3	Slope 9	Vertical 5	Horizontal 4	Inverse 3	
6	Discontinuous Slope space	Wide 1	Very dense 5	Dense 4	moderate 2	wide 1	
7	Soil type	Sandy 2	Clay 5	Sandy loam 3	Sandy 2		
8	Soil Moisture	Dry 2	Very wet 5	Wet 4	moderate 3	Dry 2	Very dry 1
9	Tafoni	No 0	Yes 4	No 0			
10	Soil depth	< 30 cm 2	> 90 cm 5	60~90 cm 4	30~60 cm 3	< 30 cm 2	
11	Topography	mild hilly 4	mild hilly 4	hilly 3	mountainous 2		
12	Topographic shape (cross-section)	Linear 1	Comlplex 4	concave 3	convex 2	linear 1	
13	Topographic shape (longitudinal section)	Linear 1	Complex 4	concave 3	convex 2	linear 1	
14	Slope	20~30 4	20~30° 4	10~20° 3	> 30° 2	< 10° 1	

다만, Fig. 1과 같이 일부지역의 경우 현장조사 야장에 의한 평가점수가 65점 이상으로 A등급임에도 불구하고 전문가 의견에 의해 B등급으로 구분되거나, B등급임에도 A등급으로 구분된 지역이 상당수 포함된 것으로 조사되었다.

각 기법별 학습을 위해 개별항목별 평가점수는 최대점수와 최소점수의 차이가 각 항목별로 다르기 때문에 기계학습을 위해서는 해당점수를 표준화(standardization) 또는 정규화(normalization)시킬 필요가 있다.

본 연구에서는 현장조사 평가점수를 MinMaxScale을 이용하여 0~1로 정규화(normalization)시켰으며, 불연속면과 관련된 항목의 경우, 일부지역의 현장조사에서 평가점수가 누락된 자료의 경우, 일괄적으로 0 값으로 전체 자료를 처리하였다.

Location	Parent Rock	Rock Weathering	Soil Depth	Soil Type	Soil Moisture	…	Slope	Direct Sign	Indirect Sign	Checklist Score	Checklist Level	Researcher Level
GG-171	5	7	3	3	4		4	22	14	80	A	B
GN-156	8	7	3	3	4		3	22	14	74	A	B
GB-11	**8**	**7**	**4**	**3**	**3**	**…**	**4**	**22**	**14**	**83**	**A**	**B**
GG-03516	5	3	3	3	4		4	22	14	77	A	B
GN-7	2	3	3	3	4		4	22	0	52	B	A
JN-104	**2**	**3**	**3**	**3**	**3**		**2**	**22**	**0**	**48**	**B**	**A**
CN-29	2	7	3	3	3	…	2	22	0	51	B	A
NEW-1	2	3	4	5	3		3	22	0	50	B	A
GW-30	5	7	4	3	3		4	0	14	57	B	C
CN-140	**5**	**3**	**4**	**3**	**1**		**4**	**22**	**14**	**62**	**B**	**C**
JN-153	2	3	3	3	2	…	2	0	14	54	B	C
JN-154	2	3	3	3	2		2	0	14	54	B	C

Fig. 1. Difference of risk level between checklist-based and researcher-based

3.2 분석절차

기계학습기법을 이용한 땅밀림 위험등급 분류를 위해 우선 Fig. 2와 같이 현장조사 평가점수 기반 자료군(C cases: C1~C4)과 전문가의견이 반영된 자료의 2개(R cases: R1~R4)로 구분하였다. 그리고 각각의 자료군은 현장조사의 14개 조사항목의 평가점수(numeric data)를 모형학습에 사용하는 경우(C1~C2, R1~R2)와 조사항목의 조사결과명(text data)을 모형학습에 사용하는 경우(C3~C4, R3~R4)에 대하여 분석하였다. 여기서 조사결과명이란 Table 1의 14개 조사항목에 포함된 현장조사를 통해 분석된 4,010개의 대상지역에 대한 지질, 토양, 지형특성을 의미한다. 또한, 현장조사 등급판정에 가장 큰 비중을 차지하는 간접징후와 직접징후가 반영되었을 경우(C1, C3, R1, R3)와 반영되지 않았을 경우(C2, C4, R2, R4,) 역시 고려하여 모형을 구축하였다.

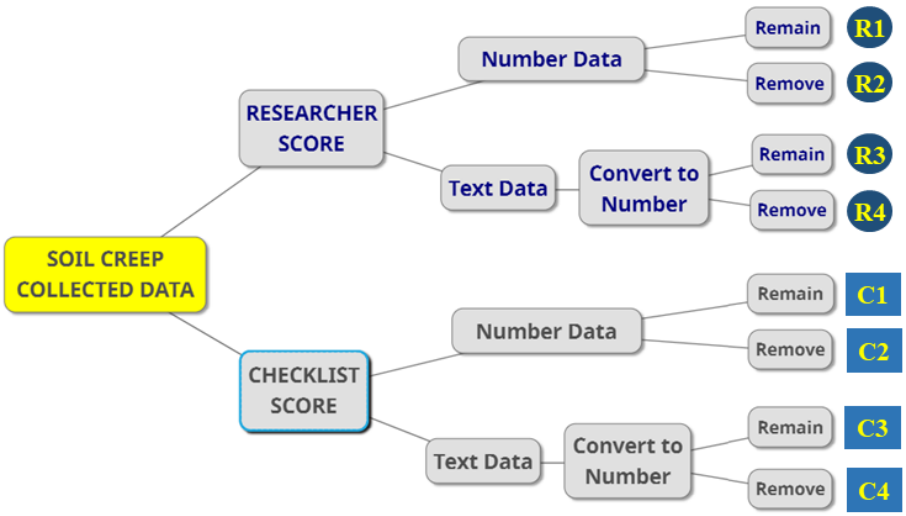


Fig. 2. Data set grouping for machine learning applications

또한, 각 기계학습기법의 학습을 위하여 전체 4,010개의 자료 중 80%를 사용하였으며, 20%는 학습된 모형의 예측 정확도를 분석하기 위해 사용하였다.

기계학습기법의 예측정확도를 높이기 위해서는 기법별 모형을 생성할 때, 사용자가 직접설정하는 초매개변수(하이퍼파라메터, hyperparameter)의 최적화가 필요하다. 이러한 초매개변수에는 DT의 개수, DT의 깊이, 학습에 사용되는 반복횟수 등이 있다. 따라서 본 연구에서는 각 기법별 필요한 초매개변수 설정을 위해 GR(grid search) 기법을 이용한다. GR은 초매개변수에 할당할 수 있는 값을 순차적으로 입력한 뒤 가장 높은 성능을 보이는 초매개변수를 최적값으로 설정하는 기법이다. 본 연구에서 사용된 기계학습기법들의 최적매개변수는 Table 2에 정리되어 있다.

Table 2. Optimal hyperparameters of 6 machine learning techniques used in this study

Technique	Optimal parameter	Remark
K-NN	n_neighbors: 4	Number of neighbors to use by default for kneighbors queries
	weights: uniform	weight function used in prediction
SVM	C: 1	Regularization parameter. The strength of the regularization is inversely proportional to C
	gamma: 0.001	Kernel coefficient for 'rbf', 'poly' and 'sigmoid'
	kernel: linear	Specifies the kernel type to be used in the algorithm. It must be one of 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' or a callable
LR	C: 10	Like in SVM, smaller values specify stronger regularization
	max_iter: 100	Maximum number of iterations taken for the solvers to converge
	penalty: l2	Used to specify the norm used in the penalization
DT	criterion: entropy	The function to measure the quality of a split
	max_depth: 20	The maximum depth of the tree
	max_leaf_nodes: 10	Best nodes are defined as relative reduction in impurity
	min_samples_split: 2	The minimum number of samples required to split an internal node
RF	n_estimators: 10	The number of trees in the forest
	max_depth: 50	The maximum depth of the tree
	max_features: auto	The number of features to consider when looking for the best split
	min_samples_leaf: 2	The minimum number of samples required to be at a leaf node
	min_samples_split: 2	Like in DT
	bootstrap: True	Whether bootstrap samples are used when building trees
XGB	n_estimators: 120	The number of boosting stages to perform
	max_depth: 50	The maximum depth of the individual regression estimators
	max_features: sqrt	Like in RF
	min_samples_split: 3	Like in DT

3.3 분석결과

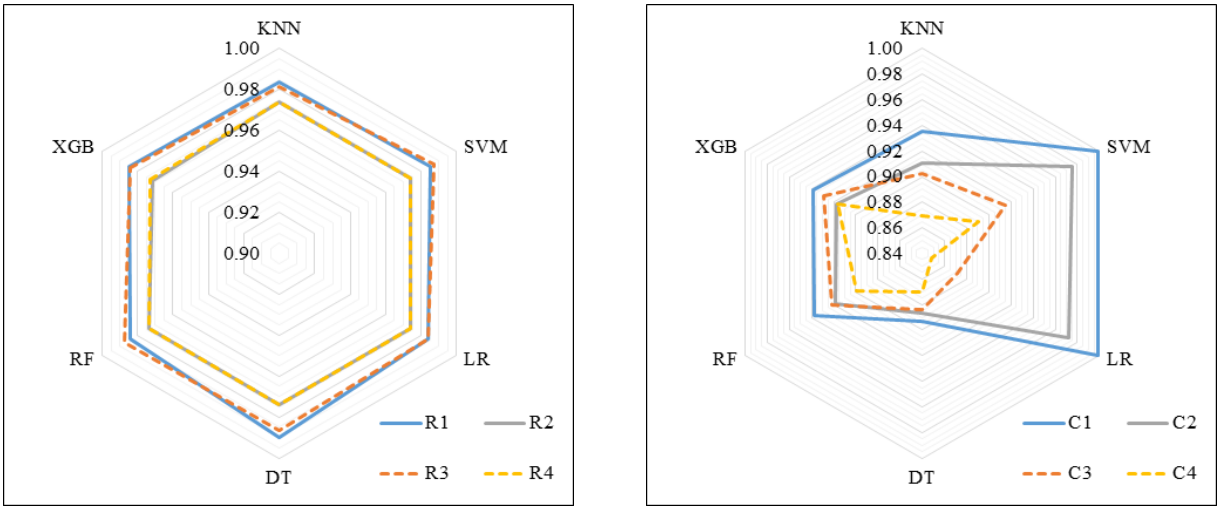
Fig. 3은 주어진 2개의 자료군(C cases, R cases)을 학습한 6개의 기계학습기법의 모든 위험등급 분류예측에 대한 종합적인 정확도(분류예측성공률=성공횟수/전체예측횟수)를 나타내고 있다. 즉, Fig. 3은 등급별 분류 정확도가 아닌 모든 등급에 대한 기법별 분류 정확도를 의미하며, 정량적 성능평가결과는 우수한 예측결과일수록 1에 가깝고 부정확할수록 0에 가깝다.

6개의 기법 모두 0.9 이상의 우수한 정확도를 보여주었으며, 수치형 자료를 학습에 적용한 경우가, 문자형 자료를 학습한 모형보다 우수한 성능을 나타냈으며, 현장조사 평가점수 자료군(C1~C4) 보다는 전문가의견이 반영된 평가점수 자료군(R1~R4)으로 학습한 모형이 정확도가 높은 것으로 분석되었다. 특히, 직접징후와 간접징후 정보를 학습에 반영한 경우가 예측정확도가 높게 나타났다.

가장 높은 예측정확도를 보여주고 있는 R1과 R3 자료군에 대해 기법별 예측실패 사례를 상세히 분석한 결과는 Fig. 4와

같으며, 현장조사 평가점수 기반 자료군 중 비교적 정확도가 높은 C1과 C2에 대한 분석결과는 Fig. 5와 같다. Fig. 4와 Fig. 5의 상단 결과는 각 위험등급별 실패개수를 나타내고 있으며, 하단은 실제 위험등급에 대해 모형에서 각기 다른 위험등급으로 구분된 결과를 나타내고 있다.

여기서 R 자료군 정확도 분석에 사용된 실측자료는 A등급: 6개, B등급: 15개, C등급: 781개이며, C 자료군은 A등급: 6개, B등급: 203개, C등급: 593개이다. 즉, C 자료군의 현장조사 평가점수를 통해 B등급으로 분류된 자료가 전문가 의견 반영을 통해 R 자료군에서는 C등급으로 재분류되었다.



(a) (b)

Fig. 3. Summary of classification accuracy by the six techniques

KNN	SVM	LR	DT	RF	XGB	Observed value	Classified value	KNN	SVM	LR	DT	RF	XGB
0	0	1	1	0	1	C (0)	B or A	4	0	0	1	0	1
8	8	9	4	10	9	B (1)	C or A	7	7	9	5	6	7
5	4	3	3	3	2	A (2)	C or B	4	3	4	5	4	5
13	**12**	**13**	**8**	**13**	**12**	**Total**		**15**	**10**	**13**	**11**	**10**	**13**

Case R1 **Case R3**

KNN	SVM	LR	DT	RF	XGB	Observed value	Classified value	KNN	SVM	LR	DT	RF	XGB
0	0	1	1	0	1	C (0)	B	4	0	0	1	0	1
0	0	0	0	0	0		A	0	0	0	0	0	0
7	5	5	3	5	4	B (1)	C	6	5	5	3	4	3
1	3	4	1	5	5		A	1	2	4	2	2	4
2	0	0	0	0	0	A (2)	C	0	0	0	0	0	1
3	4	3	3	3	2		B	4	3	4	5	4	4

Fig. 4. Number of classification failures of R1 and R3

KNN	SVM	LR	DT	RF	XGB	Observed value	Classified value	KNN	SVM	LR	DT	RF	XGB
15	0	0	27	9	12	C (0)	B or A	17	0	1	18	11	14
35	1	1	58	36	36	B (1)	C or A	49	14	15	67	48	46
2	0	0	1	5	1	A (2)	C or B	6	6	6	6	6	6
52	**1**	**1**	**86**	**50**	**49**	**Total**		**72**	**20**	**22**	**91**	**65**	**66**

Case C1 ⇩ **Case C2**

KNN	SVM	LR	DT	RF	XGB	Observed value	Classified value	KNN	SVM	LR	DT	RF	XGB
15	0	0	27	9	12	C (0)	B	17	0	1	18	11	13
0	0	0	0	0	0		A	0	0	0	0	0	1
35	1	1	56	36	36	B (1)	C	49	14	15	67	48	46
0	0	0	2	0	0		A	0	0	0	0	0	0
0	0	0	0	0	0	A (2)	C	2	2	2	3	3	3
2	0	0	1	5	1		B	4	4	4	3	3	3

Fig. 5. Number of classification failures of C1 and C2

R 자료군의 경우, 모든 등급에 대하여 기법별로 8개에서 15개의 예측 실패가 있었으며, K-NN 기법이 가장 예측 실패가 가장 많은 것으로 분석되었다. R1 자료군에서 DT는 8개의 예측 실패가 있었으나 타 기법에 비해 가장 정확도가 높은 것으로 분석되다. R3 자료군에서 K-NN은 15개가 예측실패로 나타났으며 A 등급에 대해서는 예측정확도가 타기법에 비해 매우 낮은 것으로 나타났다. 또한, 학습량이 가장 많은 C등급의 경우, 대부분의 기법이 높은 예측 정확도를 보여주었으나 학습량 자료가 절대적으로 부족한 A, B등급의 경우, 예측 정확도가 감소하는 것으로 분석되었다.

특히, 현장 자료에서는 B등급으로 분류된 지역에 대해서 기계학습에서는 A 또는 C등급 분류된 경우가 가장 많은 것으로 나타났다. 다만, 실제 A등급에 대한 예측 오류는 R3의 XGB 사례(C등급으로 분류)를 제외하고는 모든 기법에서 B등급으로 분류되었다. R 자료군에 대한 기법별 정확도 비교는 Fig. 3(a)를 통해 확인할 수 있다.

C 자료군의 경우, R 자료군에 비해 SVM과 LR을 제외하고 예측 정확도가 매우 낮게 나탔으며, R 자료군에서 가장 우수한 성능을 보인 DT 기법이 가장 부정확한 예측값(86개 예측 실패)을 도출하는 것으로 분석되었다. 특히, 수치형 자료를 이용한 C1의 경우, 학습자료가 매우 부족한 A등급의 경우에도 SVM과 LR은 6개 모두 정확하게 예측한 것으로 분석되었다. 다만, 직간접징후를 학습에 사용하지 않은 C2 자료군의 경우, A등급 예측에 모두 실패하였다.

Figs. 3~5의 분석 결과와 같이 땅밀림 예측을 위한 기계학습의 정확도를 높이기 위해서는 문자형 자료보다는 평가점수를 활용하는 수치형 자료를 활용하고, 직간접징후를 모형학습에 활용하는 것이 유리하다. 다만, 지속적인 양질의 땅밀림 자료의 구축 및 업데이트는 기계학습기법 선택에 앞서 반드시 선행되어야 한다.

4. 결 론

본 연구에서는 최근 산지토사재해 연구에서 활발하게 사용되고 있는 기계학습기법을 이용하여 국내 땅밀림 위험등급을 분류하는 모형을 개발하고, 그 정확도를 비교·분석하였다. 기계학습기법의 지도학습을 위해 2개년(2019년과 2020년)의 국내 전국 땅밀림 조사결과를 수집·가공하였으며, K-NN, SVM, LR, DT, RF, XGB 총 6개의 기계학습기법을 적용하였다. 본 연구의 주요결과는 다음과 같이 요약할 수 있다.

분류 정확도 분석결과, 6개의 기법 모두 위험등급 전체에 대해서는 평균적으로 90% 이상의 우수한 분류 정확도를 보여주

었다. 특히, SVM과 LR은 수치형 자료와 문자형 자료에 대해 모두 우수한 성능을 보이고 있어 타기법보다 안정적인 결과를 제공해주었다. 또한, 현장조사 결과의 수치형 자료를 모형 학습에 적용한 경우가, 문자형 자료를 학습한 모형보다 우수한 성능을 나타냈으며, 현장조사 평가점수만을 고려한 자료군(C1~C4) 보다는 전문가 의견이 반영된 평가점수 자료군(R1~R4)으로 학습한 모형이 정확도가 높은 것으로 분석되었다. 특히, 직접징후와 간접징후 정보를 학습에 반영한 경우가 예측정확도가 높게 나타났다. 이는 현장조사 평가점수에서 징후에 대한 전문가의 의견이 가장 높은 배점을 차지하여 A 등급의 분류정확도를 높이는 효과에 의한 것으로 판단된다.

지도학습 기계학습기법들은 자료의 양과 질이 충분히 확보될 경우, 정확도가 향상된다. 본 연구에서의 땅밀림 자료들이 대부분 위험등급이 B, C 등급에 속하다보니 각 기법들은 충분한 학습을 통해 매우 높은 분류 정확도를 제공하였다. 다만, A 등급의 땅밀림 발생위험지역의 경우, 분류에 실패한 기법들이 많았으며, 이는 실제로 A 등급에 대한 학습자료가 절대적으로 부족하여 각 기법들의 학습이 충분하지 않은 것으로 판단된다.

현재 땅밀림 조사는 매년 갱신이 되고 있으며 그 품질 또한 향상될 것으로 기대된다. 따라서 이와 같은 양질의 현장조사 자료가 지속적으로 확보될 경우, 본 연구에서 활용한 기계학습기법은 땅밀림 분석을 위한 공학적 도구로 활용이 가능할 것으로 판단된다. 다만, 현장조사 야장의 평가점수의 가중치 부여와 관계없이 대상지역의 지질, 지형, 토양 특성 정보만으로 땅밀림 위험등급을 구분할 수 기계학습기반 분류모형의 고도화가 요구된다.

Acknowledgment

This subject is supported by Korea Ministry of Environment as “The SS projects; 2019002830001”.

References

Akgun, A. (2012). A Comparison of Landslide Susceptibility Maps Produced by Logistic Regression, Multi-Criteria Decision, and Likelihood Ratio Methods: A Case Study at İzmir, Turkey. Landslides. 9(1): 93-106.

Althuwaynee, O. F., Pradhan, B., Park, H. J., and Lee, J. H. (2014). A Novel Ensemble Decision Tree-based CHi-squared Automatic Interaction Detection (CHAID) and Multivariate Logistic Regression Models in Landslide Susceptibility Mapping. Landslides. 11(6): 1063-1078.

Bergen, K. J., Johnson, P. A., Maarten, V., and Beroza, G. C. (2019). Machine Learning for Data-driven Discovery in Solid Earth Geoscience. Science. 363(6433).

Byeon, S. H., Kang, H. J., Han, J. W., and Kim, T. W. (2008). Flood Mitigation Planing for a Basin Using a Decision Tree Model. Journal of Civil and Environmental Engineering Research B. 28(1B): 33-40.

Chae, B. G., Kim, W. Y., Kim, Y. C., Kim, K. S., Lee, C. O. and Choi, Y. S. (2004). Development of a Logistic Regression Model for Probabilistic Prediction of Debris Flow. The Journal of Engineering Geology. 14(2): 211-222.

Chen, W., Peng, J., Hong, H., Shahabi, H., Pradhan, B., Liu, J., Zhu, A., Pei, X., and Duan, Z. (2018). Landslide Susceptibility Modelling using GIS-based Machine Learning Techniques for Chongren County, Jiangxi Province, China. Science of the total environment. 626: 1121-1135.

Choi, S. W., Jang, W. C. (2017). Forecasting Probabilities of Earthquake in Korea Based on Seismological Data. The Korean Journal of Applied Statistics. 30(5): 759-774.

Danneels, G., Pirard, E., and Havenith, H. B. (2007). Automatic Landslide Detection from Remote Sensing Images using Supervised Classification Methods. In 2007 IEEE International Geoscience and Remote Sensing Symposium. 3014-3017.

Ding, A., Zhang, Q., Zhou, X., and Dai, B. (2016). Automatic Recognition of Landslide Based on CNN and Texture Change Detection. In 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC). 444-448.

Kirschbaum, D. and Stanley, T. (2018). Satellite-based Assessment of Rainfall-triggered Landslide Hazard for Situational Awareness. Earth's Future. 6(3): 505-523.

Ma, Z., Mei, G., and Piccialli, F. (2020). Machine Learning for Landslides Prevention: A Survey. Neural Computing and Applications. 1-27.

National Institute of Forest Science. (2018). Field Survey Manual of Soil Creep. Seoul: NIFoS.

Park, J. H. (2015). Analysis on the Characteristics of the Landslide-with a Special Reference on Geo-topographical Characteristics. Journal of Korean Society of Forest Science. 104(4): 588-597.

Segoni, S., Lagomarsino, D., Fanti, R., Moretti, S., and Casagli, N. (2015). Integration of Rainfall Thresholds and Susceptibility Maps in the Emilia Romagna (Italy) Regional-scale Landslide Warning System. Landslides. 12(4): 773-785.

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., and Demir, I. (2020). A Comprehensive Review of DEEP Learning Applications in Hydrology and Water Resources. Water Science and Technology. 82(12): 2635-2670.

Woo, S. Y., Jung, C. G., Kim, J. U. and Kim, S. J. (2018). Assessment of Climate Change Impact on Aquatic Ecology Health Indices in Han River Basin using SWAT and Random Forest. Journal of Korea Water Resources Association. 51(10): 863-874.

Korean References Translated from the English

국립산림과학원 (2018). 땅밀림 발생지 현장조사 매뉴얼. 서울: 국립산림과학원.

박재현 (2015). 땅밀림 산사태의 발생특성에 관한 분석 - 지형 및 지질특성을 중심으로 -. 한국임학회지. 104(4): 588-597.

변성호, 강현직, 한정우, 김태웅 (2008). 의사결정나무모형을 이용한 유역내 구조적 홍수방어 대안 도출. 대한토목학회논문집 B. 28(1B): 33-40.

우소영, 정충길, 김진욱, 김성준 (2018). SWAT 및 random forest를 이용한 기후변화에 따른 한강유역의 수생태계 건강성 지수 영향 평가. 한국수자원학회논문집. 51(10): 863-874.

채병곤, 김원영, 조용찬, 김경수, 이춘오, 최영섭 (2004). 토석류 산사태 예측을 위한 로지스틱 회귀모형 개발. 지질공학. 14(2): 211-222.

최서원, 장원철 (2017). 지진 관측자료를 기반으로 한 한반도 지진 발생확률 예측. 응용통계연구. 30(5): 759-774.