의사결정나무모형을 이용한 급경사지재해 예측기법
Prediction method of slope hazards using a decision tree model

송영석[1], Young-Suk Song, 채병곤[1], Byung-Gon Chae, 조용찬[1], Yong-Chan Cho

[1] 한국지질자원연구원 산사태재해연구팀 선임연구원, Senior Researcher, Landslide Hazards Research Team, Korea Inst. of Geosci. & Mineral Res.

SYNOPSIS : Based on the data obtained from field investigation and soil testing to slope hazards occurrence section and non-occurrence section in gneiss area, a prediction technique was developed by the use of a decision tree model. The slope hazards data of Seoul and Kyonggi Province were 104 sections in gneiss area. The number of data applied in developing prediction model was 61 sections except a vacant value. The statistical analyses using the decision tree model were applied to the entrophy index. As the results of analyses, a slope angle, a degree of saturation and an elevation were selected as the classification standard. The prediction model of decision tree using entrophy index is most likely accurate. The classification standard of the selected prediction model is composed of the slope angle, the degree of saturation and the elevation from the first choice stage. The classification standard values of the slope angle, the degree of saturation and elevation are 17.9°, 52.1% and 320m, respectively.

Key words : slope hazards, gneiss area, prediction technique, decision tree model, entrophy index

1. 서 론

우리나라 연평균 강우량중 절반이상이 7월과 8월에 집중되며, 이 시기에 토석류급경사지재해가 대부분 발생한다. Olivier(1994)는 24시간 동안의 강우량이 연평균 강우량의 20%를 초과할 경우 대형 급경사지재해가 일어날 수 있다고 보고한 바 있다. 그리고 Brand(1981)는 짧은 시간에 내리는 집중강우는 지질조건이나 수문지질 조건과 관계없이 대형 급경사지재해를 일으킬 수 있다고 보고한 바 있는데 이는 집중강우가 지표물질을 완전히 포화시킬 수 있는 상태의 강우량을 의미한다.

그런데 동일한 강우량을 갖는 지역에서도 급경사지재해가 발생되는 지역과 발생되지 않는 지역으로 구분된다. 이는 강우량이 급경사지재해를 발생시키는 가장 큰 요인임에도 불구하고 지반 및 지질매체의 특성에 따라 급경사지재해 발생정도가 다름을 의미한다. 즉, 지반 및 지질매체의 공학적 특성에 따라 동일한 강우조건에서도 급경사지 재해가 발생되는 경우와 발생되지 않는 경우로 나눌수 있다. 따라서 일정 강우조건하에서 대상지역의 어떤 지반조건 및 지질조건일 때 과연 급경사지재해가 발생하며 정량적으로 급경사지재해 발생가능성을 예측하는 것은 매우 중요한 사항이다.

김원영 등(2003)은 지질조건별 국내에서 발생된 자연사면의 산사태 발생특성 및 원인을 규명하고, 이에 대한 자료를 조사하였다. 이를 토대로 광역적인 지역을 대상으로 산사태 발생가능성을 예측하기 위하여 로지스틱 회귀모델을 이용한 산사태 예측모델을 개발하여 일부지역에 대한 산사태 예측지도를 작성한 바 있다. 그러나 김원영 등(2003)의 산사태 예측모델은 전문적인 지식을 가진 지질 및 GIS전문가에 의해서만 수행이 가능하므로, 본 기술을 범용화 및 실용화하기에는 여러 가지 문제점이 있다. 그러므로 일반 지질 및 토목기술자가 쉽게 활용할 수 있는 단순하고 정확한 예측모델의 개발이 필요하다.

따라서 본 연구에서는 기 조사된 편마암 지역에서의 급경사지재해 발생지역 및 미발생지역에 대한 현장조사자료 및 토질시험자료를 토대로 통계적인 분석방법인 의사결정나무모형(decision tree model)을 이용하여 급경사지재해 정밀예측모델을 개발하고자 한다. 이를 위하여 의사결정나무모형의 분석방법 가운데 엔트로피 지수를 활용하고자 한다. 그리고 예측모델에 대한 정확성을 검증하기 위하여 정오분류표를 적용하고자 한다.

2. 급경사지재해 자료조사

새로운 급경사지재해 예측모델 개발하기 위하여 가장 먼저 수행되어야 할 사항은 현재까지 발생된 급경사지재해에 대한 자료수집 및 발생특성을 분석하는 것이다. 본 연구에서는 최근 10년간 급경사지재해가 발생된 지역가운데 서울 및 경기지역을 대상으로 조사된 자료를 활용하였다. 이들 자료는 자연사면에서의 급경사지재해 발생지역에 대한 야외 정밀조사 및 토질시험결과를 토대로 수집된 것이다.

대상지역인 서울 및 경기지역은 서울, 포천, 성동, 문산 등으로서 1998년 8월 4일부터 7일까지 최고 588.5mm의 집중호우가 내렸으며, 이로 인하여 많은 급경사지 재해가 발생된 지역이다. 대상지역의 지질조건은 모두 편마암이며, 총 104개소의 현장정밀 조사자료 및 토질시험자료를 활용하였다(김원영 등, 2000).

표 1은 본 연구에서 수집된 지역별 급경사지재해 발생구간 및 미발생구간의 개소수를 정리한 것이다. 표에서 보는 바와 같이 급경사지재해 발생구간은 77개소이고, 미발생구간은 27개소로서 총 104개소에 대한 자료를 수집하였다. 이들 정밀조사된 자료를 토대로 의사결정나무모형을 이용한 새로운 급경사지재해 정밀예측모델을 개발하였다.

표 1. 급경사지재해 발생 및 미발생구간의 개소수

지역	급경사지재해		합계
	발생구간	미발생구간	
서울 및 경기지역	77	27	104

3. 의사결정나무모형 이론

의사결정나무는 분석대상에 대한 분류나 예측을 수행하기 위해서 사용되는 분석기법으로 대용량의 데이터 내에 존재하는 관계, 패턴 및 규칙 등을 탐색하고 모형화하는 역할을 수행하며, 신경망이나 판별분석 등에 의한 방법과는 달리 적용결과에 의해 규칙을 명확하게 나타낼 수 있다. 또한, 예측모형 자체뿐만 아니라 최적의 결과를 검색하거나 분석에 필요한 변수 간의 교호효과, 즉 두 개 이상의 입력변수가 결합하여 목표변수에 어떻게 영향을 주는지를 찾아내는데 이용된다(김종규 외, 2006). 특히, 나무모형구조로 표현되기 때문에 다른 기법들과 비교하여 쉽게 이해되고 설명할 수 있으며, 임의의 데이터 범주에서 동일한 특성을 갖는 집합으로 구분하여 특성을 정의하고, 목표변수에 대한 규칙을 추론하여 미래에 대한 예측을 할 경우 유용하게 활용할 수 있다(최기헌, 1995).

본 연구에서는 서울 및 경기 편마암 지역에서의 급경사지재해 발생지역 및 미발생지역에 대한 현장조사자료 및 토질시험자료를 토대로 의사결정나무모형을 이용하여 급경사지재해 정밀예측모델을 개발하였다. 그리고 의사결정나무모형을 이용한 급경사지재해 예측결과의 정확성을 검증하기 위하여 정오분류표를 이용하였다.

3.1 의사결정나무모형 알고리즘

의사결정나무모형은 의사결정규칙(decision rule)을 나무구조로 도표화하여 관심대상이 되는 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 이 방법은 분류 또는 예측의 과정이 나무구조에 의한 추론규칙에 의해 표현되기 때문에 신경망, 판별분석 등에 비해 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있는 분석방법이다.

의사결정나무분석을 위해서 CHAID, CART, QUEST 등과 같은 다양한 알고리즘이 제안되어 있으며 최근에는 이들의 장점을 결합하여 보다 개선된 알고리즘들이 제안되고 상용화되고 있다. 의사결정나무모형의 대표적인 알고리즘은 CHAID (Chi-squared Automatic Interaction Detection) 알고리즘(Kass, 1980)으로 명목형, 순서형, 연속형 등 모든 종류의 목표변수와 분류변수에 적용이 가능하며, Exhaustive CHAID 알고리즘(Biggs et al, 1991)으로 발전하였다. 그 밖에 CART(Classification and Regression Tree), QUEST(Quick, Unbiased, Efficient, Statistical), C5.0, C4.5 알고리즘 등이 있다.

순수도(purity) 또는 불순도(impurity)를 기준으로 자식마디를 형성해 나가는 순수도 지수(purity index)중 목표변수가 이산형인 경우에는 목표변수의 각 범주에 속하는 빈도(frequency)에 기초하여 분리가 일어난다. 이때 사용되는 주요 분리기준(partitioning criterion)으로는 카이제곱 통계량(chi-square statistic)의 p-값, 지니 지수(gini index), 엔트로피 지수(entropy index) 등이 있다. 특히 엔트로피 지수는 식 (1)과 같이 표현되며, 다항분포에서의 우도비 검정통계량을 사용하는 것과 같은 것으로 알려져 있고 최근에 널리 알려진 알고리즘인 C4.5는 엔트로피 지수를 분리기준으로 사용한다.

$$E = -\sum_{j=1}^{c} P(j)\ln P(j) \qquad (1)$$

여기서, j : 1, 2, …, c로서, c는 목표변수의 범주수

$P(j)$: 해당마디에서의 j번째 그룹에 속하는 자료의 비율을 추정치로 사용

3.2 예측모형의 평가방법

일반적인 모형평가의 기준으로는 모형이 얼마나 효과적으로 구축되었는가 즉, 얼마나 적은 입력변수로 모형을 구축했는가의 문제나 혹은 같은 모집단 내의 다른 데이터에 적용하는 경우 얼마나 안정적인 결과를 제공해 주는가 즉 일반화의 가능성 등 여러 각도에서 생각할 수 있다. 그러나 무엇보다도 우선적으로 고려되어야 할 사항은 구축된 모형이 얼마나 예측과 분류에서 뛰어난 성능을 보이는가를 알아보는 것이다. 이는 아무리 안정적이고 효과적인 모형도 실제 문제에 적용했을 경우 빗나간 결과만을 양산한다면 아무런 의미가 없기 때문이다. 따라서 모형의 평가는 예측을 위해 만든 모형이 임의의 모형보다 우수한지, 고려된 다른 모형과 비교하여 어느 것이 가장 우수한 예측력을 보유하고 있는지를 비교 분석하는 과정이라 할 수 있다. 모형의 평가방법으로는 정오분류표((mis)classification table), Lift Chart의 %Response 이익도표, ROC 도표 등이 있다.

오분류표 평가방법은 목표변수가 범주형인 경우에 적용할 수 있을 것이다. 통계모형의 평가분석을 위해 사후확률(posterior probability)을 비교할 수 있다. 일반적으로 분류의 기준으로 삼는 사후확률(posterior probability)의 경계는 “1/(목표변수의 범주 개수)”로 삼는 것이 보통이다. 또한 구축된 모형에 대하여 예측과 분류가 얼마나 뛰어난 성능을 보이는가, 그리고 얼마나 안정적인가를 비교하기 위해서 training data(분석용 자료)와 validation data(평가용 자료)의 정분류율(판별력)을 비교하여 validation data의 오분류율을 선정한다. 추가적으로 구축된 모형별 오분류율에 대해서도 검토한다. 식 (2) 및 식 (3)은 정분류율 및 오분류율을 산정하는 방법이다.

$$\text{정분류율} = \frac{(\text{실제0, 예측0})\text{의 빈도} + (\text{실제1, 예측1})\text{의 빈도}}{\text{관찰치의 빈도}} \times 100(\%) \quad (2)$$

$$\text{오분류율} = \frac{(\text{실제0, 예측1})\text{의 빈도} + (\text{실제1, 예측0})\text{의 빈도}}{\text{관찰치의 빈도}} \times 100(\%) \quad (3)$$

4. 의사결정나무모형을 이용한 급경사지재해 예측모델

4.1 분석자료

본 연구에서는 기 조사된 서울 및 경기지역의 급경사지재해 발생지역 및 미발생지역에 대한 현장조사자료 및 토질시험자료를 토대로 의사결정나무모형(decision tree model)을 이용하여 급경사지재해 정밀예측모델을 개발하였다.

표 2는 예측모델개발을 위해 서울 및 경기지역을 대상으로 조사된 자료수를 나타낸 것으로 총 104개소의 조사자료 가운데 결측치를 제외한 61개소의 자료를 활용하였다. 그리고 표 3은 각 지역별로 의사결정나무모형 분석에 포함된 변수를 나타낸 것이다. 표에서 *표시는 범주형 변수를 나타낸 것이며, 그 외는 연속형의 변수를 나타낸 것이다.

표 2. 분석자료수

지 역	총자료	분석용 자료
서울 및 경기지역	104	61

표 3. 변수항목

구분	변수	설명	서울 및 경기지역
목표변수	산사태 발생 여부	1:발생, 0:미발생	○
입력변수	lithology (*)	암석종류	○
	weathering (*)	풍화정도	×
	elevation	지형고도	○
	slope direction	사면방향	×
	slope angle	사면경사	○
	length	산사태 길이	×
	width	산사태 폭	×
	thickness	토층두께	×
	specific gravity	비중	○
	moisture content	함수비	○
	void ratio	간극비	○
	porosity	공극률	○
	degree of saturation	포화도	○
	wet(bulk) density	전체밀도	○
	saturation density	포화밀도	○
	dry density	건조밀도	○
	USCS (*)	입도분포	×
	permeability	투수계수	○
	triggering position	산사태발생위치	×
	gravel	자갈	×
	sand	모래	×
	silt / clay	실트/점토	×
	liquid limit	액성한계	○
	plastic limit	소성한계	○
	plasticity index	소성지수	○
	shear strength-cohesion	점착력	×
	shear strength-friction angle	내부마찰각	×

4.2 예측모델

편마암지역에서의 급경사지재해 예측모델을 개발하기 위하여 전술한 분석자료(n=61)를 토대로 의사결정나무모형을 이용한 통계적인 분석을 실시하였다. 의사결정나무모형을 이용한 통계적인 분석은 엔트로피 지수를 적용하였다. 본 의사결정나무모형에 대한 통계분석에는 SAS와 SAS/E-miner 프로그램을 사용하였다.

그림 1은 엔트로피 지수를 이용하여 의사결정나무모형 예측모델을 분석한 결과이다. 그림에서 보는 바와 같이 예측모델의 최상위 분리기준변수로는 사면경사가 선택되었으며, 하위 분리기준변수로는 포화도가 선택되었고, 최하위 분리기준변수로는 사면고도가 선택되었다. 급경사지재해 발생을 일으키는 사면경사의 기준은 17.9°인 것으로 나타났으며, 사면경사가 17.9°이상인 경우 급경사지재해 발생을 일으키는 토층의 포화도 기준은 52.1%인 것으로 나타났다. 그리고 토층의 포화도가 52.1%이상인 경우 급경사지재해 발생을 일으키는 사면고도의 기준은 320m인 것으로 나타났다. 이와 같은 의사결정나무모형을 이용한 예측모델을 평가하기 위하여 정오분류표를 활용하였다. 표 4는 엔트로피 지수를 이용한 의사결정나무모형 예측모델의 정오분류표를 나타낸 것이다. 그리고 식(2) 및 식(3)을 이용하여 정오분류표의 정분류율을 계산해보면 91.80%, 오분류율을 계산해보면 8.20%로 나타났다.

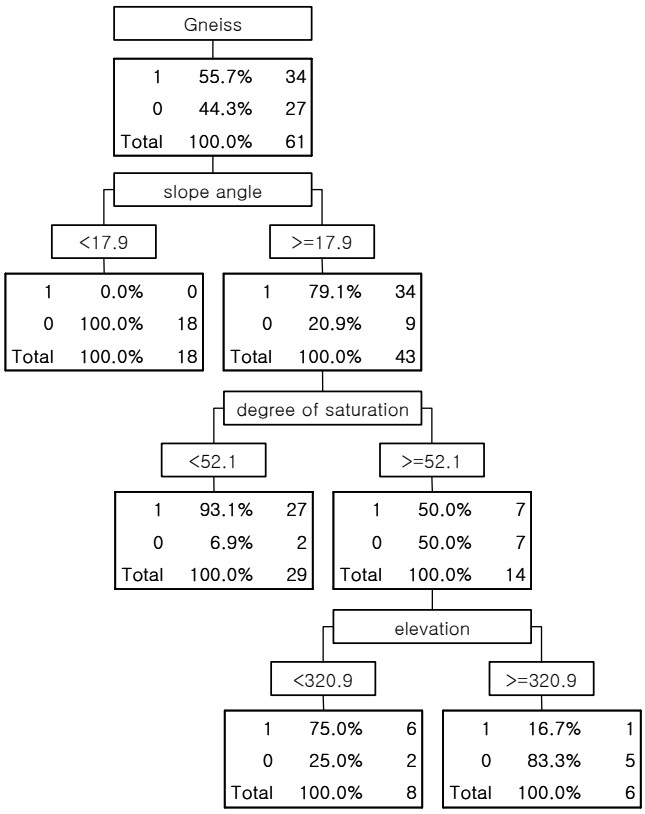

그림 1. 엔트로피지수를 이용한 의사결정나무모형

표 4. 정오분류표

예측값 / 실제 관측된 값	급경사지재해 미발생	급경사지재해 발생	합 계
급경사지재해 미발생	23	4	27
급경사지재해 발생	1	33	34
합 계	24	37	61

- 정분류율 $= \frac{(23+33)}{61} \times 100(\%) = 91.80(\%)$

- 오분류율 $= \frac{(4+1)}{61} \times 100(\%) = 8.20(\%)$

이상의 분석결과를 살펴보면 그림 1의 급경사지재해 예측모델은 정확도가 높은 것으로 평가되었으며, 분리기준도 합리적이라고 판단되었다. 따라서 이를 편마암지역에서의 급경사지재해 예측모델로 제안하고자 한다.

5. 결 론

본 연구에서는 의사결정나무모형을 이용한 편마암지역에서의 급경사지재해 예측모델을 개발하였다. 먼저 기 조사된 편마암 지역에서의 급경사지재해 발생지역 및 미발생지역에 대한 현장조사자료 및 토질시험자료를 토대로 통계적인 분석방법인 의사결정나무모형을 이용하여 급경사지재해 예측모델을 개발하였다. 이를 위하여 엔트로피 지수를 활용하여 분석을 실시하였으며, 이들 결과를 정리하면 다음과 같다.

1. 대상지역은 서울, 포천, 성동, 문산 등 서울 및 경기지역으로서 1998년 8월 4일부터 7일까지 최고 588.5mm의 집중호우로 인하여 급경사지재해가 발생된 구간이다. 대상지역의 지질조건은 모두 편마암으로서, 예측모델을 개발하기 위하여 활용된 조사자료수는 현장조사 및 토질시험 결측치를 제외한 61개소이다. 이 가운데 급경사지재해 발생구간은 34개소이고, 미발생구간은 27개소이다.
2. 의사결정나무모형을 이용한 통계적인 분석은 엔트로피 지수를 적용하여 실시하였다. 엔트로피 지수를 이용한 분석결과의 경우 사면경사, 포화도 및 사면고도가 분리기준으로 선택되었다.
3. 선정된 급경사지재해 예측모델의 분리기준은 최상위부터 사면경사, 포화도 및 사면고도의 순서로 선택되었으며, 각각의 분리기준치는 사면경사의 경우 17.9°, 포화도의 경우 52.1%, 사면고도의 경우 320m로 결정되었다.
4. 정오차분류법에 의한 예측모델의 정확성을 평가한 결과 엔트로피 지수를 이용한 의사결정나무모형 예측모델의 정분류율은 91.80%로서 높게 나타났다.

감사의 글

본 연구는 2006 건설기술혁신사업인 '국가 주요시설물 안전관리 네트워크 시범구축 및 운영시스템 개발'의 세부협동과제인 'GIS기반 급경사지 재해위험 취약지구 선정기법 연구'의 일환으로 수행되었습니다.

참고문헌

1. 김원영, 채병곤, 김경수, 기원서, 조용찬, 최영섭, 이사로, 이봉주 (2000), **산사태 예측 및 방지기술연구**, 과학기술부, 한국자원연구소, KR-00-(T)-09, 642p.
2. 김원영, 채병곤, 김경수, 조용찬, 최영섭, 이춘오, 이철우, 김구영, 김정환, 김준모 (2003), **산사태 예측 및 방지기술연구**, 과학기술부, 한국지질자원연구원, KR-03-(T)-03, 339p.
3. 김종규, 사공명, 이준석, 이용주 (2006), "의사결정트리 기법을 이용한 터널 보조공법 선정방안 연구", **대한토목학회논문집**, 제26권 제4C호, pp.255-264.
4. 최기헌 (1995), **데이터 마이닝: 개념 및 기법**, 자유아카데미.
5. Biggs, D., de Ville, B. and Ville, E. (1991), "A method of choosing multiway partitions for classification and decision tree", *Journal of Applied Statistics*, Vol.18, pp.46-62.
6. Brand, E. W. (1981), "Some thoughts on rainfall-induced slope failures", *Proceedings of 10th International Conference on Soil Mechanics Foundation Engineering*, Stockholm, The Netherlands, pp.373-376.

7. Kass, G. (1980), *An exploratory technique for investigating large quantities of categorical data*, Applies Statistics.
8. Olivier, M. Bell, F. G. and Jemy, C. A. (1994), "The effect of rainfall on slope failure, with examples from the Greater Durban area", *Proceedings of 7th intern. Cong. IAEG*, Vol.3, pp.1629-1636.