

의사결정트리에서 공간사건 예측을 위한 리프노드 등급 결정 방법 분석

연영광^{1*}

Analysis of Leaf Node Ranking Methods for Spatial Event Prediction

Young-Kwang YEON^{1*}

요 약

공간사건들은 데이터마이닝 분류알고리즘을 이용하여 예측 가능하며, 의사결정 트리는 대표적인 분류알고리즘들 중 하나로 사용되고 있다. 의사결정 트리는 레이블 값을 갖는 분류작업에 주로 사용되었으나 규칙평가 기법을 트리 리프노드 등급 계산에 응용하면서부터 공간사건 예측에 이용되고 있다. 이 논문에서는 의사결정 트리에서 사용되는 규칙평가 방법들을 공간예측에 적용하여 비교하였다. 실험을 위해 의사결정 트리 알고리즘인 C4.5 알고리즘과 규칙 평가기법인 Laplace, M-estimate 및 m-branch 기법들을 구현하여 자연환경에서 발생하는 대표적인 공간예측 응용분야인 산사태에 적용하였다. 적용한 규칙 평가 기법들의 정확도 평가결과, 그 특성에 따라 정확도의 차이가 있었으며 m-branch가 가장 높은 성능을 보였다. 그러나 m-branch 및 M-estimate와 같이 별도의 파라미터를 갖는 경우 반복적으로 최적의 파라미터 값을 찾는 과정을 요구하였다. 따라서 적용 대상에 따라 선택적으로 활용할 수 있다. 이러한 의사결정 트리를 이용한 공간예측은 예측 결과뿐만 아니라 특정 위치에서의 예측결과에 대한 원인분석을 가능하게 함으로 다양한 응용을 가능하게 한다.

주요어 : 의사결정트리, 공간예측, 리프노드 등급결정, 예측정확도

ABSTRACT

Spatial events are predictable using data mining classification algorithms. Decision trees have been used as one of representative classification algorithms. And they were normally used in the classification tasks that have label class values. However since using rule ranking methods, spatial prediction have been applied in the spatial

2014년 9월 16일 접수 Received on September 16, 2014 / 2014년 11월 12일 수정 Revised on November 12, 2014 /
2014년 11월 24일 심사완료 Accepted on November 24, 2014

¹ 한국지질자원연구원 국토지질연구본부 Geoscience Research Division, Korea Institute of Geoscience and Mineral Resources

* Corresponding Author E-mail : ykyeon@kigam.re.kr

prediction problems. This paper compared rule ranking methods for the spatial prediction application using a decision tree. For the comparison experiment, C4.5 decision tree algorithm, and rule ranking methods such as Laplace, M-estimate and m-branch were implemented. As a spatial prediction case study, landslide which is one of representative spatial event occurs in the natural environment was applied. Among the rule ranking methods, in the results of accuracy evaluation, m-branch showed the better accuracy than other methods. However in case of m-brach and M-estimate required additional time-consuming procedure for searching optimal parameter values. Thus according to the application areas, the methods can be selectively used. The spatial prediction using a decision tree can be used not only for spatial predictions, but also for causal analysis in the specific event occurrence location.

KEYWORDS : *Decision Tree, Spatial Prediction, Leaf Node Ranking, Prediction Accuracy*

서론

공간사건들은 특정 환경적 조건에 의해 발생되며, 예측은 사건 발생위치의 환경적 조건을 학습하여 이미 발견된 사건 발생환경과 유사한 조건을 갖는 위치에 대하여 단위면적당 상대적 등급으로 그 결과가 표현된다. 이러한 공간사건의 예측 응용은 자연 분야의 광물탐사, 산사태 예측분석과 같이 어떠한 원인들에 의해 발생하는 공간 사건의 발생 패턴 분석에 적용할 수 있다. 공간사건의 예측은 데이터마이닝 분류기법을 이용하나 결과표현에 있어서 전통적인 분류기법에서의 기 정의된 클래스로의 분류대신 타깃 클래스로의 분류정도를 이용하여 상대적 등급으로 표현한다.

의사결정 트리는 분류기법들 중 하나로 학습된 트리 결과가 규칙형태로 직관적으로 변환가능하다. 이러한 특징은 예측분야에서 예측결과를 도출뿐만 아니라, 결과에 대한 원인분석을 가능하게 한다. 또한 다른 기법들에 비해 정확도와 빠른 처리 결과로 인해 널리 응용되고 있다(Ferri *et al.*, 2003; Pal and Mather, 2003; We *et al.*, 2008). 일반적인 트리기반 알고리즘은 트리성장 및 트리전지과정으로 구성된다. 트리알고리즘으로부터 구축되는 트리 구조의 차이는 주로 트리 성장과정에서 이용되

는 선택 속성기준, 트리전지기준으로부터 비롯된다.

Yeon *et al.*(2010)은 의사결정 트리를 이용한 공간사건 예측과정을 그림 1과 같이 개념화하였다. 그 과정은 입력데이터로 공간사건 발생 위치와 이와 관련된 유발 요인들을 학습하여 트리를 구축하는 단계와 구축된 트리의 리프노드 등급 결정 단계와 리프노드 등급을 사건발생 유발 환경요인과 대응되는 위치에 매핑하는 단계로 구성한다.

트리구축단계는 Entropy(Quinlan, 1993) 혹은 Gini index(Breiman *et al.*, 1984)와 같은 분류기준을 이용하여 가장 잘 분류가 되는 속성을 선택하며 자노드를 생성하며, 생성된 자노드로 분류된 훈련데이터들이 이동한다. 이러한 과정은 정지조건이 만족될 때 까지 지속된다. 분류된 훈련데이터들은 최종적으로 트리의 리프노드에 남는다. 공간사건 예측에 트리를 이용하기 위해서는 사건 클래스에 대한 상대적 등급형태로 그 결과가 표현되어야 한다. 리프노드에 포함된 클래스 분포를 이용하여 리프노드 등급단계에 이용한다.

리프노드 등급 결정방법은 의사결정트리에서 유도된 규칙간 상대적 우선순위를 선정하기 위한 방법인 Laplace(Zadrozny and Elkan, 2001), M-estimate(Cussents, 1993) 및 m-branch(Ferri *et al.*, 2003)기법들이 있다.

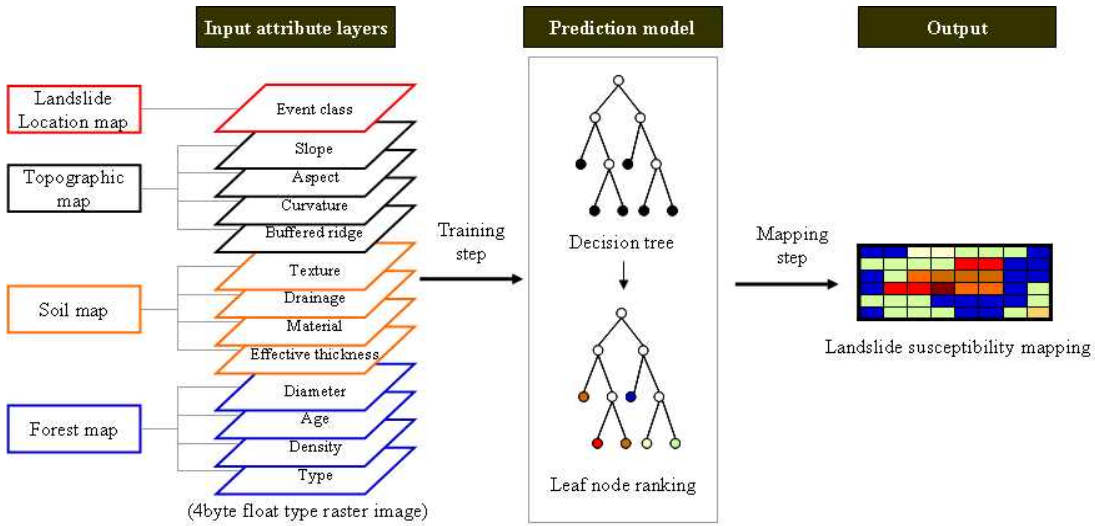


FIGURE 1. Spatial prediction using a decision tree(source : Yeon *et al.*, 2010)

그러나 공간예측 응용을 위해 규칙평가 방법들을 직접적으로 적용하는 것은 적합하지 못하다. 전통적인 의사결정 트리는 트리구축과정을 마친 후 각 리프노드에서 다수의 클래스로 출력 클래스가 결정된다. 반면 공간예측은 리프노드마다 공간사건 클래스를 출력클래스로 결정하고 이에 대한 상대적인 비율형태로 표현되어야 한다. 따라서 공간예측을 위해서는 규칙평가 방법을 공간사건에 해당하는 특정 클래스에 대한 상대적 비율로 수정이 필요하다. 이러한 계산적 특징으로 현재까지 공간 예측응용을 위한 정량적인 규칙평가 방법들의 성능 평가 사례가 없다.

이 논문에서는 대표적인 자연환경에서 발생되는 공간사건인 산사태 사례를 통해 기존의 규칙평가 방법들을 수정·응용하여 비교해 보고자 한다. 실험을 위해 의사결정 트리 알고리즘은 규칙평가 방법에서 이용되어온 Quinlan (1993)의 C4.5 알고리즘과 규칙평가 방법들을 구현하여 적용한다. 이 논문을 통해 기존 규칙평가 기법들에 대하여 공간 예측 적용방법과 각 방법들에 대한 성능적 특징을 분석해 보고자 한다.

의사결정 트리에서의 공간 사건 예측 응용

1. Entropy 기반의 속성 선택기법

의사결정트리의 노드는 속성선택 기준에 의해 해당 노드에 포함된 하나의 속성을 선택하여 분기되어 트리가 성장한다. 이 논문에서의 사용하는 의사결정 트리는 C4.5에서 사용하고 있는 Entropy기반의 속성 선택기준을 이용한다. 이 방법은 노드에 포함된 속성간의 무질서도를 낮추면서 진행한다. 무질서도를 산정하는 정보량은 Entropy로 정의되며 임의의 노드 N에서의 정보량이며 식 (1)과 같다. 이하 정의된 식은 Quinlan(1993)에 의해 설명되었다.

$$Entropy(N) = - \sum_j p(C_j|N) \log_2 p(C_j|N) \quad (1)$$

여기서 $p(C_j|N)$ 는 N에서의 상대 빈도이다. N에 포함된 k개의 속성에 대하여 속성 A를 선택하게 될 정보량(Entropy)은 다음 식 (2)와 같이 정의 된다.

$$Entropy_A(N) = \sum_{j=1}^k \frac{|N_j|}{|N|} \times Entropy(N_j) \quad (2)$$

정보이익은 원래 노드에 있었던 정보량에 대한 새롭게 분류한 정보량의 차에 대한 이익으로 InfoGain으로 정의되며, 이는 식 (3)과 같다.

$$InfoGain(A) = Entropy(N) - Entropy_A(N) \quad (3)$$

InfoGain에 의해 정보량이 가장적거나, 정보이익이 가장 큰 속성을 선택하게 되며 이는 분리를 많이 갖는 속성을 선택하는 경향이 있다. 이러한 경향은 연속형 속성을 포함하는 경우 해당 속성에 포함된 모든 값 사이가 분리점이기 때문에, 연속형 속성 속성에 편향되어 성장하게 된다. 이러한 문제를 회피하기 위해 분리정보를 이용하여 정보이익을 정규화 한다. 분리정보는 정보량과 유사하게 분기점에 따른 정보량으로 많은 수의 분기를 갖는 속성에 대하여 높은 값의 분리정보를 갖게 된다. 분리정보는 식 (4)와 같이 정의된다. 이하 정의된 식은 Quinlan(1993)에 의해 설명되었다.

$$SplitInfo_A = \sum_{j=1}^v \frac{|N_j|}{|N|} \times \log_2 \frac{|N_j|}{|N|} \quad (4)$$

따라서 정보량을 분리정보로 보상해준 값을 이득비(GainRatio)라 하며, 식 (5)로 정의된다.

$$GainRatio(A) = \frac{InfoGain(A)}{SplitInfo(A)} \quad (5)$$

2. 리프노드 등급 방법

의사결정트리에서 규칙은 리프노드로부터 루트노드에 이르기까지 일련의 'And' 조합으로 직관적으로 생성된다. C4.5와 같이 멀티클래스를 지원하는 트리 알고리즘은 규칙의 출력클래스로 리프노드에 포함된 다수의 클래스로 결정된다. 따라서 출력클래스를 대상으로 한 규칙평

가 기법들을 공간예측 응용을 위해서 특정 타깃 클래스에 대한 기준으로 수정되어 응용할 수 있다.

공간 예측에서 의사결정 트리는 별도로 규칙을 생성하지 않더라도 앞의 그림 1과 같이 구축된 트리에서 직접 리프노드의 등급을 평가하여 예측분석에 적용할 수 있다. 따라서 공간 사건 예측을 위해 규칙평가 방법들을 특정 리프노드 등급은 리프노드의 전체 클래스 분포에 대한 타깃 클래스의 비율로 식 (6)과 같이 정의될 수 있다.

$$P(node) = \frac{n_{event}}{n_{event} + n_{non_event}} \quad (6)$$

의사결정 트리의 구축과정은 분류기준에 의해 순수한 분류를 지향하여 궁극적으로 하나의 노드는 하나의 클래스만을 포함하기 위해 트리가 성장된다. 따라서 타깃 클래스에 대한 리프노드의 비 0 또는 1로 수렴한다. 이러한 문제를 해결하기 위해, Laplace 방법은 리프노드의 등급계산을 위해 리프노드에 포함된 사건클래스의 빈도를 라플라스 수정(Cestnik, 1990; Zadrozny and Elkan, 2001)을 응용하여 기존 확률 추정문제를 완화하기 위해 사용되며, 공간예측 응용을 위해 식 (7)과 같이 수정될 수 있다.

$$R(node) = \frac{n_{event} + 1}{n_{event} + n_{non_event} + c} \quad (7)$$

여기서 c 는 전체데이터 셀에서 클래스 수이다.

또 다른 리프노드 평가방법인 M-estimate (Cussens, 1993)은 사건클래스에 대한 선행 확률을 리프노드 등급산정에 사용할 수 있다. 파라미터 상수 b , m 에 대하여 b 가 긍정클래스의 선행 확률이며 특정 클래스를 대상으로 한 공간예측을 위해 식 (8)과 같이 수정된다. m 과 b 는 일정한 비율의 관계를 갖고 있다.

$$R(\text{node}) = \frac{n_{\text{event}} + bm}{n_{\text{event}} + n_{\text{non_event}} + m} \quad (8)$$

Ferri *et al.*(2003)은 m-branch 기법을 제안하였다. 이 방법은 루트-리프노드의 재귀적으로 M-estimate 확장한 것으로, 각 경로상, 부모 노드의 확률 추정치가 모든 자노드로 전파된다. m-branch 도 공간 예측을 위해 식 (9)와 같이 수정할 수 있다. 여기서, m은 노드의 깊이 혹은 차수, n_{c_j} 는 클래스 c_j 에 포함된 샘플의 수이다. s_t 레이블이 정해지지 않는 샘플이다.

$$R(\text{node.child}) = \frac{n_{\text{event}} + mR(\text{node.parent})}{n_{\text{event}} + n_{\text{non_event}} + m} \quad (9)$$

여기서, 파라미터 m은 $M+(d-1)/d \times M\sqrt{N}$ 으로 계산된다.

이와 같이 각 규칙평가 기법들을 공간예측에 적용하기 위해 출력클래스가 아닌 공간사건인 타킷 클래스에 대한 상대적인 비를 이용할 수 있도록 응용하였다. 이와 같은 규칙평가 방법들은 공통적으로 C4.5 알고리즘을 기반으로 응용되었다. 이 논문에서도 C4.5 알고리즘을 기반으로 공간사건에 응용해보고자 한다.

실험 및 평가

1. 데이터셋

산사태는 자연사면에서 발생되어 토층이나 풍화대 또는 토층과 풍화 암편이 집적된 풍적층 등의 미고결 물질이 집중 강우나 지진 등 지반진동 등에 의해 전단력이 약화되어 파괴가 발생사는 현상이다(Casale *et al.*, 1994). 산사태는 국내의 지형학적 및 지질학적 특성으로 인해 매년 인명 및 재산피해가 반복되고 있다. 산사태는 집중 강우 혹은 지진에 의한 외부적인 요소에 의해 발생되지만 지형, 임상, 토양 및 지질 환경적 요인에 의해 발생 가능성 혹은 규모가 달라진다(Dikau *et al.*, 1996). 이러한

특징을 이용하여 산사태의 예측 분석과 관련된 다양한 연구들(Lee *et al.*, 2002; Jo and Jo, 2009; Yeon, 2011; Park *et al.*, 2012)이 진행되었다. 선정된 연구지역은 강원도의 인제읍과 북면 사이에 지역이며, 산사태는 2006년 7월 11일부터 18일간 내린 집중호우로 인해 발생하였다. 이 지역의 평균 강우량은 1995년부터 2005년까지 약 1,400mm이며, 산사태가 발생한 2006년의 연간 강우량은 1740mm로, 산사태가 발생한 8일간의 559mm의 강우량이 당해년도 강우량에 큰 영향을 주었다.

산사태 발생 위치는 항공영상 및 수치고도 모델(DEM) 자료를 분석하여 수집되었다. 이외에 산사태 유발 환경요인으로 표 1과 같이 지형도, 산림도 및 토양도를 이용하였다. 지형적 요인들로 경사방위, 곡률, 능선, 경사를 이용했으며 이들은 산사태 발생과 관련된 지형학적 요인들이다. 산림도에서 연급, 밀도, 직경 및 형태를 이용하였으며, 임상의 형태나 상태에 따라 산사태 방지에 영향을 줄 수 있다. 토양 요인들로 토질, 배수, 모재, 유효토심두께를 이용하였으며, 이들은 강우로 인한 산사태에서 배수에 영향을 준다. 산사태 위치를 포함한 13개의 레이어들을 5m×5m의 해상도로 변환하여 전체 1,385,973개의 픽셀을 구성하였다.

의사결정 트리의 궁극적인 목적은 예측 정확도이며, 예측 정확성을 높이기 위해 트리 구축 과정에서 공간사건 발생 유무에 따라 사건 발생 환경 요인을 선택하며 트리를 구축한다. 이러한 특징은 확률 혹은 빈도를 기반으로 하는 기법들에서 인자들간 독립성 검정과 같은 상관성 분석을 별도로 요구하지 않는다. 즉 트리 구축 과정에서 사건 발생 유무에 대한 분류를 잘 수행할 수 있는 속성을 선별하여 노드가 성장한다. 또한 연속형 속성인 경우 분리가 잘 되는 구간을 선정하여 트리구축과정에서 연속형 데이터의 구간화가 수행된다. 이러 특징은 연속형 데이터의 임의 구간화로 인한 모델의 예측성능 저하를 방지할 수 있다. 따라서 이 논문에서는 연속형 속성자료를 사전에 구간화 없이 그대로 사용하였다.

TABLE 1. Configuration of dataset

Map Source	Thematic Layer	Type	Scale(resolution)
Airborne Image	Landslide	class	0.4m
Topographic Map	Slope	continuous	1:5,000
	Aspect	discrete	
	Curvature	continuous	
	Ridge	continuous	
Soil map	Texture	discrete	1:25,000
	Drainage	discrete	
	Material	discrete	
	Thickness	discrete	
Forest Map	Type	discrete	1:25,000
	Diameter	discrete	
	Density	discrete	
	Age	discrete	

2. 적용 및 결과

의사결정 트리는 Quinlan의 C4.5 알고리즘과 동일한 결과를 얻도록 알고리즘을 구현하여 트리를 구축하였다. C4.5는 Entropy 기반의 속성 선택과정을 수행하며, 연속형 속성을 사용하기 위해 Gain ratio를 이용한다. 또한 트리 구축 후 추가적으로 오분류율이 적은 트리를 얻기 위해 트리 전체를 순회하며 트리 부노드보다 자노드에서 오분류율이 높게 산정되는 경우를 찾아 자노드를 제거하는 Node-Collapsing 과정을 수행한다.

공간사건 데이터는 중요한 의미를 갖는 소수의 발생 클래스와 그렇지 않은 다수의 미발생 클래스 분포를 갖는다. 대부분의 분류기법들과 마찬가지로 의사결정 트리도 중요한 의미를 갖는 소수의 클래스가 모델 구축과정에서 무시되어 간결한 트리구조를 취하는 경향이 있으며, 이러한 경우 다수의 클래스로 일반화될 수 있다. 이러한 특징은 Provost and Domingos (2003), Liang and Yan(2006) 및 Alvarez *et al.*(2007)의 연구들에서 클래스 불균형 데이터에서 트리 일반화 과정을 적용하지 않은 큰 트리에서 보다 좋은 성능을 증명하였다. 따라서 이 논문에서의 트리구축은 앞선 과정을 통해 C4.5과 동일한 트리를 구축하였으며 별도의 전지과정을 거치지 않고 완전 성장토록 하

여 구축하였다.

이와 같은 방법으로 얻은 트리를 이용하여 적용한 리프노드를 등급 평가방법간의 성능을 비교하기 위해 산사태 발생 위치를 기반으로 2-fold 교차검정을 수행하였다. 이과정은 산사태 발생 위치를 랜덤하게 두 개의 그룹으로 이등분 한다. 첫 번째 단계에서는 첫 번째 그룹으로 모델을 만들고 두 번째 그룹으로 검증용, 두 번째 단계에서는 두 데이터 그룹의 역할을 바꿔 수행하여 그 결과의 평균으로 정확도를 계산한다. 이러한 교차검증 방법은 예측모델간의 평가에서 객관성으로 인해 Davis *et al.*(2006), Jiménez *et al.*(2009), Peng *et al.*(2009), Bonachea *et al.*(2009), Meusburger and Alewell(2009), Su and Cui(2009), Clerici *et al.*(2010), Rossi *et al.*(2010)의 연구에서 적용하였다.

성능 평가 척도는 누적이득차트를 이용하여 상대적 등급으로 표현된 예측 도면을 백분위 단위로 변환하여 교차평가를 위해 별도로 분류한 검증용 공간데이터의 위치에 해당하는 백분위 등급 순으로 정렬하여 이를 누적하여 그리며 2차원 그래프로 표현한다. 즉 백분위에 해당하는 단위 면적당 상위 분류 영역에 대한 누적 비율을 2차원 그래프로 묘사하기 위해 다음 식 (10)에 의해 백분위 등급이 결정된다.

$$Percentile = \frac{n_{upper\ rank} + n_{same\ rank} \times 0.5}{n_{total\ instance}} * 100 \quad (10)$$

리프노드 평가방법별 성능에 평가에서 M-estimate와 m-branch는 최적의 성능을 나타내는 파라미터 값을 실험에 의해 경험적으로 찾아야 한다. m-branch에서 상수 M의 변화에 따른 예측 정확치는 M 값이 500 이상에서 부터 누적이득 차트(Brandenburger and Furth, 2009)의 AUC(Area Under Curve)값이 약 86%에서 큰 변화를 보이지 않았으며 적용한 사례에서는 그 값이 8,000일 때 가장 높은 AUC를 얻을 수 있었다. 또한 M-estimate의

경우 mb 값이 10일 때 AUC가 84%에서 작은 편차를 보였으며 mb=20일 때 가장 높은 정확도를 보였다(그림 2).

최고의 예측성능에서의 파라미터를 적용한 M-estimate와 m-branch 더불어 Laplace에 대한 교차평가에 대한 성능 결과는 그림 3과 같이 누적이득 차트로 표현 하였다. 각 기법별 그래프의 하부영역으로 상대적인 비교에서, m-branch, Laplace 및 M-estimate 순으로 정확도를 보였다. 교차평가에 대한 결과는 그림 4와 같으며, 교차평가에서 사용한 산사태 발생 위치는 두 그룹은 사각형(□)으로 묘사된 LandslideSetA, 원(○)으로 묘사된 LandslideSetB로 표현되어 있다.

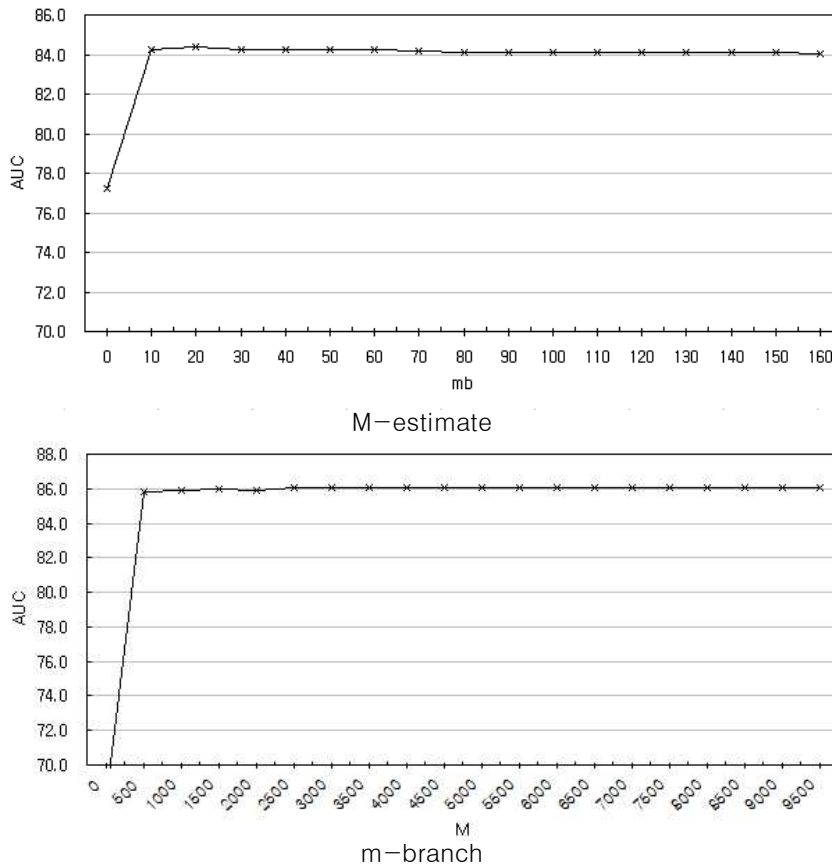


FIGURE 2. Trends of AUC values according to applied methods

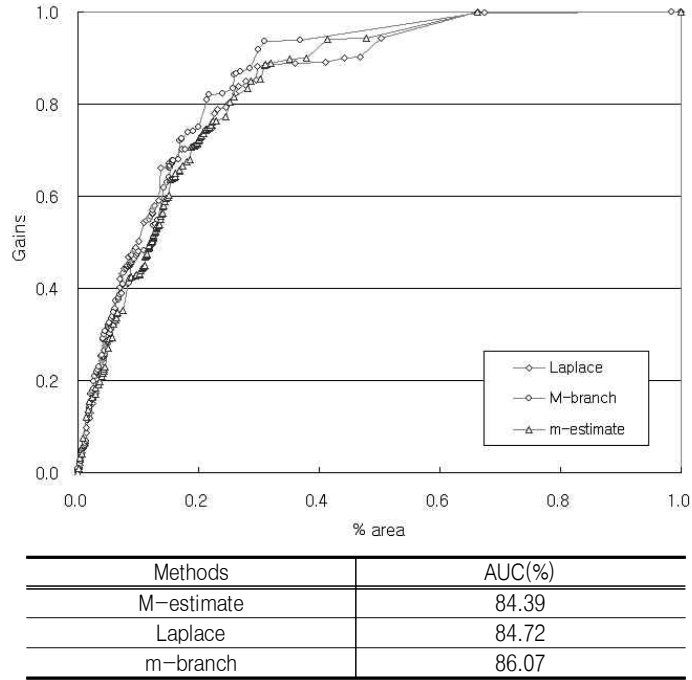


FIGURE 3. Cumulative gain chart of applied methods and AUC values of each curve

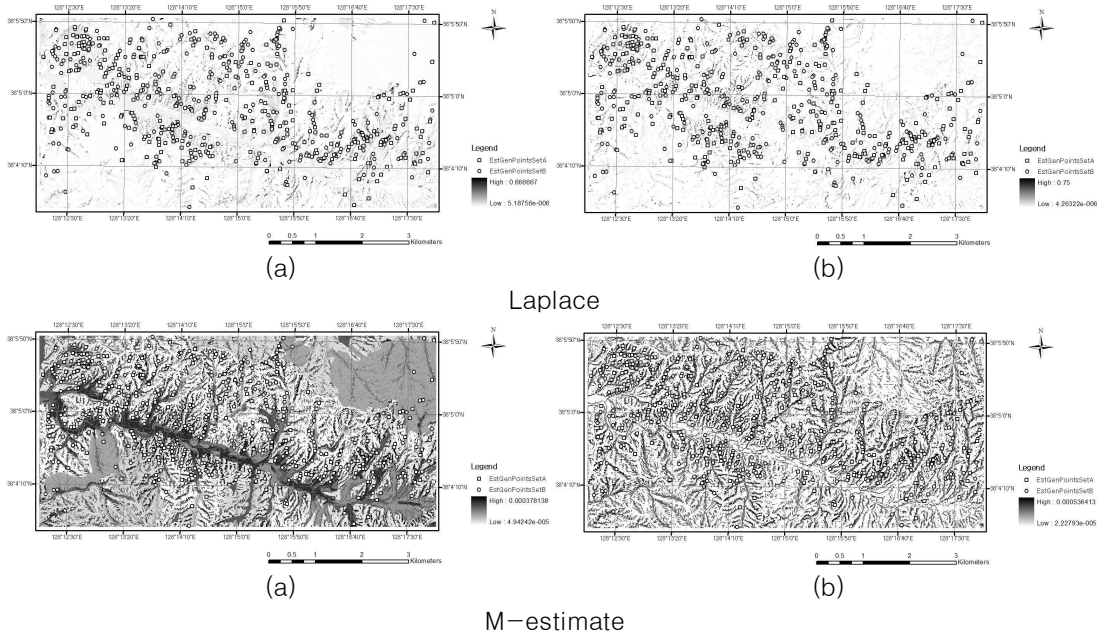


FIGURE. 4 First fold(a) and Second fold(b) results of the two fold cross validation according to the each method

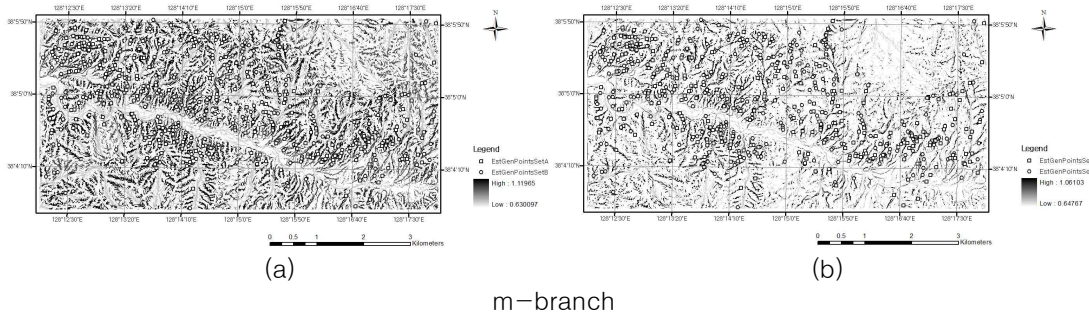


FIGURE. 4 Continued

결론 및 토의

의사결정 트리는 원인분석 능력 및 적용의 유연성으로 인해 다양한 예측분야에서 널리 사용되고 있으나 공간예측 분야에서는 그 활용이 활발하지 못했다. 이 논문에서는 의사결정 트리 알고리즘들 중에서 가장 널리 알려진 C4.5 알고리즘을 이용하여 공간사건 예측분야에 응용할 수 있도록 리프노드 평가방법의 적용방법과 더불어 이들에 대한 성능 평가를 수행하였다.

리프노드 평가방법은 의사결정 트리 구축 후 리프노드에 포함된 클래스 분포를 이용하여 상대적 등급을 산정하기 위한 방법이다. 트리는 순수한 분류를 지향하기 때문에 궁극적으로 리프노드에 클래스에서의 확률적 표현은 0 또는 1로 산정 된다. 또한 일반적인 리프노드 평가방법은 다수의 클래스가 출력 레이블이 되기 때문에 공간예측분야에 직접적으로 적용할 수 없었다. 이러한 특징으로 공간예측응용을 위해 타깃 클래스에 대한 분포비를 이용할 수 있도록 계산식을 변형하여 산사태 예측분석에 적용할 수 있었다.


이 논문에서 적용한 리프노드 평가방법의 사례적용 결과, 누적이득 차트의 AUC기준으로 M-estimate와 Laplace의 차이는 거의 유사했으며 m-branch가 가장 우수한 예측 결과를 도출하였다. 리프노드 평가방법인 Laplace는 주로 리프노드에 포함된 클래스 분포를 이용하여 계산되며 별도의 파라미터가 포함되어 있지 않아 빠른 계산결과를 유도할 수 있는 장점이 있었다. 그러나 리프노드의 클래스 분포를 이용하여 등급이 계산되기 때문에 트리의 구조에

따라 민감하게 반응할 가능성을 배제할 수 없다. M-estimate와 m-branch의 경우 최적의 예측정확도를 보이는 파라미터 탐색 과정이 필요하다. 특히 M-estimate의 경우 계산이 Laplace와 같이 리프노드의 클래스 분포를 이용하기 때문에 Laplace와 동일한 문제점이 내재될 수 있다. m-branch의 경우 리프노드 평가는 트리의 전체 노드의 클래스 분포를 이용한다. 이러한 특징은 앞선 두 방법보다 트리 구조에 덜 민감한 예측 결과를 도출할 수 있다는 장점이 있으며 보다 나은 예측 성능을 기대할 수 있다. 또한 적용 사례에서도 가장 좋은 성능을 나타내었다. 그러나 리프노드 등급을 계산하기 위해서는 트리의 각 노드마다 사건 발생에 대한 클래스 분포 값을 저장하고 있어야 한다.

이 논문에서는 공간예측 분야의 의사결정 트리는 공간예측 응용을 위해, 기존의 규칙평가 기법을 응용하여 적용할 수 있었다. 특히 의사결정 트리는 각 리프노드의 등급 값이 해당 조건과 일치하는 공간 좌표상에 매칭 될 수 있기 때문에 특정 위치에서의 원인분석을 수행할 수 있다. 따라서 이 논문에서 적용한 리프노드 평가방법은 적용한 산사태 사례뿐만 아니라 다양한 공간 예측 분야에서 예측 결과 도출 및 결과 해석에 활용할 수 있다.

감사의 글

이 연구는 한국지질자원연구원 주요사업 “광화대 멀티스케일 3D 지질모델링 구현” 과제로

수행되었습니다. 

REFERENCES

- Bonachea, J., J. Remondo, J.R.D. De Terán, A. González-Díez and A. Cendrero. 2009. Landslide risk models for decision making. *Risk Analysis* 29(11):1629-1643.
- Brandenburger, T. and A. Furth. 2009. Cumulative gains model quality metric. *Journal of Applied Mathematics and Decision Sciences* 2009:1-14.
- Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone. 1984. *Classification and Regression Trees*, Chapman & Hal, Wadsworth, Inc, New York.
- Casale, R., R. Fantechi and J.C. Flageolet. 1994. Temporal occurrence and forecasting of landslides in the European community. Final Report, European Community Programme Epoch. 957pp.
- Cestnik, B. 1990. Estimating probabilities: a crucial task in machine learning. *Proceedings of 9th European Conference on Artificial Intelligence* 1990, pp.147-149.
- Clerici, A., S. Perego, C. Tellini and P. Vescovi. 2010. Landslide failure and runout susceptibility in the upper T. Ceno valley (Northern Apennines, Italy). *Natural Hazards* 52(1):1-29.
- Cussents, J. 1993. Bayes and pseudo-bayes estimates of conditional probabilities and their reliabilities. *Proceedings of European Conference on Machine Learning*, pp.136-152.
- Davis, J.C., C.J. Chung and G.C. Ohlmacher. 2006. Two models for evaluating landslide hazards, *Computers & Geosciences* 32(8):1120-1127.
- Dikau, R., L. Schrott, D. Brunsden and M.L. Ibsen. 1996. *Landslide recognition: Identification, Movement and Causes*, John Wiley & Sons: Chichester, UK. pp.122-136.
- Ferri, C., P.A. Flach and J. Hernández-Orallo. 2003. Improving the AUC of probabilistic estimation trees. In: N. Lavrač *et al.* (Eds.) *Machine Learning: ECML*. Springer Berlin Heidelberg, pp.121-132.
- Jiménez-Perálvarez, J.D., C. Irigaray, R. El Hamdouni and J. Chacón. 2009. Building models for automatic landslide-susceptibility analysis, mapping and validation in ArcGIS. *Natural Hazards* 50(3):571-590.
- Jo, M.H. and Jo, Y.W. 2009. Developing forecast technique of landslide hazard area by integrating meteorological observation data and topographical data -a case study of Uljin area-. *Journal of the Korean Association of Geographic Information Studies* 12(2): 1-10 (조명희, 조윤원. 2009. 기상과 지형 자료를 통합한 산사태 위험지 예측 기법 개발 -울진지역을 대상으로-. *한국지리정보학회지* 12(2):1-10).
- Lee, J.D., S.H. Yeon, S.G. Kim and H.C. Lee. 2002. The application of GIS for the prediction of landslide - potential area. *Journal of the Korean Association of Geographic Information Studies* 5(1):38-47 (이진덕, 연상호, 김성길, 이호찬).

2002. 산사태의 발생가능지 예측을 위한 GIS의 적용. 한국지리정보학회지 5(1):38-47.
- Liang, H. and Y. Yan. 2006. Improve decision trees for probability-based ranking by lazy learners. Tools with Artificial Intelligence, 2006. ICTAI'06. 18th IEEE International Conference on IEEE, pp.427-435.
- Meusburger, K. and C. Alewell. 2009. On the influence of temporal change on the validity of landslide susceptibility maps. Natural Hazards and Earth System Science 9(4):1495-1507.
- Pal, M. and P.M. Mather. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment 86:554-556.
- Park, J.S., K.T. Kim and Y.S. Choi. 2012. Landslide risk assessment using HyGIS-landslide. Journal of the Korean Association of Geographic Information Studies 15(1):119-132 (박정술, 김경탁, 최윤석. 2012. HyGIS-Landslide를 이용한 산사태 발생 위험도 평가. 한국지리정보학회지 15(1):119-132).
- Peng, W.F., C.L. Wang, S.T. Chen and S.T. Lee. 2009. Incorporating the effects of topographic amplification and sliding areas in the modeling of earthquake-induced landslide hazards, using the cumulative displacement method. Computers and Geosciences 35(5):946-966.
- Provost, F.J. and P. Domingos. 2003. Tree induction for probability-based ranking. Machine Learning 52:199-215.
- Quinlan, J.R. 1993. Programs for Machine Learning. Morgan Kaufmann, 302pp.
- Rossi, M., F. Guzzetti, P. Reichenbach, A.C. Mondini and S. Peruccacci. 2010. Optimal landslide susceptibility zonation based on multiple forecasts. Geomorphology 114(3):129-142.
- Su, F. and P. Cui. 2009. GIS-based susceptibility mapping and zonation of debris flows caused by Wenchuan earthquake. Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on IEEE, pp.1-5.
- Wu, X., V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D.J. Hand and D. Steinberg. 2008. Top 10 algorithms in data mining. Knowledge and Information Systems 14(1):1-37.
- Yeon, Y.K. 2011. Evaluation and analysis of Gwangwon-do landslide susceptibility using logistic regression. Journal of the Korean Association of Geographic Information Studies 14(4):116-127 (연영광. 2011. 로지스틱 회귀분석 기법을 이용한 강원도 산사태 취약성 평가 및 분석. 한국지리정보학회지 14(4):116-127).
- Yeon, Y.K., J.G. Han and K.H. Ryu. 2010. Landslide susceptibility mapping in Injae, Korea, using a decision tree. Engineering Geology 116(3):274-283.
- Zadrozny, B. and C. Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. Proceedings of 18th International Conference on Machine Learning. 2001, pp.609-616. 