Journal of Wetlands Research
Vol. 21, No. 2, May 2019, pp. 157-162

ISSN 1229-6031 (Print) / ISSN 2384-0056 (Online)
DOI https://doi.org/10.17663/JWR.2019.21.2.157

랜덤포레스트를 이용한 대설피해액에 대한 범주형 예측 및 개선방안 검토

이형주·정건희†

호서대학교 토목공학과

Categorical Prediction and Improvement Plan of Snow Damage Estimation using Random Forest

Lee Hyeong Joo·Gunhui Chung†

Department of Civil Engineering, Hoseo university, Korea
(Received : 22 April 2019, Revised: 16 May 2019, Accepted: 16 May 2019)

요 약

최근 세계적인 기상이변으로 이례적인 대설과 한파의 발생 빈도가 증가하고 있다. 이로 인해 대설피해 저감에 대한 연구가 진행되고 있으나, 우리나라는 시군구 별 과거 피해이력이 적고, 피해 발생지역과 관측소 간의 거리가 멀어 정확한 피해예측이 어려운 상황이다. 따라서 본 연구에서는 대설피해에 영향을 미친다고 생각되는 인자들의 데이터를 수집한 뒤 랜덤포레스트 모형의 설명변수로 설정하여 추정되는 대설피해액을 범주형태로 예측하고자 하였다. 또한 설명변수 중 취약성 분석을 통해 도출된 취약성 지수를 설명변수로 이용함으로써 지역적 특색과 특성을 반영하였다. 지금은 과거 피해 자료의 부족, 비닐하우스 설계 기준의 변화 등으로 인해 예측 정확도가 높지 않지만, 피해가 발생한 지역의 정확한 기상자료가 확보되고, 변수로 사용한 데이터의 업데이트가 진행된다면 본 연구 결과의 정확도 향상과 재난 발생 전 피해규모 및 범위에 대한 신속한 예측을 통해 대비차원의 재난관리 대처능력이 향상될 것으로 기대된다.

핵심용어 : 대설피해, 랜덤포레스트, 피해예측

Abstract

Recently, the occurrence of unusual heavy snow and cold are increasing due to the unusual global climate change. In particular, the temperature dropped to minus 69 degrees Celsius in the United States on January 8, 2018. In Korea, on February 17, 2014, the auditorium building in Gyeongju Mauna Resort was collapsed due to the heavy snowfall. Because of the tragic accident many studies on the reduction of snow damage is being conducted, but it is difficult to predict the exact damage due to the lack of historical damage data, and uncertainty of meteorological data due to the long distance between the damaged area and the observatory. Therefore, in this study, available data were collected from factors that are thought to be corresponding to snow damage, and the amount of snow damage was estimated categorically using a random forest. At present, the prediction accuracy was not sufficient due to lack of historical damage data and changes of the design code for green houses. However, if accurate weather data are obtained in the affected areas. the accuracy of estimates would increase enough for being used for be the degree preparedness of disaster management.

Key words : snow damage, estimation of damage, random forest

1. 서 론

최근 전 세계적으로 지구온난화에 따른 급격한 기후변화로 인해 전 세계적으로 해수면 상승, 한파, 폭설, 가뭄, 국지적 집중호우, 홍수 등과 같은 자연재해가 다양하게 존재하며, 이러한 자연재해의 강도와 빈도가 점차 증가하고, 사회·경제적으로 심각한 피해를 주고 있다(Ha et al, 2007). 우리나라 또한 2014년 2월 17일 오후 9시 11분 무렵, 경상북도 경주시 양남면 신대리에 위치한 마우나오션리조트의 강당 건물이 폭설로 무너져 신입생 오리엔테이션을 진행 중이던 부산외국어대학교 학생들이 매몰되는 사고가 발생했다. 이로 인해 사망 10명, 부상자 204명이 발생한 대형 폭설 피해가 발생하였다. 최근 발생한 대설피해로는 2018년 2월 제주도에서 발생한 41년 만의 폭설로 인해 제주 지역의 도로가 마비되고 제주공항 활주로가 폐쇄되었던 것이 대표적이다. 국민안전처에 따르면 최근 10년 연 평균 자연재해 피해액이 약 1.4조 원이며, 이 중 대설에 의한 피해액은 약 1,700억 원으로 전체 자연재해 피해액의 약 12%를

† To whom correspondence should be addressed.
Department of Civil Engineering, Hoseo university, Korea
E-mail: gunhuic@gmail.com

차지하고 있다(MPSS, 2014). 우리나라의 대설의 경우 지역적으로 집중되어 나타나는 것이 특징이다. 영동지역에서는 5~10 cm이상 혹은 30 cm이상의 눈이 쌓이는 모습을 자주 볼 수 있다. 그러나 남부지역은 많은 양의 눈이 내리지 않으므로 1~2 cm적설에도 교통두절 등 간접적인 피해가 많이 발생하고 있다(Kwon et al, 2016). 이처럼, 대설과 같이 지역별 격차가 심한 경우에는 각 지역의 대설피해의 특성을 충분히 고려하여 연구를 진행할 필요가 있다고 판단되었다. 현재 대설피해 저감에 대한 연구가 다수 진행되고 있으나, 다른 재해사상에 비해 과거 피해이력이 현저히 적고, 피해 발생 지역과 관측소 사이의 거리가 멀어 정확한 피해예측이 어려운 상황이다. 이러한 대설피해에 대비하기 위해서는 재해 발생 전 피해규모에 대한 신속한 예측을 통해 피해를 방지하는 대책수립을 위한 노력과 연구가 필요하다.

기존 대설피해에 대한 연구를 살펴보면, Kwon and Chung(2017)은 회귀분석을 통하여 강원도지역의 대설피해액 추정 모형을 개발하였고, Oh and Chung(2017)은 대설피해가 자주 발생하는 강원도, 전라도, 충청도 지역을 대상을 대설피해 예측 모형을 개발하고 모형의 적용성을 검토하였다. Oh and Chung(2018)은 선행연구의 연구 결과를 인용하여 취약성 등급별 대설피해 예측함수를 개발하고 적용성을 검토하였다.

이와 같이 재해발생 전에 대설피해를 예측하고자 하는 선행 연구들은 다수 존재하지만, 현재 대설피해를 정확하게 예측할 수 있는 대책이나 방안은 마땅히 없는 실정이다. 또한 지역별 특성을 제대로 고려하지 않고, 특정 시도만을 대상으로 했다는 점에서 전국 적용성의 한계가 존재하였고, 시군구 별 과거 피해이력이 적고, 피해발생지역과 관측소 간의 거리가 멀어 수치형으로 대설피해액을 예측하기에는 어렵다는 점도 해결해야 할 문제이다. 따라서 본 연구에서는 전국의 대설피해를 사전에 대비하기 위해 대설피해액을 범주형태로 구분하고, 대설피해에 영향을 미친다고 생각되는 설명변수 선정, 관련 자료 데이터를 수집하고 랜덤 포레스트를 이용하여 대설피해액을 범주형태로 예측하였다.

2. 방법론

2.1 랜덤포레스트 모형

랜덤 포레스트(Random Forest)는 Breiman(2001)에 의해 제안된 앙상블(ensemble) 기반 모형으로 의사결정나무 모형에 배깅(Bagging)의 기본 원리와 임의성을 더한 형태이다. 배깅의 상관된 예측 값에 대한 문제점을 랜덤 포레스트에서는 설명변수를 임의로 선택하는 과정을 추가하여 문제를 해결하고자 하였다(Choi et al., 2017). 랜덤 포레스트는 일반적으로 변수의 개수가 m개이면 각 분할에서 랜덤으로 m/3개의 변수를 선택하여 트리를 만든다(Choi et al, 2017).

랜덤 포레스트에서 변수의 중요성은 훈련 데이터에서 j번째

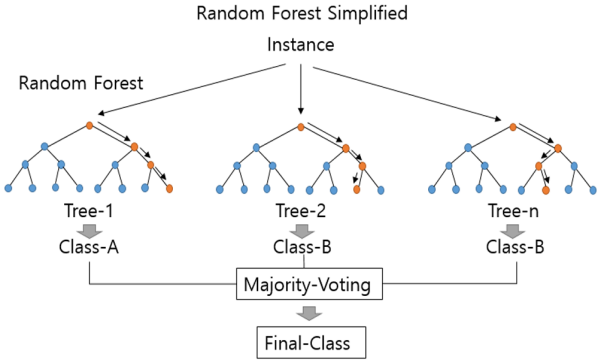

Fig. 1. Random Forest Structure

특징의 값을 치환하고, 다시 데이터에 대하여 OOB-오차(Out-of-bag error)를 구하여 치환 전의 OOB-오차의 차이를 평균하여 정의한다(Choi et al., 2017). 큰 중요도 점수를 가지는 변수는 작은 값을 갖는 변수보다 높은 순위의 중요성을 갖게 된다(Choi et al., 2017). 특히 랜덤 포레스트는 데이터와 변수를 샘플링하여 서로 조금씩 다른 나무들로 구성되었기 때문에 각 나무들의 예측 값은 비 상관하게 되어 일반화 성능을 향상시킨다(Choi et al., 2017). Fig. 1은 랜덤 포레스트 모형의 구조를 나타낸 것이다.

2.2 P-S-R 구조체계

P-S-R 구조는 OECD(Organisation of Economic Cooperation and Development)에서 1993년에 개발하여 국제기구나 각국의 지표설정에 주로 활용되고 있다. P-S-R 구조는 현치수특성 평가문제를 인과관계로 분석할 수 있으며, 사회, 경제 및 여타 쟁점사항간의 상호 연관된 관점에서의 파악이 가능하다. P-S-R 구조를 바탕으로 한 압력지표(PF), 현상지표(SF), 대책지표(RF) 등 3개 지표의 평가를 위한 세부 평가항목의 선정에 있어서, 대설피해 취약성을 대표할 수 있는 평가기준을 선정하는 것이 중요하지만, 앞서 언급한 바와 같이 모든 대상지역별로 분석 및 적용이 가능한 데이터를 수집하는 것은 쉽지 않다. Fig. 2는 P-S-R 구조체계의 이해를 돕기 위해 흐름을 나타낸 것이다.

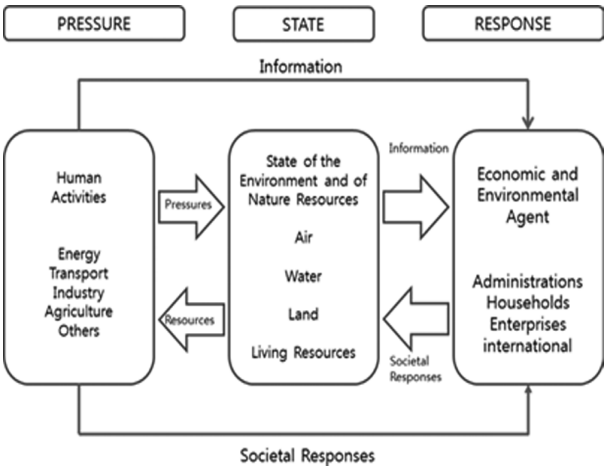

Fig. 2. P-S-R Structure System

2.3 엔트로피 이론

엔트로피 방법은 정보이론을 활용하여 각 속성의 가중치를 결정하는 방법이다(Shannon and Weaver, 1949). 특정 속성에 포함된 자료들이 상이한 정도가 작을수록, 즉, 분산이 작을수록 불확실성인 엔트로피가 증가하며, 충분한 정보가 전달되고 있지 않다고 가정하여 작은 가중치를 적용한다. 이는 확보된 자료들의 분포만을 고려하여 가중치를 산정하기 때문에 개인의 주관이 개입되지 않는다는 장점이 있다. 속성 i ∈ I, i = 1,2,..., m에 세부지표 j ∈ J, j = 1,2,..., n개가 조사되어, 다음과 같은 속성정보 매트릭스(R)를 Fig. 3과 같이 구성한다고 가정하였을 때, 각 속성별 세부지표 정보는 표준화를 통해 0~1사이의 값으로 변환한다.

$$R=\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{bmatrix}$$

Fig. 3. Property Information Matrix

식 (1)은 0~1 사이의 값으로 변환하는 표준화 방법이다.

$$r_{ij}=\frac{x_{ij}}{\sum_{j=1}^{n} x_{ij}}, i=1,2,.....,m; j=1,2,...n \quad (1)$$

각 속성 i의 엔트로피(H_i)는 식 (2)으로 계산한다.

$$H_i=-k\sum_{j=1}^{n} r_{ij}\ln r_{ij}, i=1,2,....,m \quad (2)$$

여기서, $k=1/\ln$을 나타낸다.

속성 i의 엔트로피 가중치(w_i)는 식 (3)과 같으며, 각 속성에 포함된 자료들의 다양성을 나타내는 척도인 $d_i=1-H_i$를 계산하여 결정한다.

$$w_i=\frac{d_i}{\sum_{i=1}^{m} d_i} \quad (3)$$

여기서, $0 \le w_i \le 1$이고, $\sum_{i=1}^{m} w_i = 1$이며, 엔트로피 가중치 추정 방법의 흐름을 이해하기 쉽게 Fig. 4에 나타내었다.

3. 분석 방법 및 자료 구축

3.1 대설피해이력 조사

행정안전부에서 매년 발간하는 재해연보는 피해 원인·지역·수계·기간 등의 범주로 분류되어 있고, 피해액의 경우

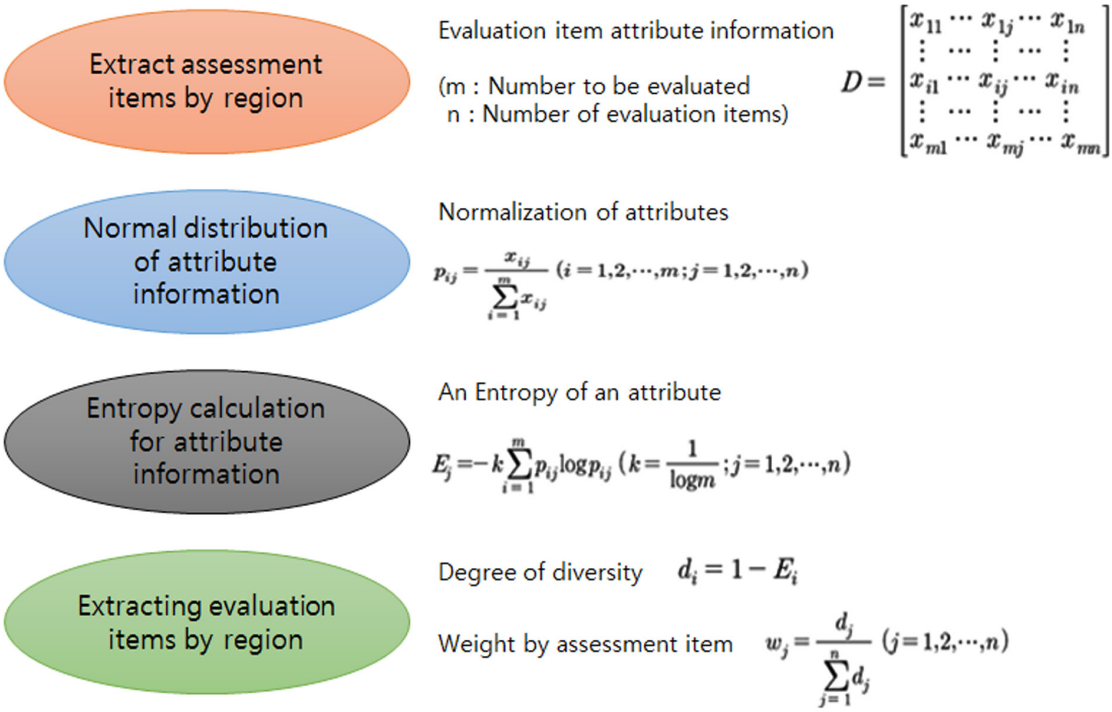

Fig. 4. Entropy Weight Estimation Process

공공시설 피해와 사유시설 피해로 구분되어 있다. 따라서 대설피해이력은 재해 관련 자료의 기초 자료가 되고 있는 재해연보의 자료를 이용하여 조사하였다. 본 연구에서는 공공시설 피해액과 사유시설 피해액을 구분 짓지 않고 재해연보에 집계된 대설피해이력과 종속변수로 활용할 대설피해액 자료(1994년부터 2016년까지)를 수집하였다.

3.2 설명변수 데이터 구축

랜덤 포레스트 방법에 사용될 설명변수로 대설피해의 핵심적인 기상요소와 각 지역의 특색과 특성을 반영할 수 있는 사회·경제적 요소를 고려하였다. 기상요소로 최심신적설량, 최저기온, 일평균 일조시간을 고려하였고, 사회·경제적 요소로는 취약성 지수를 고려하였다. 또한 피해발생지역의 과거 기상 특색을 반영하기 위해 대설피해횟수와 연간 평균온도 0℃ 이하인 날 수를 고려하였다.

본 연구에서 사용되고 있는 기상자료는 종관기상관측장비(Automated Synoptic Observing System, ASOS)를 사용하여 데이터를 구축하였다. 그 이유는 기상청관리자가 직접 기상 요소를 관측하는 유인 기상측정 시스템이기 때문에 무인으로 운영되는 방재기상관측장비(Automatic Weather System, AWS)보다 정확한 자료를 얻을 수 있다고 판단되었기 때문이다. Table 1은 모형에서 사용하는 변수 명칭과 설명을 나타내었다.

3.3 취약성 분석

우리나라의 대설의 경우 지역적으로 집중되어 나타나는 것이 특징이다. 영동지역에서는 5~10 cm 이상 혹은 30 cm 이상의 눈이 쌓이는 모습을 자주 볼 수 있다. 그러나 남부지역은 많은 양의 눈이 내리지 않으므로 1~2 cm 적설에도 교통두절 등 간접적인 피해가 많이 발생하고 있다(Kwon et al., 2016). 이처럼, 대설과 같이 지역별 격차가 심한 경우에는 각 지역의 대설피해의 특성을 충분히 고려하여 설명변수를 구축할 필요가 있다고 판단되었다.

본 연구에서는 랜덤포레스트 방법을 사용하기 위해 종속변수 및 설명변수 데이터 구축이 필수적이다. 앞서 언급한 바와 같이 사회경제적 요소로 대설피해횟수와 연간 평균온도 0℃ 이하인 날 수를 고려하였다. 그러나 이 두 가지 설명변수로 229개 시군구의 특성과 지역적 특색을 반영하기에 무리가 있다고 판단되어 P-S-R 구성 체계를 이용하여 지표를 선정하고, 엔트로피(Entropy) 가중치 추정법을 이용하여 지표의 가중치를 산정하였다. 또한 취약성 분석을 실시하여 도출된 취약성 지수를 설명변수로 활용하였다. Table 2는 취약성 분석의 입력 자료를 나타냈고, Fig. 5는 대설 취약성 지수 계산 결과를 나타내었다.

Table 1. Variable Name and Description

Name	Description
D1	Historical Number of Snowfall damage
D2	Number of days when the average temperature is below zero
D3	Daily fresh snow depth during the disaster periods
D4	Daily lowest temperature
D5	Annual average daylight time
D6	Snow vulnerability Index

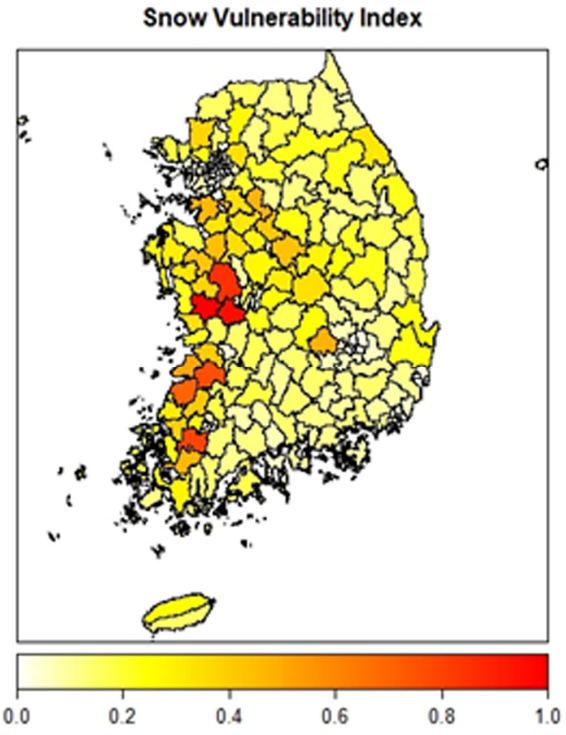

Fig. 5. Calculated Snow Vulnerability Index

Table 2. Input Data for Snow Vulnerability Index

Division	Data	Unit (Data Period)
P(Pressure Index)	Altitude deviation	m
	Day maximum fresh snow depth	cm
	Number of days when the average temperature is below zero	Days/Year
S(Status Index)	Average amount of snow damage in the past year	1,000 won(1996~2016)
	Number of snowfall damage	Number of times(1996~2016)
R(Response Index)	Number of snow-removal equipment	The number
	The density of public officials	Number of people/km^2
	Forest and land area	km^2
	Annual average daylight time	Time/Days

3.4 대설피해액 범주형태 변환

본 연구에서는 수치형으로 집계되어 제공되는 재해연보상 대설피해액을 범주 형태로 변환하여 피해액을 예측하였다. 본 과정을 진행하기 위해서는 수치형으로 집계된 피해액을 범주 형태로 변환하는 과정이 필요하다. 피해액의 범주는 총 5가지로 분류하고 각 범주마다 피해액의 범위를 지정하였다. A범주는 10억 원~100억 원, B범주는 1억 원~10억 원, C범주는 천만 원~1억 원, D범주는 백만 원~천만 원 마지막으로 E범주는 십만 원~백만 원으로 임의적으로 범주를 지정하여 피해액이 어느 범주에 포함되는지 예측하였다. 각 범주의 범위와 범주에 속하는 피해이력의 개수를 Table 3에 정리하였다.

Table 3. Historical Number of Snow Damage in terms of Damage Categories

Category	Count	Heavy snow damage(×1,000 KRW)
A	55	1,000,000~10,000,000
B	171	100,000~1,000,000
C	336	10,000~100,000
D	311	1,000~10,000
E	120	100~1,000

4. 결과 및 고찰

수치 형태로 제공되는 대설피해액을 보다 정확성 높은 추정을 위해 피해액을 범주 형태로 구분짓고, 랜덤 포레스트 모형을 사용하여 예측피해액을 추정하였다. 피해액이 어떤 범주에 속하는지 추정하고 모형의 오류율을 평가하였다. 설명변수의 개수는 총 6개이고 550개의 의사결정나무를 만들어 모형을 구축하였다. 랜덤 포레스트 모형에서 사용된 설명변수의 중요도는 Fig. 6과 같으며, 가장 중요한 변수는 D3인 최심신적설량인 것으로 나타났다.

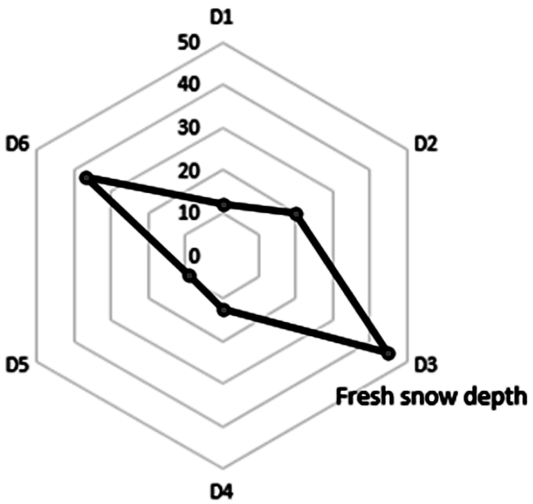

Fig. 6. Importance of Variables

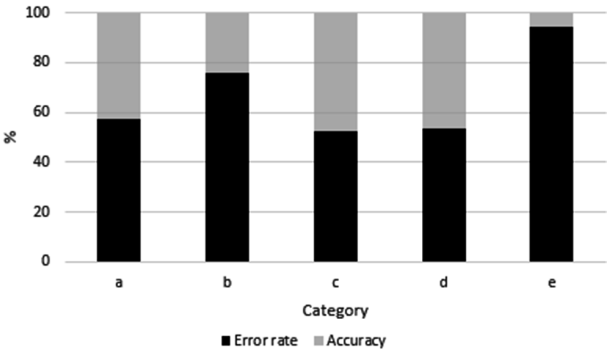

Fig. 7. Accuracy and Error Rate

본 연구에서는 총 표본의 개수를 100%로 봤을 때 학습을 위한 자료로 70%를 활용하고 예측을 위한 자료로 30%를 활용하였다. 이와 같은 과정을 수행하는 이유는 전체 자료를 모두 포함하여 정확도 검증을 할 경우 정확도가 항상 높게 나오기 때문에 정확도 검증을 수행하는 의미가 없다. 따라서 본 연구에서 개발한 모형은 1994년부터 2016년까지 발생한 대설피해이력 중 최심신적설심이 0인 표본을 제외한 994개 대설피해이력을 평가용 구간(334개)과 학습용 구간(660개)으로 분류하고 랜덤 포레스트 모형 구축 후 정확도 검증 결과 62.13%의 오류율을 나타냈다. 각 범주의 정확도와 오류율을 Fig. 7에 나타냈다.

5. 결 론

본 연구에서는 국내에서 대두되고 있는 자연재난인 대설피해를 대상으로 랜덤 포레스트 모형을 적용하여 범주 형태로 대설피해액을 예측하였다. 현재 대설피해 예측 및 저감에 대한 연구가 다수 진행되고 있으나, 대설의 경우 시·군·구 별 과거 피해이력이 적고, 피해지역과 관측소 사이의 거리가 멀기 때문에 데이터 구축에서 많은 어려움이 있다. 따라서 수치형으로 대설피해액을 예측하기에는 어려움이 있다고 판단되어 랜덤 포레스트 모형을 이용하여 범주 형태로 피해액을 예측하였다. 종속변수는 대설피해이력 기간에 발생한 대설피해액으로 선정하였고, 설명변수로 직접적인 영향을 미치는 기상요소와 지역적 특색을 반영할 수 있는 사회·경제적 요소를 고려하였다.

우선 행정안전부에서 매년 발간하는 재해연보를 이용하여 1994년부터 2016년까지 발생한 대설피해이력과 피해이력 기간의 대설피해액을 조사하고 대설피해에 영향을 미친다고 생각되는 설명변수 데이터를 수집하였다. 종속변수는 대설피해액으로 선정하였고, 설명변수로는 대설피해횟수, 연간 평균온도 0℃ 이하인 날 수, 최심신적설, 최저기온, 일조시간, 취약성 지수를 고려하였다. 특히 기상요소만으로는 229개 시군구의 차별화가 어려워 각 시·군·구의 특성과 지역적 특색을 반영할 수 있는 대설 취약성 지수를 계산하여 설명변수로 사용하였다.

랜덤 포레스트 모형을 이용하기 위해 수치형으로 제공되는 대설피해액을 임의로 범주를 지정하여 5가지 범주에 대

설피해액을 포함시키고 모형을 구축하였다.

1994년부터 2016년까지 발생한 대설피해이력 중 최심신적설심이 0 cm인 표본을 제외한 994개 대설피해이력을 평가용 구간과 학습용 구간으로 분류하여 모형을 구축하여 변수중요도 검사를 실시한 결과 설명변수 중 최심신적설 변수의 중요도가 가장 중요한 변수임을 나타냈다. 이를 이용하여 향후 연구에서 설명변수 설정 시에 대설피해에 영향력이 높은 변수를 설정하는 방향성을 제시하였다. 또한 정확도 검증 결과 62.13%의 오류율을 나타냈다. 낮은 오류율은 보이는 이유는 2010년대에 발생한 대설 피해 자료와 1990년대, 2000년대 초반의 대설피해액을 비교한 결과 같은 적설심에도 대설피해액의 차이가 크다는 것을 확인하였다. 과거 1990년대와 2000년대 초반의 대설 피해액은 대설에 가장 취약한 비닐하우스의 설계기준 변화, 농업 기술의 발달 등 외부요인의 변화로 인해 자료의 동질성을 확보하는 것이 매우 어렵다. 또한 대설피해가 발생했음에도 불구하고 적설심이나, 온도 등 기상자료 누락으로 인한 자료의 불확실성 또한 존재한다. 향후 과거 자료와 최근 10년 자료의 동질성을 확보하고 기상요소와 사회·경제적 요소 데이터의 최신화, 과거 자료의 편의를 보정한다면 보다 높은 정확성을 보일 것으로 기대된다.

감사의 글

본 연구는 행정안전부 재난예측및저감연구개발사업의 지원을 받아 수행된 연구임 (MOIS-재난-2015-05)

References

Breiman, L, (2001). Random forests machine learning, 45(1), pp. 5-32. DOI : https://doi.org/10.1023/A:1010933404324

Choi, CH, Park, KH, Park, HK, Lee, MG, Kim, JS, Kim, HS, (2017). Development of heavy rain damage prediction function for public facility using machine learning, J. of Korean Society of Hazard Mitigation, 17(6), pp. 443-450. [Korean Literature]. DOI : https://doi.org/10.9798/KOSHAM.2017.17.6.443

Ha, R, Shin, HJ, Kim, SJ, (2007). Proposal of prediction technique for future vegetation information by climate change using satellite image, J. of Korean Association of Geographic Information Studies, 10(3), pp. 58-69. [Korean Literature] DOI : http://www.koreascience.or.kr/article/JAKO200721761942397.page

Kwon, SH, Park, HS, Chung, GH, (2016). Analysis of snow vulnerability and adaptation policy for heavy snow, J. of Korean Society of Hazard Mitigation, 16(2), pp. 363-368. [Korean Literature]. DOI : https://doi.org/10.9798/KOSHAM.2016.16.2.363

Kwon, SH, Chung, GH, (2017). Estimation of snow damages using multiple regression model: The Case of Gangwon Province, J. of the Korean Society of Civil Engineers, 37(1), pp. 61-72. [Korean Literature]. DOI: https://doi.org/10.9798/KOSHAM.2016.16.2.437

MPSS(Ministry of Public Safety and Security)(2014). The 2013 Annual Natural Disaster report, Ministry of Public Safety and Security. [Korean Literature]

Oh, YR, Chung, GH, (2017). Estimation of snow damage and proposal of snow damage threshold based on historical disaster data, J. of the Korean Society of Civil Engineers, 37(2), pp. 325-331. [Korean Literature]. DOI : https://doi.org/10.12652/ksce.2017.37.2.0325

Oh, YR, Chung, GH, (2018). Multiple regression models of snow damage prediction according to the snow damage vulnerability groups Korea, J. of Korean Society of Hazard Mitigation, 18(2), pp. 355-359. [Korean Literature]. DOI: https://doi.org/10.9798/KOSHAM.2018.18.2.355

Shannon, CE, Weaver, W, (1949). The mathematical theory of communication, The University of Illinois Press, Urbana, U.S.A