J. Korean Soc. Hazard Mitig.
Vol. 19, No. 6 (Nov. 2019), pp.115~127
https://doi.org/10.9798/KOSHAM.2019.19.6.115

ISSN 1738-2424(Print)
ISSN 2287-6723(Online)
www.kosham.or.kr

풍수해방재

머신러닝 기반의 호우피해 발생확률 예측 모형 개발

Developing a Prediction Model (Heavy Rain Damage Occurrence Probability) Based on Machine Learning

김종성* · 이준형** · 김동현*** · 최창현**** · 이명진***** · 김형수******
Kim, Jongsung*, Lee, Junhyeong**, Kim Donghyun***, Choi, Changhyun****, Lee, Myungjin*****, and Kim, Hung Soo******

Abstract

In this study, a prediction model (heavy rain damage occurrence probability or PM-HDOP) was developed for a metropolitan area. The heavy rain damage and rainfall data were collected as dependent and independent variables, respectively. The dataset was divided into training (2005-2016) and test sections (2017). We developed the PM-HDOP using machine learning methods such as logistic regression, artificial neural network, bagging, random forest, and boosting to predict the occurrence of nonlinear natural disasters. An architectural model with the best performance was selected, and the PM-HDOP was subsequently used to predict the probability of occurrence. As a result, a boosting scheme showed the best performance in Gyeonggi-do and Seoul, and a bagging scheme showed the best performance in Incheon. If the results of this study are used to predict the occurrence of heavy rain damage, which is not currently being serviced in Korea, it is possible to effectively reduce the damages.

Key words : Machine Learning, Disaster Management, Natural Disaster, PM-HDOP

요 지

본 연구에서는 수도권 지역을 대상으로 사전에 피해 발생 여부를 파악하기 위하여 머신러닝 기반의 호우피해 발생확률 예측 모형을 개발하였다. 종속변수로써 활용하는 호우피해 자료와 독립변수로써 활용하는 강우자료를 수집하였고, 학습기간(2005-2016)과 평가기간(2017)으로 자료를 구분하였다. 로지스틱 회귀모형, 인공신경망, 배깅, 랜덤포레스트, 부스팅 등의 머신러닝 기법들을 적용하여 모형을 개발하였다. 평가기간의 자료를 이용하여 각 모형들에 대해 피해 발생 여부를 예측하고, F1-Score를 통해 성능이 가장 우수한 모형을 선별하였다. 그 결과 경기도 지역과 서울 지역에서는 부스팅이 가장 우수한 성능을 보였고, 인천 지역에서는 배깅이 가장 우수한 성능을 나타냈다. 본 연구 결과를 활용하여 현재 국내에서 서비스 되고 있지 않는 호우피해 발생 예측에 대한 서비스가 이루어진다면, 사전 대비 차원의 재난관리를 통해 효과적으로 피해를 저감할 수 있을 것이다.

핵심용어 : 머신러닝, 재난관리, 자연재난, 호우피해 발생확률 예측 모형

*정회원, 인하대학교 토목공학과 박사과정(E-mail: kjjs0308@naver.com)
Member, Doctor's Course, Department of Civil Engineering, Inha University
**정회원, 인하대학교 토목공학과 석사과정
Member, Master's Course, Department of Civil Engineering, Inha University
***정회원, 인하대학교 토목공학과 박사
Member, Ph.D., Candidate. Department of Civil Engineering, Inha University
****정회원, 인하대학교 토목공학과 박사과정
Member, Doctor's Course, Department of Civil Engineering, Inha University
*****정회원, 인하대학교 토목공학과 박사과정
Member, Doctoer's Course, Department of Civil Engineering, Inha University
******교신저자, 정회원, 인하대학교 사회인프라공학과 교수(Tel: +82-32-874-0069, Fax: +82-32-876-9787, E-mail: sookim@inha.ac.kr)
Corresponding Author, Member, Professor, Department of Civil Engineering, Inha University

1. 서 론

최근 세계적인 기후변화로 인해 국지성 집중호우, 태풍 등의 자연재난이 빈번하게 발생하고 있으며, 이로 인해 피해의 규모도 점점 대형화 되고 있다. 국내에서도 연평균 약 3천 4백억 원의 피해가 발생하고 있으며, 그 중 호우로 인한 피해는 약 1천 5백억 원의 피해를 차지하고 있다(MOIS, 2018). 현재 우리나라의 경우 기상청에서 발표하는 기상특보를 통해 예측되는 자연재난에 대해 파악할 수 있지만, 이는 피해가 아닌 기상정보에 대한 예측 정보이기 때문에 피해 발생 여부는 확인하기 어렵다. 행정구역별로 자연재난으로 인한 피해 발생 여부를 예보할 수 있다면 재난관리 담당자는 대비 차원의 다양한 재난관리를 수행할 수 있을 것이며, 국민들은 스스로 자발적으로 대책을 마련함으로써 피해를 최소화할 수 있을 것이다. 또한 기존 기상 특보는 피해의 개념을 직접적으로 언급하지 않기 때문에, 심각성을 깨닫지 못하고 이를 무시하여 인명피해가 발생한 사례가 수차례 존재했다. 자연재난으로 인한 피해 발생 여부를 예보하게 된다면 기존 기상특보 보다 피해의 개념을 직접적으로 언급하기 때문에 예보에 대한 심각성을 깨닫지 못해 발생하는 피해사례를 줄여 줄 수 있을 것이다.

기후변화가 심화됨에 따라 기상정보를 예측하기 위한 연구가 많이 수행되었다. 먼저 기후변화로 인한 자연재난을 파악하기 위해 장기간 기상정보를 예측한 연구사례를 살펴보면 주로 수치예보자료를 이용하여 다양한 모형에 적용함으로써 장기간 기상정보를 예측하였다(Kannan et al., 2010; Abbot and Marohasy, 2012; Mekanik et al., 2013; Kim et al., 2013). 장기간 예측의 경우 불확실성으로 인해 정확도를 보장할 수 없다는 문제가 제기됨에 따라, 단기간 기상정보 예측에 관심이 집중되었다(El-Shafie et al., 2011; Abhishek et al., 2012; Cramer et al., 2017; Chatterjee et al., 2018; Mishra et al., 2018). 이와 같은 기상정보를 예측한 연구사례에서는 간접적인 효과로써 자연재해를 줄이는데 기여를 했지만, 피해의 개념이 고려되지 않았기 때문에 직접적으로 자연재난으로 인한 피해를 저감하는데 기여하지는 못했다. 자연재해를 사전에 예측하기 위한 연구사례를 살펴보면 강수량, 풍속 등의 기상인자와 피해이력을 연계하여 다양한 통계적 기법을 적용하였다(Dorland et al., 1999; Jang and Kim, 2009; Zhai and Jiang, 2014; Lee et al., 2016; Choi et al., 2017; Choo et al., 2017; Kim et al., 2017). 이러한 연구에서는 대부분 선형적인 모형인 회귀모형 등을 이용하였기 때문에, 비선형적 형태는 고려할 수 없었다. 또한 피해가 발생했던 사례만을 이용하여 분석을 수행하였기 때문에, 실제로 예측에 활용할 때에는 재난의 강도와는 상관없이 항상 예측피해액이 도출된다는 문제가 있다. 즉, 이와 같은 피해 금액을 예측하기 위해서는 선행적으로 피해 발생 여부를 판단할 수 있는 모형을 추가적으로 고려할 필요가 있다.

자연재해 발생 여부를 판단하기 위한 연구를 살펴보면 피해자료와 기상정보를 연계하여 피해가 발생하는 임계기준을 설정하는 연구가 주를 이루었다(Montesarchio et al., 2009; Park and Kang, 2014; Song et al., 2016; Jeong et al., 2017; Cho et al., 2018; Kim et al., 2018; Lee et al., 2018). 그러나 기상정보에 대한 임계 기준을 설정할 경우 기후변화로 인해 변동하는 기상 정보를 반영하기에는 무리가 있다. 또한 임계 기준을 설정할 때 평균적인 개념으로 설정하기 때문에 비선형적으로 발생하는 자연재난은 설명할 수 없다. 따라서 본 연구에서는 국내에서 발생하는 자연재난 중 많은 부분을 차지하는 호우피해에 대해 발생 여부를 파악할 수 있는 모형을 개발하고자 하였다.

Choi et al. (2018)는 본 연구와 가장 유사한 목적으로 선행되었던 호우피해 발생을 예측하기 위한 연구사례이다. 해당 연구에서도 호우피해 자료를 사용하였기 때문에 불균형 자료를 사용한 것으로 판단된다. 불균형 자료의 경우 예측시점에서 올바르게 피해를 분류하기 위한 확률에 대한 경계를 결정해야하며, 상대적으로 개수가 적은 피해가 발생한 날에 면밀하게 평가하기 위해서는 F1-Score를 활용해야 한다. 하지만 해당 연구사례에서는 이러한 부분에 대한 언급이 부족하다. 본 연구에서는 불균형 자료를 사용하기 때문에 각 모형들에 대한 확률 경계를 최적화 하였고, 적합한 평가지표를 사용하였다. 또한 자연재난의 비선형적 특성을 파악할 수 있는 로지스틱 회귀모형, 인공신경망, 배깅, 랜덤포레스트, 부스팅 등의 다양한 머신러닝 기법을 적용하여 호우피해 발생확률 예측 모형을 개발하였다.

2. 호우피해 발생확률 예측모형 방법론

2.1 호우피해 가능성 예측 모형 개발 절차

본 연구에서는 호우피해가 발생하는지를 파악할 수 있는 호우피해 발생확률 예측모형(Prediction Model of Heavy Rain Damage Occurrence Probability, PM-HDOP)을 개발하기 위하여 Fig. 1과 같이 3단계에 걸쳐 분석을 수행하였다. 붉은 테두리는 학습기간, 하늘색 테두리는 평가기간을 의미하며, 이탤릭체로 표기한 부분은 평가 지표를 의미한다.

2.1.1 자료 구축

호우피해 자료를 수집하여 종속변수로 설정하고, 강우자료를 수집하여 선행강우(1~7일), 지속시간별 최대강우(1~24시간)를 1일단위로 계산하여 독립변수로 설정한다. 또한 전체 자료 중 주로 피해가 발생하는 기간을 파악하고, 해당하는 기간의 자료를 추출하여 호우피해 발생확률 예측모형(PM-HDOP)을 개발하는데 활용하였다.

2.1.2 확률경계 및 매개변수 최적화를 통한 모형 학습

전체 자료를 학습기간(2005~2016)과 평가기간(2017)으로

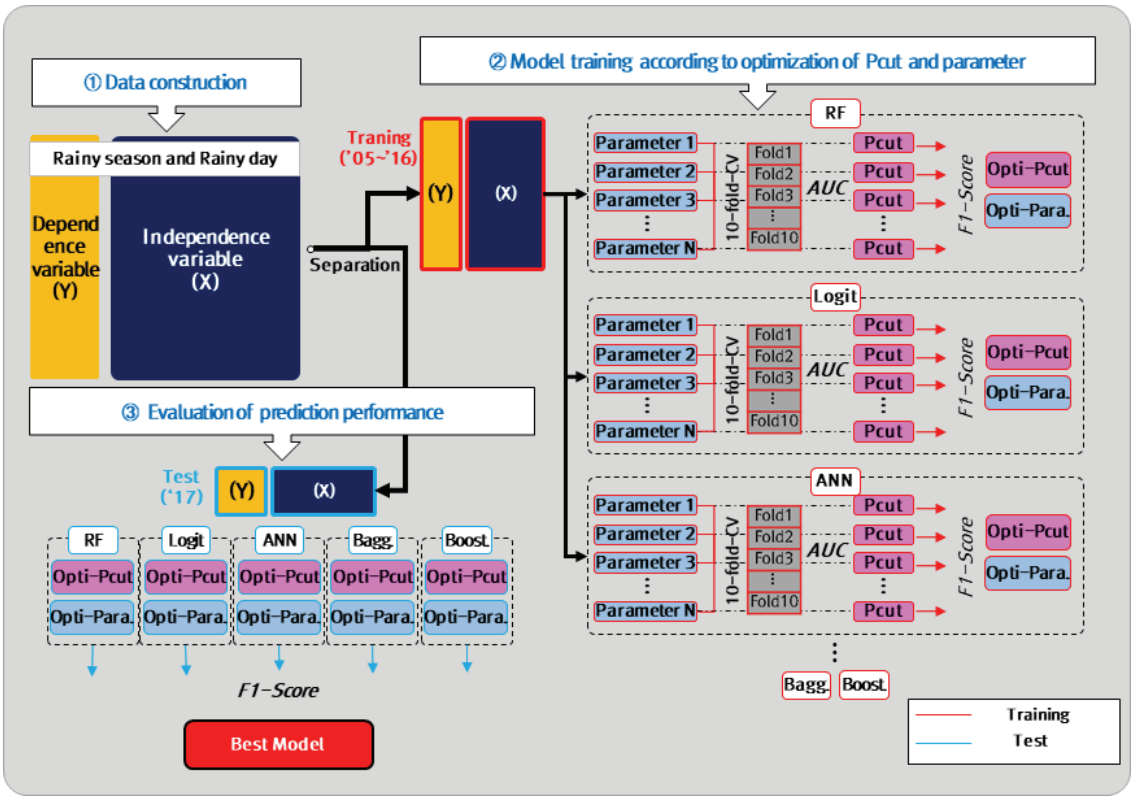


Fig. 1. Procedure to Develop PM-PHD

구분하고, 학습기간 자료를 활용하여 각 모형을 학습하였다. 모형은 랜덤포레스트(Random Forest, RF), 로지스틱 회귀모형(Logistic Regression, Logit), 인공신경망(Artificial Neural Network, ANN), 배깅(Bagging, Bagg.), 부스팅(Boosting, Boost.)을 고려하였다. 모형들의 각 매개변수 마다 10-Fold-Cross Validation을 통해 데이터를 분할하여 모형을 학습하고, 도출되는 확률값 중 Area Under ROC (AUC)를 최대로 하는 확률 경계를 도출하여 각 매개변수의 최적 확률경계로 정의하였다. 해당하는 확률 경계값을 통해 도출되는 호우피해 발생 예측결과를 이용하여 F1-Score를 계산하고, F1-Score의 성능이 가장 우수한 매개변수를 도출하였다.

2.1.3 예측 성능 평가

각 모형들의 최적 매개변수에 대하여 평가기간 자료를 적용하여, 예측 확률 값을 최적 Pcut과 비교하여 피해 발생을 예측하였다. 또한 실제 피해와 비교하여 분류 성능을 평가하고, 분류 성능이 가장 우수한 모형을 최종 모형으로 선정하였다. 여기서 분류 성능 평가지표로써 F1-Score를 활용하였다.

2.2 로지스틱 회귀모형

로지스틱회귀는 확률 모델로서 독립변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계기법이다(Cox, 1958). 로지스틱 회귀의 목적은 일반적인 회귀 분석의 목표와 동일하게 종속변수와 독립변수 간의 관계를 구체적인 함수로 나타내어 향후 예측 모형에 사용하는 것이다. 이는 독립변수의 선형결합으로 종속변수를 설명한다는 관점에서 선형 회귀분석과 유사하지만, 크게 다른 점으로는 종속변수가 범주형 데이터를 하여 입력 데이터가 주어졌을 때 해당 데이터의 결과를 분류하는 분류 기법이다. 로지스틱 회귀 분석은 데이터 마이닝과 같은 다양한 분야에서 분류 및 예측을 위한 모델로서 폭넓게 사용되고 있다. 로지스틱 회귀 기법의 분석 과정은 독립변수를 로짓(logit) 변환 후 최대우도추정법(Maximum likelihood estimation)을 이용한다. 이러한 방식에서 로지스틱 회귀모형은 특정 사건 발생의 확률을 추정한다(Dai and Lee, 2002). 로지스틱 회귀모형은 독립변수에 대한 선형식으로 다음과 같이 표현할 수 있다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

여기서 β_0은 모형의 절편, n은 독립변수의 개수, β_i는 각 x_i의 계수를 의미한다. Eq. (1)을 통해 도출되는 값은 로짓으로서, 이를 다시 역로짓으로 변환하는 작업을 통해 0과 1사이의 확률 값으로 변환된다.

$$P = \frac{1}{1+e^{-y}} \quad (2)$$

2.3 인공신경망

인공신경망은 인간의 뇌가 수많은 신경들로부터 입력과 출력의 신호를 전달하는 과정을 착안하여 모델화한 방법이다(Dreyfus, 1990). Fig. 2에서 인공신경망의 구조를 나타냈다. 먼저 입력층과 은닉층, 출력층으로 구성되며, 각 노드간의 가중치를 통해 결과값을 결정하게 된다.

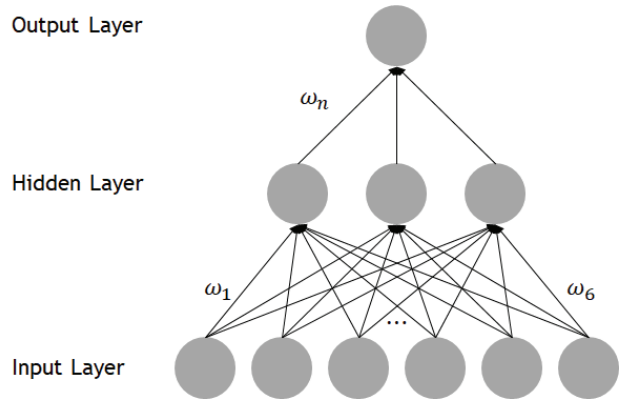


Fig. 2. Concept of Artificial Neural Network

인공신경망이 다른 모형과 가장 큰 차이점은 학습과정에서 있다. 변수들간의 관계를 의미하는 가중치를 학습의 반복을 통한 보정으로 결정되며, 학습방법으로는 역전파(Backpropagation) 알고리즘이 사용된다. 인공신경망의 예측 성능은 은닉층의 노드 수, 학습 횟수 등의 매개변수에 따라 결정된다.

2.4 배깅

배깅(Bagging)은 Boostrap aggregating의 줄임말로써 주어진 데이터에 대한 여러 부스트랩 자료를 생성하고 각 부스트랩 자료를 모델링 한 후 결합하여 최종 예측 모형을 산출하는 방법이다(Aslam et al., 2007). 배깅은 랜덤포레스트, 부스팅과 같이 기계학습(Machine learning) 알고리즘 중 앙상블(Ensemble) 학습기법이다. 앙상블 알고리즘은 더 좋은 예측 성능을 얻기 위하여 다수의 학습 알고리즘을 사용하는 방법이다. 배깅은 전체 자료에서 동일한 크기의 무작위 추출된 하위 표본자료를 생성하고, 각 하위 표본자료마다 학습 모델을 만들어준다. 이때 각 모델은 서로 다른 알고리즘을 적용한다. 각 하위 모형에서 도출된 결과들을 종합하여 최종 결과값을 표현한다. 최종 결과값을 도출하는 방법은 평균값 혹은 투표를 통해 최종 결과를 표출한다. 배깅은 여러 번의 샘플링을 통해 분산을 줄여주기 때문에 모델의 변동성을 감소시킬 수 있다. 일반적으로 데이터가 적을 경우에 우수한 성능을 보인다. 배깅의 조절 매개변수는 샘플링의 개수로써, 많은 샘플링을 수행할수록 더 좋은 결과를 나타낸다. Fig. 3에서는 배깅의 기본적인 개념을 나타냈다.

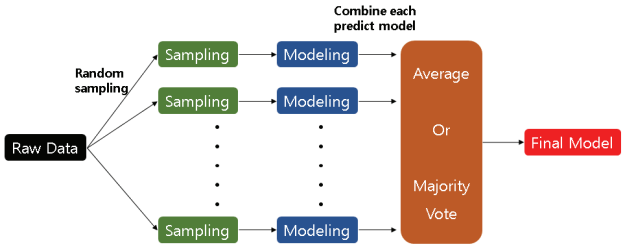


Fig. 3. Concept of Bagging

2.5 랜덤포레스트

랜덤포레스트(Random forest)는 앙상블 기반 모형으로 다수의 의사결정나무로부터 분류 또는 평균 예측치를 출력하는 모형이다. 즉 랜덤포레스트는 여러 의사결정나무를 만들고, 가장 많은 등급으로 분류된 결과를 예측 결과로 결정하는 방법이다. 여러 샘플결과를 통해 모형을 학습하고, 최종 결과를 투표를 통해 표현하는 점에서 배깅과 유사한 개념이다. 랜덤포레스트는 데이터를 부스트랩(Bootstrap)을 통해 숲을 결정하게 되며, 전체 데이터를 전부 이용하는 것이 아니라 샘플의 결과물을 각 트리의 입력 값으로 넣어 학습하는 방식을 의미한다. 랜덤포레스트는 간편하고 빠른 학습과정에도 높은 정확도를 가지며, 일반화 할 수 있는 성능을 보유한 기법이다. 랜덤포레스트는 일반적으로 변수의 개수가 m개이면 각 분할에서 랜덤으로 m/3개의 변수를 선택하여 트리가 만들어진다. 랜덤포레스트는 데이터와 변수를 샘플링하여 서로 조금씩 다른 나무들로 구성되었기 때문에 각 나무들의 예측값은 과적합 문제가 발생하지 않는다. 일반적으로 랜덤포레스트는 나무의 개수와 샘플링의 개수를 매개변수로서 고려하며, 본 연구에서는 나무의 개수는 독립변수에 따라 고정하여 분석을 수행하였다. Fig. 4에서는 랜덤포레스트의 개념을 나타냈다.

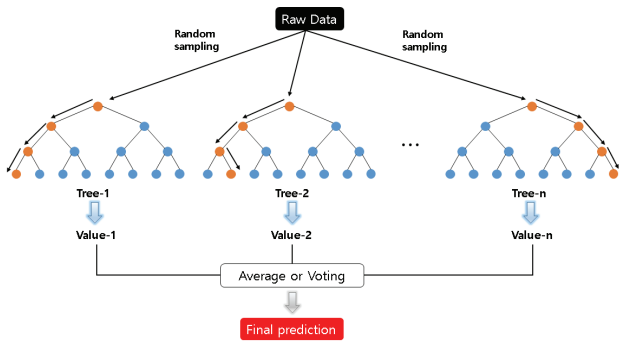


Fig. 4. Concept of Random Forest

2.6 부스팅

부스팅(Boosting) 알고리즘도 앙상블 기반의 모형으로써 순차적으로 약한 분류기를 생성하여 강한 분류기를 만들어내는 학습기법이다. 부스팅 알고리즘의 분류 시스템을 구성하는 분류기의 수는 실험자에 의해 결정되며, 이를 매개변수

로써 조절할 수 있다(Wang and Japkowicz, 2010). 부스팅 알고리즘은 매 반복학습 단계마다 이전 단계에서 만들어진 수행결과를 기반으로 관측치들의 가중치를 재조정하게 되는데, 자동 가지치기로 인해 과적합(Overfitting) 문제가 적게 발생하는 장점이 있다. 이전 단계에서 만들어진 분류기에 대하여 잘 맞춘 분류기의 가중치는 높이고, 틀린 분류기는 가중치를 낮춰서 전체적으로 성능을 향상시키는 방향으로 업데이트를 수행한다. 이러한 일련의 과정을 특정 조건이 될 때까지 반복하고, 반복이 끝나면 누적된 각 분류기의 가중치를 이용하여 최종 학습 모델을 만들게 된다(Fig. 5).

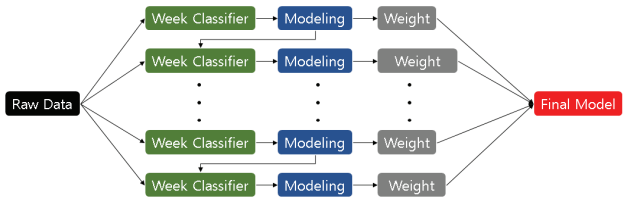


Fig. 5. Concept of Boosting

앞서 설명한 배깅과 랜덤포레스트는 무작위 샘플링을 기반으로 하나의 분류기가 하나의 결과를 도출하며, 해당 결과를 투표 혹은 평균 개념으로 최종 결과를 표현한다. 부스팅은 첫 번째 모형이 두 번째 모형에 영향을 미치는 방식으로 점차 성능을 높여가는 과정을 거치고, 최종결과를 표현할 때에도 모든 모형에 대한 가중치를 적용하여 결과를 도출한다. 부스팅은 모델의 정확도를 향상시키는 방향으로 반복하기 때문에, 데이터가 다수 존재할 경우 우수한 성능을 보인다.

2.7 분류 성능 평가

본 연구에서는 호우피해가 발생할 경우 1, 발생하지 않을 경우를 0이라고 정의하였다. 이와 같은 분류에 대한 검증은 혼동행렬(Confusion Matrix)이라고 불리는 이진 분류 결과표를 작성하여 다양한 성능지표를 통해 수행할 수 있다(Fawcett, 2006). 혼동행렬은 Table 1과 같이 표현할 수 있다.

Table 1. Confusion Matrix

Total Damage		Predicted(pred.)	
		0	1
Observed(obs.)	0	TN	FP
	1	FN	TP

- TP : pred. is 1 and obs. is 1
- TN : pred. is 0 and obs. is 0
- FP : pred. is 1 and obs. is 0
- FN : pred. is 0 and obs. is 1

단순히 Confusion Matrix만을 이용하여 분류 성능을 나타내기에는 어려움이 있다. Table 1을 활용하여 다양한 성능 평가 지표를 산정할 수 있고, 활용 여부는 목적에 따라 결정할 수 있다(Table 2). 성능 평가 지표의 종류로는 Accuracy(정확도), Error Rate(오류율), Sensitivity(민감도), Precision(정밀성), Specificity(특이도), AUC, F1-Score 등이 있다(Powers, 2011).

Table 2. Equation of Each Evaluation Index

Evaluation index	Equation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Error Rate	$\frac{FN+FP}{TP+TN+FP+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
Specificity	$\frac{TN}{TN+FP}$

먼저 정확도는 전체 데이터 중에 실제 1을 1로 예측한 비율을 말하고, 오류율은 전체 데이터 중에 실제 0을 0이라고 예측한 비율을 말한다. 민감도는 실제 값이 1인 것 중 1이라고 예측한 비율을 의미하며, 정밀성은 1이라고 예측한 것 중 실제 값이 1인 비율을 의미한다. 또한 특이도는 실제 값이 0인 것 중 0이라고 예측한 비율을 말한다. 민감도와 1-특이도의 그래프를 통해 Receiver Operating Characteristics (ROC) 곡선을 그릴 수 있고, 그 곡선의 아래 면적을 Area Under Curve (AUC)라고 한다. 정밀성과 민감도를 활용하여 F-Score를 산정할 수 있으며, F-Score는 불균형적인 자료에 적합하다. 본 연구에서는 확률 경계를 도출할 때에는 AUC를 활용하였으며, 분류의 성능을 평가할 때 F1-Score 평가지표를 활용하였다.

$$F-score = \frac{(1+\beta^2)(\mathrm{Prec.}\times Sens.)}{(\beta^2\mathrm{Prec.}+Sens.)} \quad (3)$$

여기서 F-Score는 일반적으로 β는 1로 표기하고, F1-Score라고 말한다.

3. 호우피해 발생확률 예측 모형 개발

3.1 대상지역 선정

본 연구에서는 수도권 지역에 대하여 시범적으로 호우피해 발생확률 예측 모형(PM-HDOP)을 개발하고자 하였다. 수도권 지역에는 경기도, 서울특별시, 인천광역시가 포함되며, 각각 31개, 25개, 10개의 행정구역이 포함된다. 또한 수도권 지역에 영향을 미치는 자동기상관측장비(Automatic

Weather System, Automatic Weather Station, AWS)는 총 102개소로 파악되며, Fig. 6과 같이 표현하였다.

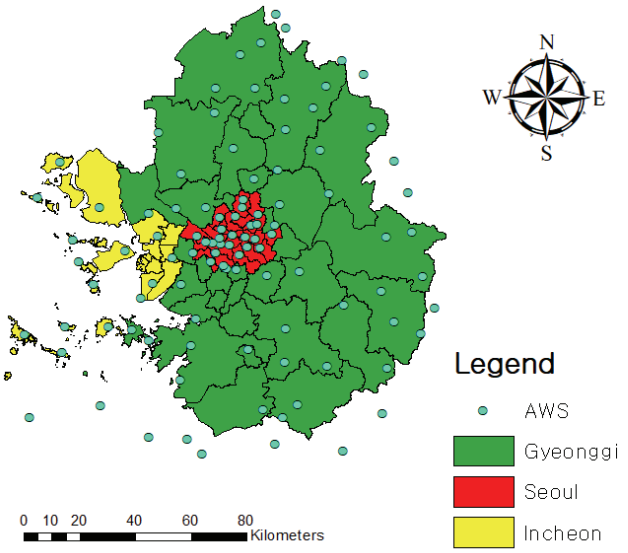


Fig. 6. Study Area

3.2 자료 구축

3.2.1 종속 변수 구축

본 연구에서는 PM-HDOP을 개발하기 위한 종속변수로써, 호우피해 자료를 고려하였다. 호우피해 자료는 행정안전부(Ministry of the Interior and Safety, MOIS)에서 지난해의 피해 현황을 정리한 보고서인 재해연보(MOIS, 2018)를 통해 수집하였다. 재해연보는 1985년부터 기상 원인별, 시군구별로 구분하여 피해 현황을 정리하고 있으며, 각각의 재해는 불규칙하게 발생하기 때문에 재해 시작일시와 종료일시를 함께 기록하고 있다. 종속변수와 독립변수의 전체 자료 기간을 동일하게 구축해야 하기 때문에, 본 연구에서는 2005년부터 2017년까지 자료를 수집하였다. 또한 호우피해가 발생한 기간이 중요한 요소이기 때문에, 시군구 단위로 호우피해가 발생한 기간을 수집하였다. 이를 통해 1일단위로 호우피해가 발생한 경우 "1"로 표기하고, 호우피해가 발생하지 않은 경우 "0"으로 표기하였다. 최종적으로 시군구별 1일 단위 피해발생 여부를 나타낸 자료를 종속 변수로 활용하였다.

3.2.2 독립 변수 구축

기상자료는 기상청(KMA), 국토교통부(Ministry of Land, Infrastructure and Transport, MOLIT), 한국수자원공사(K-Water) 등에서 수집 및 관리하고 있다. 그 중에서 가장 신뢰도가 높은 기관은 기상청으로써, 오랜 기간 동안 관측 자료에 대한 관리가 꾸준하게 이루어지고 있다. 기상청에서 관리하는 관측 장비는 종관기상관측장비(Automated Surface Observing System, ASOS)와 무인으로 운영되는 자동기상관측장비(Automatic Weather System, Automatic Weather Station, AWS)로 두 가지 종류가 있다. ASOS의 경우 1970년대부터 관측을 시작하여 자료의 보유기간이 길다는 장점이 있지만, 총 101개소가 설치되어 조밀하지 않다는 단점이 존재한다. AWS의 경우 대부분 2000년대 초반부터 관측을 시작하여 자료의 보유기간이 짧다는 단점이 있지만 총 501개소가 설치되어 비교적 조밀하다는 장점이 있다. 본 연구에서는 자료의 보유기간이 짧더라도 조밀하게 분포하여, 각 지역마다 발생하는 강우특성을 상세히 파악할 수 있는 AWS를 사용하는 것이 적합하다. 또한 본 연구에서는 지점 강우자료를 그대로 사용하는 것이 아니라, Thiessen 다각형법을 활용하여 시군구 단위 면적 강우로 환산하였다. 시군구 면적 강우자료를 2005년부터 2017년까지 1시간 단위로 구축하였고, 이를 1일 단위로 선행강우량(7일~1일)과 1일 총강우량, 지속시간별 최대강우량(1~24시간)을 계산하였다. 최종적으로 앞서 설명한 32개의 강우변수를 독립변수로써 고려하였다.

3.2.3 자료 전처리

호우피해 발생확률 예측 모형 개발을 위해 구축된 자료를 Table 3에서 살펴보면 피해가 발생한 날 대비 피해가 발생하지 않는 날이 현저하게 많다.

Table 3. Number of Damage Data Each Region

Total	Number of data	
	1	0
Gyeonggi-do	2,658	144,530
Seoul-city	737	117,963
Incheon-city	573	46,907

이 경우 자료의 형태가 매우 불균형하다고 볼 수 있으며, 반응변수의 불균형한 비율이 분류 모형의 성능에 문제를 준다(Longadge et al., 2013). 예를 들어 0의 비율이 90%일 때 모든 예측을 0으로 하는 모형의 정확도는 90%이기 때문에 마치 좋은 예측 모형인 것처럼 보일 수 있지만, 이는 실제로는 1에 대한 예측력이 전혀 없는 모형이라고 할 수 있다. 1(호우피해 발생)과 0(호우피해 없음)이 균등하게 분포하는 자료를 이용하여, 모형을 개발할 경우 가장 효과적일 것이다. 이러한 불균형적인 데이터에서의 1과 0의 분류모형의 성능을 개선하기 위한 방법으로 다양한 샘플링 기법들이 제안되었다. 본 연구에서는 호우피해 발생을 나타내는 1에 대하여 적절하게 예측할 수 있는 모형을 개발하기 위해 호우피해가 주로 어느 기간에 발생하는지를 파악하고, 해당하는 기간의 자료를 추출하여 활용하였다. Fig. 7과 같이 호우피해는 주로 우기(6월~9월)에서 발생하였으며, 강우가 발생한 날 피해가 발생한 것을 확인할 수 있었다.

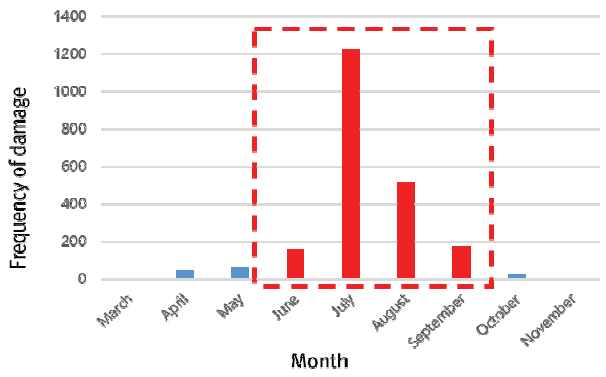


Fig. 7. Frequency of Damage in Each Month

따라서 본 연구에서는 우기시즌에 해당하며 강우가 발생한 날을 추출하여 데이터를 구축하였고, 자료의 개수를 Table 4와 같이 정리하였다.

Table 4. Number of Damage Data After Sampling

Rainy season & Rainy days	Number of data	
	1	0
Gyeonggi-do	2,597	19,988
Seoul-city	737	15,900
Incheon-city	570	5,706

본 연구에서는 호우피해 발생확률 예측 모형을 개발하기 위하여, 앞서 구축한 자료를 총 2가지(Training section, Test section)로 구분하였다. Training section (2005년~2016년)에서는 각 모형에 대하여 매개변수 및 확률경계를 최적화하였고, Test section (2017년)에는 성능평가를 통해 최종모형을 선정하였다.

3.3 모형 학습을 통한 매개변수 및 확률 경계 최적화

Training section (2005년~2016년)의 자료를 이용하여 각각의 모형을 학습하였고, 학습과정에서 교차검증을 통해 매개변수와 확률 경계를 최적화하였다. 일반적으로 학습에 사용하지 않은 표본에 대해 종속 변수를 알아내고자 하는 것을 예측이라고 하며, 예측 성능이 표본 내에서는 좋지만 표본 외에서 좋지 않은 경우를 과적합화(Overfitting)이라고 한다. 과적합문제가 발생할 경우 새로운 독립변수 데이터에 대해 전혀 예측하지 못하기 때문에 목적에 맞지 부합하지 못하게 된다. 이를 해결하기 위하여 교차 검증을 고려하게 되며, 교차 검증은 수차례 모형 학습과정 및 검증과정을 통해 예측 모형의 통계적 신뢰도를 높일 수 있는 방법이다. 또한 데이터의 수가 적은 경우에도 검증 데이터를 적게 성능을 평가하면 검증 성능의 신뢰도가 떨어질 수 있고, 검증 데이터를 늘리면 학습용 데이터가 적어지기 때문에 정상적인 학습이 이루어지지 않는다. 이러한 문제를 해결하기 위한 방법이 K-fold-Cross Validation이다. 10-Fold-Cross Validation은 학습용 데이터를 10%씩 겹치지 않는 10개의 조각으로 분할하여, 90%의 데이터를 이용하여 10%에 해당하는 데이터를 예측하여 성능을 검증하는 방식이다(Fig. 8).

3.3.1 매개변수별 확률경계 최적화

본 연구에서는 PM-HDOP을 개발하기 위해 총 5개의 모형(랜덤포레스트(RF), 로지스틱 회귀모형(Logistic), 인공신경망(ANN), 배깅(Bagging), 그레디언트 부스팅(Boosting))을 적용하였다.

모든 모형에서 1로 분류할 확률을 예측할 수 있도록 설정하여 모형을 학습하였다. 각 모형들의 매개변수에 대하여 매번 10-fold-CV를 수행하여 최적의 확률경계(pcut)을 추출하였고, 최적의 확률 경계를 도출하는 평가지표로써 Area Under the Curve (AUC)를 고려하였다. 적합된 모형을 10%의 검증용 데이터 조각에 적용하여 1(호우피해 발생)에 대한 확률 예측값을 계산하고, 특정 확률 경계값과 비교하여 호우피해 발생여부를 판단하여 민감도와 1-특이도로 만들어지는 2차원 곡선을 작성하며 이를 Reciever Operating Characteristic (ROC)곡선이라고 한다. 이 곡선의 아래 면적을 AUC라고 한다(Fig. 9).

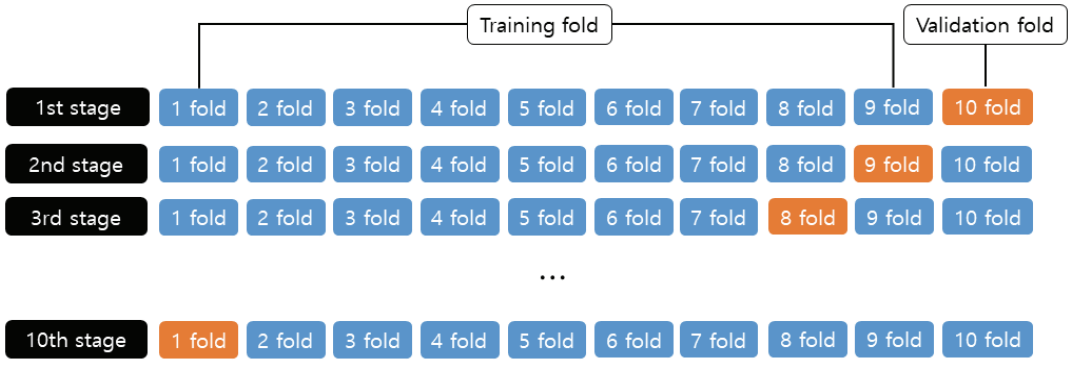


Fig. 8. Concept of 10-Cross-Validation

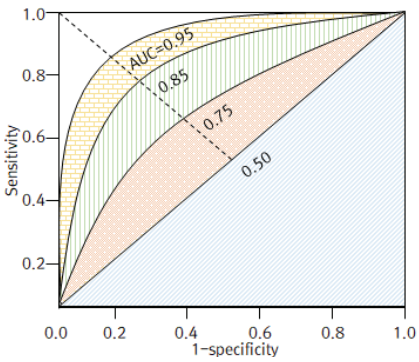


Fig. 9. Concept of ROC and AUC

AUC의 값은 0.5~1의 값을 가지며, 1에 가까울수록 예측력이 우수한 모형이라고 판단할 수 있다. AUC는 확률경계에 관계없이 이항 분류 모형 간의 상대적인 예측력을 비교할 수 있다는 장점 때문에 대표적인 성능 비교 척도로써 널리 사용되고 있다. 균형 데이터일 경우 확률 경계를 따로 설정하지 않아도 무방하지만, 본 연구에서 활용한 자료와 같이 불균형한 자료의 경우 확률 경계를 추가적으로 설정해야 한다. Fig. 10에서 예시로써 경기도 지역에 대한 지역별 각 모형의 매개변수별 확률 경계(Pcut)를 추정한 결과를 나타냈다. ROC 곡선은 각 매개변수 마다 10-fold-CV를 통해 예측된 10개의 조각들을 병합함으로써 그릴 수 있다. 로지스틱 회귀모형의 경우에는 매개변수를 고려할 필요가 없기 때문에 1번의 ROC 곡선을 나타냈으며, 나머지 모형들은 모두 총 8번씩의 모형을 표현하였다. 본 연구에서 추정한 각 모형들의 매개변수를 정리하여 Table 5에 제시하였다.

Table 5. Parameter Combination of Each Model

Parameter	ANN (size)	Bagging (nbagg)	RF (ntree)	Boosting (mfinal)
1	3	5	150	5
2	6	10	200	10
3	9	15	250	15
4	12	20	300	20
5	15	25	350	25
6	18	30	400	30
7	21	35	450	35
8	24	40	500	40

Fig. 10에서 각 모형들의 매개변수마다 최적의 확률경계(Pcut)를 도출하는 과정에 대하여 예시로 나타냈으며, 대부분 AUC가 80% 이상의 분류 성능을 나타내고 있었다. 여러 모형들에 대한 ROC 곡선은 경기도 지역에 대한 결과이다.

3.3.2 모형별 매개변수 최적화

모형들의 매개변수를 결정하기 위한 평가지표로는 1(피해발생)에 대하여 면밀하게 검토할 수 있는 방법인 F1-Score를 고려하였다. 모형별 매개변수마다 최적의 확률 경계를 이용하여 F1-Score를 산정하고, 최종적으로 선정된 모형별 매개변수 및 확률경계를 Tables 6, 7, 8에서 각각 경기도, 서울, 인천에 대하여 정리하였다. 경기도 지역의 경우 랜덤포레스트의 최적 매개변수가 450일 때 가장 높은 성능을 보였으며, Pcut은 0.1733일 때 가장 높은 성능을 보였다. 로지스틱 회귀모형은 Pcut이 0.1037일 때 가장 높은 성능을 보였으며, ANN은 노드 수가 12개와 Pcut이 0.1464일 때 가장 높은 성능을 보였다. 배깅은 nbagg가 5이며, Pcut이 0.01일 때 가장 높은 성능을 보였다. 부스팅의 경우 mfinal이 25이며, Pcut이 0.2956일 때 가장 높은 성능을 보였다.

Table 6. Optimum Pcut and Parameter of Each Model in Gyeonggi-do Province

Model	AUC	Optimum Pcut	F1-score	Optimum Parameter
Logistic	84.59%	0.1037	47.35%	0
ANN	84.64%	0.1464	50.28%	12
Bagging	74.70%	0.0100	52.37%	5
RF	88.70%	0.1733	55.80%	450
Boosting	87.62%	0.2956	48.73%	25

Table 7. Optimum Pcut and Parameter of Each Model in Seoul City

Model	AUC	Optimum Pcut	F1-score	Optimum Parameter
Logistic	87.92%	0.0332	27.40%	0
ANN	87.68%	0.0669	35.74%	21
Bagging	77.58%	0.0100	49.26%	5
RF	93.34%	0.0750	43.76%	200
Boosting	93.80%	0.2899	36.75%	30

Table 8. Optimum Pcut and Parameter of Each Model in Incheon City

Model	AUC	Optimum Pcut	F1-score	Optimum Parameter
Logistic	82.49%	0.0840	39.15%	0
ANN	84.29%	0.0675	35.57%	18
Bagging	79.93%	0.0200	45.09%	5
RF	87.51%	0.1540	47.12%	500
Boosting	88.13%	0.3344	45.50%	30

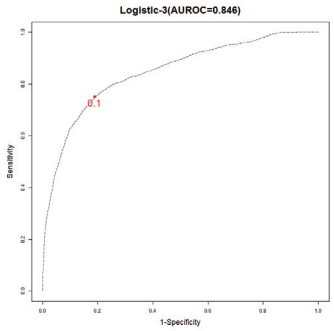
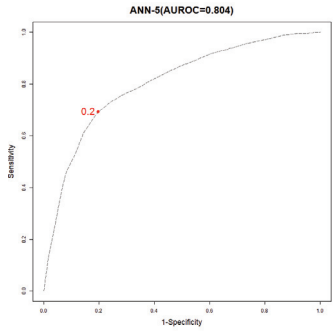
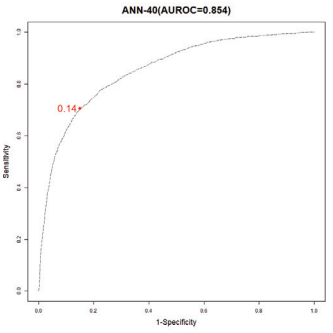
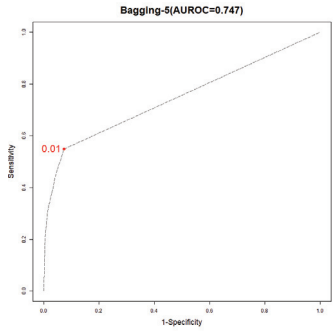
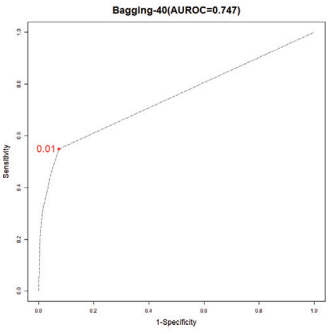
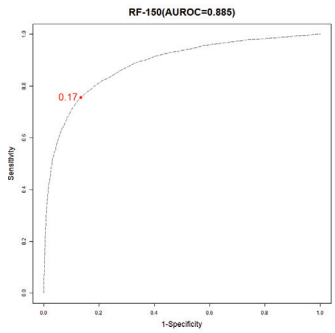
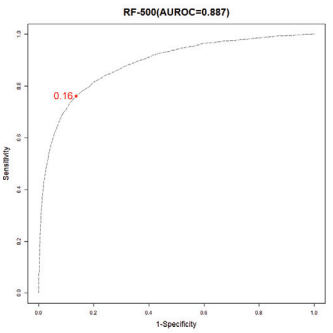
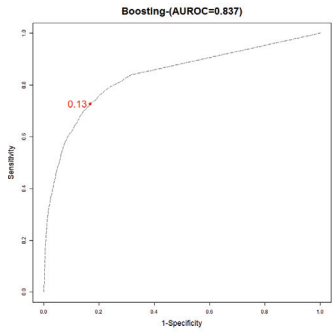
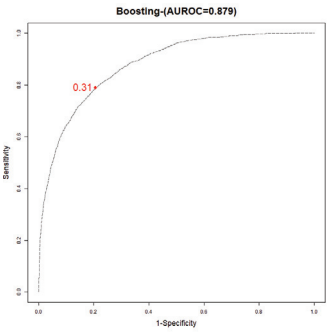
Model	Parameter 1	···	Parameter 8
Logistic Regression			
ANN		···	
Bagging		···	
Random Forest		···	
Boosting		···	

Fig. 10. ROC Plot to Optimize Pcut for Each Parameter of Models in Gyeonggi-do Province

모든 지역에서 배깅의 확률 경계는 다른 모형들에 비해 설정되었는데, 이는 호우피해가 발생할 확률이 낮게 설정되어 약한 강도에도 피해가 발생한다고 분류할 수 있음을 의미한다.

부스팅의 확률 경계는 다른 모형들에 비해 항상 높게 설정된 것을 확인할 수 있는데, 이는 모든 모형들 중에서 가장 강한 강도의 재난일 때를 경계로 설정한 것을 의미한다. Tables 6~8까지의 결과를 이용하여 최종 모형을 선정하기에는 무리가 있다. 이는 각 모형마다 매개변수의 성능을 평가한 결과이며, 모형의 학습 과정에서 교차 검증을 수행하였기 때문에 자료의 구분 없이 재학습하여 모형들의 계수 및 가중치 등을 재산정할 필요가 있다. 모형들의 직접적인 성능을 비교하기 위해 지역마다 앞서 결정된 최적의 확률 경계와 매개변수를 이용하여 2005년부터 2016년까지 총 5개의 모형(RF, Logistic, ANN, Bagging, Boosting)을 재학습하였다.

3.4 예측력 평가 및 최종 모형 선정

각 지역마다 최종 모형을 선정하기 위해 재학습 과정을 거친 5개의 모형들을 Test section 자료를 적용하여, 성능을 비교하였다. 단, Test section에서는 실제로 예측환경과 동일하게 유지하기 위하여, 우기시즌과 강우가 발생한 날을 추출하지 않았다. 2017년 1월 1일부터 12월 31일까지 1일 단위로 지역마다 5개 모형에 자료를 적용하여 호우피해 발생확률을 예측하였고 확률경계와 비교를 통해 호우피해가 발생여부를 분류하였다. 그 결과를 Tables 9 ~ 11과 같이 Confusion Matrix를 통해 정리하였다.

Table 9. Confusion Matrix Each Model in Gyeonggi-do Province

Logistic		Predict	
		0	1
Obs.	0	10353	627
	1	127	208

ANN		Predict	
		0	1
Obs.	0	10647	333
	1	149	186

Bagging		Predict	
		0	1
Obs.	0	10869	111
	1	235	100

RF		Predict	
		0	1
Obs.	0	10666	314
	1	143	192

Boosting		Predict	
		0	1
Obs.	0	10621	359
	1	118	217

Table 10. Confusion Matrix Each Model in Seoul City

Logistic		Predict	
		0	1
Obs.	0	5188	3893
	1	16	28

ANN		Predict	
		0	1
Obs.	0	7532	1549
	1	27	17

Bagging		Predict	
		0	1
Obs.	0	9032	49
	1	39	5

RF		Predict	
		0	1
Obs.	0	8822	259
	1	22	22

Boosting		Predict	
		0	1
Obs.	0	8824	257
	1	20	24

Table 11. Confusion Matrix Each Model in Incheon City

Logistic		Predict	
		0	1
Obs.	0	2234	1343
	1	40	33

ANN		Predict	
		0	1
Obs.	0	3287	290
	1	46	27

Bagging		Predict	
		0	1
Obs.	0	3531	46
	1	46	27

RF		Predict	
		0	1
Obs.	0	3452	125
	1	46	27

Boosting		Predict	
		0	1
Obs.	0	3492	85
	1	52	21

실제로 피해가 발생한 날 중 예측을 올바르게 한 결과(TP)를 붉게 표기하였다. 경기도, 서울, 인천에서의 결과를 살펴보면 로지스틱 회귀모형은 모든 지역에서 실제로 피해가 발생하지 않은 날임에도 피해가 발생한다고 예측한 결과(FP)가 다른 모형들에 비해 현저하게 많은 것을 확인할 수 있다. 이는 로지스틱회귀모형에서의 Pcut이 적절하게 선정되지 않았을 가능성이 있으며, 다른 모형에 비해 단순하기 때문에 성능이 떨어지는 것으로 판단된다. 또한 경기도의 경우에는 다른 지역에 비해 행정구역이 많이 분포하고 실제로 피해가 발생한 날도 많기 때문에, 대부분의 모형의 성능이 유사하며 비교적 좋은 성능을 나타나는 것을 확인할 수 있다. 하지만 서울의 경우 다른 지역에 비해 피해가 발생하지 않은 날 대비 피해가 발생한 날이 더 적게 분포하기 때문에, 예측 성능이 비교적 떨어지는 것을 확인할 수 있다.

위의 결과만을 이용하여 지역별 최종 모형을 선정하기에는 어려움이 있기 때문에, 정량적인 평가 지표로써 F1-Score를 산정하여 Table 12에서 각 모형에 대한 결과를 비교하였다.

Table 12. Comparison of Quantitative Evaluation Indices

Region	F1-Score				
	Logit.	ANN	Bagg.	RF	Boost.
Gyeonggi	35.6%	43.6%	36.6%	45.6%	47.6%
Seoul	1.4%	2.1%	10.2%	13.5%	14.7%
Incheon	4.6%	13.8%	36.9%	24.0%	23.4%

경기도와 서울 지역에서는 부스팅의 예측 성능이 가장 높게 나타났으며, 인천 지역에서는 배깅의 예측 성능이 가장 높게 나타났다. 인공신경망의 경우 모든 지역에서 중간 정도의 예측 성능을 보이는 것을 확인할 수 있었다. 배깅의 경우 인천의 경우 특별하게 예측 성능이 좋게 나타났으며, 나머지 지역에서는 중간 정도의 성능을 보였다. 랜덤포레스트의 경우 비교적 간단히 수행할 수 있는 머신러닝 계열임에도 불구하고, 예측 성능이 모든 지역에서 두 번째로 좋은 결과를 나타냈다. 본 연구에서 고려했던 머신러닝 계열 중 가장 발전된 버전인 부스팅은 인천을 제외한 지역에서 모두 가장 좋은 성능을 나타냈다. 따라서 경기도와 서울 지역에서는 부스팅 모형을 최종 모형을 선정하였으며, 인천 지역에서는 배깅을 최종 모형으로 선정하였다.

4. 요약 및 결론

본 연구에서는 사전에 피해가 발생하는지를 파악하기 위하여 다양한 머신러닝 기법을 활용한 호우피해 발생확률 예측 모형(PM-HDOP)을 개발하였다. 먼저 호우피해 자료를 이용하여 종속변수를 구축하고, 강우자료를 이용하여 다양한 독립변수들을 구축하였다. 또한 자료의 특성을 고려하여 샘플링을 수행하였고, 추출된 자료를 이용하여 모형의 학습과정을 수행하였다. Training section에서는 10-fold-Cross-Validation을 통해 각 모형별 매개변수와 Pcut을 최적화 하였고, Test section에서는 각 지역마다 최종 모형을 선정하기 위하여 본 연구에서 고려한 5개의 머신러닝에 대한 예측 성능을 비교하였다. 본 연구의 주요 결과를 요약하여 아래와 같이 정리하였다.

(1) 경기도와 서울에서는 F1-score가 47.67%, 14.77%로 부스팅이 최종 모형으로 선정되었으며, 인천에서는 36.99%로 배깅이 최종 모형으로 선정되었다.
(2) 로지스틱 회귀모형의 경우 가장 예측성능이 떨어지는 것으로 확인되었으며, 이는 다른 머신러닝 계열의 모형에 비해 실제는 0인데 1로 예측한 사상(FP)이 현저하게 많기 때문이라고 판단된다.
(3) 배깅의 경우 부스팅과 더불어 앙상블 계열의 머신러닝 모형임에도 불구하고, 인천을 제외한 다른 지역에서는 신경망과 비슷한 성능을 나타냈다.
(4) 신경망의 경우 Hidden layer를 1개로 설정하여 노드 수만을 조절할 경우 예측 성능이 그다지 높지 않은 것을 확인할 수 있었다.
(5) 랜덤포레스트의 경우 앙상블 계열의 머신러닝으로써 최종모형으로 채택되지는 않았지만, 모든 지역에서 우수한 성능을 보이는 것을 확인할 수 있었다.

현재 국내에서는 호우피해가 발생하는지에 대한 주의 및 경보는 이루어지고 있지 않다. 이러한 부분에서 본 연구는 강점으로써 두드러질 수 있다고 판단되지만, 아직까지 예측에 대한 성능이 월등히 높다고 판단하기에는 무리가 있다. 이는 대부분의 모형에서 TP가 FP나 FN 보다 적기 때문이라고 할 수 있으며, 향후 연구에서는 TP를 높이기 위해 추가로 다양한 요소들을 고려하여 예측 모형을 개발할 필요가 있다. 마지막으로 신경망 모형에서 Hidden layer를 1개가 아니라 증폭시켜 비선형적 특성을 반영한 Deep learning 모형을 개발한다면 예측 성능이 큰 효과를 불러올 수 있을 것으로 판단된다. 본 연구 결과를 활용하여 현재 국내에서 서비스 되고 있지 않는 호우피해 발생 예측에 대한 서비스가 이루어진다면, 피해예상지역 긴급 대피 및 위험지역 주민 사전대비 등의 효율적인 재난관리에 기여할 수 있을 것으로 판단된다.

감사의 글

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

References

Abbot, J., and Marohasy, J. (2012). Application of artificial neural networks to rainfall forecasting in Queensland, Australia. *Advances in Atmospheric Sciences*, Vol. 29, pp. 717-730.

Abhishek, K., Kumar, A., Ranjan, R., and Kumar, S. (2012). A rainfall prediction model using artificial neural network. *Proceedings of 2012 IEEE Control and System Graduate Research Colloquium*, pp. 82-87.

Aslam, J.A., Popa, R.A., and Rivest, R.L. (2007). On estimating the size and confidence of a statistical audit. *EVT '07 Proceedings of the USENIX Workwhop on Accurate Electronic Voting Technology*, p. 8.

Chatterjee, S., Datta, B., Sen, S., Dey, N., and Debnath, N.C. (2018). Rainfall prediction using hybrid neural network approach. *Proceedings of 2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)*, pp. 67-72.

Cho, J., Bae, C., and Kang, H. (2018). Development and application of urban flood alert criteria considering damage records and runoff characteristics. *Journal of Korea Water Resources Association*, Vol. 51, No. 1, pp. 1-10.

Choi, C., Kim, J., Kim, J., Kim, D., Bae, Y., and Kim, H.S. (2018). Development of heavy rain damage prediction model using machine learning based on big data. *Advances in Meteorology*. Vol. 2018, Article ID 5024930. doi:10.1155/2018/5024930

Choi, C.H., Kim, J.S., Kim, J.H., Kim, H.Y., Lee, W.J., and Kim, H.S. (2017). Development of heavy rain damage prediction function using statistical methodology. *J. Korean Soc. Hazard Mitig.*, Vol. 17, No. 3, pp. 331-338.

Choo, T.H., Kwak, K.S., Ahn, S.H., Yang, D.U., and Son, J.K. (2017). Development for the function of wind wave damage estimation at the Western coastal zone based on disaster statistics. *Journal of the Korea Academia-Industrial cooperation Society*, Vol. 18, No. 2, pp. 14-22.

Cox, D.R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 20, No. 2, pp. 215-232.

Cramer, S., Kampouridis, M., Freitas, A.A., and Alexandridis, A.K. (2017). An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Systems with Applications*, Vol. 85, pp. 169-181.

Dai, F.C., and Lee, C.F. (2002). Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. *Geomorphology*, Vol. 42, No. 3-4, pp. 213-228.

Dorland, C., Tol, R.S., and Palutikof, J.P. (1999). Vulnerability of the Netherlands and Northwest Europe to storm damage under climate change. *Climatic Change*, Vol. 43, No. 3, pp. 513-535.

Dreyfus, S.E. (1990). Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. *Journal of Guidance, Control, and Dynamics*, Vol. 13, No. 5, pp. 926-928.

El-Shafie, A., Jaafer, O., and Seyed, A. (2011). Adaptive neuro-fuzzy inference system based model for rainfall forecasting in Klang River, Malaysia. *International Journal of Physical Sciences*, Vol. 6, No. 12, pp. 2875-2888.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, Vol. 27, No. 8, pp. 861-874.

Jang, O.J., and Kim, Y.O. (2009). Flood risk estimation using regional regression analysis. *J. Korean Soc. Hazard Mitig.*, Vol. 9, No. 4, pp. 71-80.

Jeong, M.S., Oak, Y.S., Lee, Y.K., Lee, Y.S., Park, M.R., and Lee, C.H. (2017). Estimation of disaster prevention target rainfall according to urban disaster prevention performance. *Journal of the Korea Academia-Industrial cooperation Society*, Vol. 18, No. 4, pp. 101-110.

Kannan, M., Prabhakaran, S., and Ramachandran, P. (2010). Rainfall forecasting using data mining technique. *International Journal of Engineering and Technology*, Vol. 2, No. 6, pp. 397-401.

Kim, J.S., Choi, C.H., Kim, D.H., Joo, H.J., Kim, J.W., and Kim, H.S. (2018). Establishment of hazard-triggering rainfall according to heavy rain damage scale: Focused on Gyeonggi-do. *Journal of Climate Research*, Vol. 13, No. 4, pp. 297-311.

Kim, J.S., Choi, C.H., Lee, J.S., and Kim, H.S. (2017). Damage prediction using heavy rain risk assessment: (2) Development of heavy rain damage prediction function. *J. Korean Soc. Hazard Mitig.*, Vol. 17, No. 2, pp. 371-379.

Kim, Y., Kang, N., Kim, S., and Kim, H. (2013). Evaluation for snowfall depth forecasting using neural network and multiple regression models. *J. Korean Soc.*

Hazard Mitig., Vol. 13, No. 2, pp. 269-280.

Lee, J.S., Eo, G., Choi, C.H., Jung, J.W., and Kim, H.S. (2016). Development of rainfall-flood damage estimation function using nonlinear regression equation. *Journal of the Korean Society of Disaster Information*, Vol. 12, No. 1, pp. 74-88.

Lee, S.H., Kang, D.H., and Kim, B.S. (2018). A study on the method of calculating the threshold rainfall for rainfall impact forecasting. *J. Korean Soc. Hazard Mitig.*, Vol. 18, No. 7, pp. 93-102.

Longadge, R., Dongre, S., and Malik, L. (2013). Class imbalance problem in data mining: Review. *International Journal of Computer Science and Network*, Vol. 2, No. 1, Retrieved from https://arxiv.org/abs/1305.1707

Mekanik, F., Imteaz, M.A., Gato-Trinidad, S., and Elmahdi, A. (2013). Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. *Journal of Hydrology*, Vol. 503, pp. 11-21.

Ministry of the Interior and Safety (MOIS). (2018). *The 2017 statistical yearbook of natural disaster*.

Mishra, N., Soni, H.K., Sharma, S., and Upadhyay, A.K. (2018). Development and analysis of artificial neural network models for rainfall prediction by using time-series data. *International Journal of Intelligent Systems and Applications*, Vol. 10, No. 1, pp. 16-23.

Montesarchio, V., Lombardo, F., and Napolitano, F. (2009). Rainfall thresholds and flood warning: An operative case study. *Natural Hazards and Earth System Sciences*, Vol. 9, pp. 135-144.

Park, S.S., and Kang, B.S. (2014). Differentiating scheme for the storm warning criteria considering the regional disaster prevention capacity. *J. Korean Soc. Hazard Mitig.*, Vol. 14, No. 5, pp. 67-76.

Powers, D.M. (2011). Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technoligies*, Vol. 2, No. 1, pp. 37-63.

Song, Y.S., Lim, C.H., Joo, J.G., and Park, M.J. (2016). A study on heavy rain forecast evaluation and improvement method. *J. Korean Soc. Hazard Mitig.*, Vol. 16, No. 2, pp. 113-121.

Wang, B.X., and Japkowicz, N. (2010). Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, Vol. 25, No. 1, pp. 1-20.

Zhai, A.R., and Jiang, J.H. (2014). Dependence of US hurricane economic loss on maximum wind speed and storm size. *Environmental Research Letters*, Vol. 9, No. 6, doi:10.1088/1748-9326/9/6/064019

Received	October 13, 2019
Revised	October 15, 2019
Accepted	November 11, 2019