J. Korea Water Resour. Assoc. Vol. 52, No. 12 (2019), pp. 963-973
doi: 10.3741/JKWRA.2019.52.12.963
pISSN 1226-6280
eISSN 2287-6138

Applicability study on urban flooding risk criteria estimation algorithm using cross-validation and SVM

Lee, Hanseung[a] · Cho, Jaewoong[b]* · Kang, Hoseon[c] · Hwang, Jeonggeun[d]

[a]Senior Researcher, Disaster Prevention Research Division, National Disaster Management Research Institute, Ulsan, Korea
[b]Research Officer, Disaster Prevention Research Division, National Disaster Management Research Institute, Ulsan, Korea
[c]Principal Researcher, Disaster Prevention Research Division, National Disaster Management Research Institute, Ulsan, Korea
[d]Researcher, Disaster Prevention Research Division, National Disaster Management Research Institute, Ulsan, Korea

Paper number: 19-057
Received: 30 July 2019; Revised: 14 November 2019 / 19 November 2019; Accepted: 19 November 2019

Abstract

This study reviews a urban flooding risk criteria estimation model to predict risk criteria in areas where flood risk criteria are not precalculated by using watershed characteristic data and limit rainfall based on damage history. The risk criteria estimation model was designed using Support Vector Machine, one of the machine learning algorithms. The learning data consisted of regional limit rainfall and watershed characteristic. The learning data were applied to the SVM algorithm after normalization. We calculated the mean absolute error and standard deviation using Leave-One-Out and K-fold cross-validation algorithms and evaluated the performance of the model. In Leave-One-Out, models with small standard deviation were selected as the optimal model, and models with less folds were selected in the K-fold. The average accuracy of the selected models by rainfall duration is over 80%, suggesting that SVM can be used to estimate flooding risk criteria.

Keywords: Risk criteria, Watershed characteristic, Limit rainfall, Support Vector Machine, Cross-validation

교차검증과 SVM을 이용한 도시침수 위험기준 추정 알고리즘 적용성 검토

이한승[a] · 조재웅[b]* · 강호선[c] · 황정근[d]

[a]국립재난안전연구원 방재연구실 선임연구원, [b]국립재난안전연구원 방재연구실 시설연구사,
[c]국립재난안전연구원 방재연구실 책임연구원, [d]국립재난안전연구원 방재연구실 연구원

요 지

본 연구는 도시침수 위험기준이 산정되지 않은 지역의 예·경보 기준을 예측하기 위해 유역특성 자료와 피해이력 기반으로 산정된 한계강우량을 활용하여 도시침수 위험기준을 추정하는 모델을 검토하였다. 위험기준 추정모델은 머신러닝 알고리즘의 하나인 Support Vector Machine을 이용하여 설계하였으며, 학습자료는 지역별 한계강우량과 유역특성으로 구성하였다. 학습자료는 정규화 한 후 SVM 알고리즘에 적용하였으며, SVM에 적용시 Leave-One-Out과 K-fold 교차검증 알고리즘을 이용하여 절대평균오차와 표준편차를 계산한 후 모델의 성능을 평가하였다. Leave-One-Out의 경우 표준편차가 작은 모델이 최적모델로 선정되었으며, K-fold의 경우 fold의 개수가 적은 모델이 선정되었다. 선정된 모델의 지속시간별 평균 정확도는 80% 이상으로 나타나 침수 위험기준 추정을 위해 SVM을 활용가능 할 것으로 판단된다.

핵심용어: 위험기준, 유역특성, 한계강우량, Support Vector Machine, 교차검증

***Corresponding Author.** Tel: +82-52-928-8174
E-mail: jwcho80@korea.kr (J. Cho)

1. 서 론

우리나라는 매년 태풍과 집중호우로 인하여 많은 피해가 발생하고 있다. 최근 10년간('08~'17) 주요 자연재난 원인을 피해액으로 분류하였을 때 태풍과 호우는 전체 피해액의 88.4%를 차지할 만큼 많은 피해를 발생시키는 자연재난이다(MOIS, 2018).

강우로 인한 자연재난을 대비하기 위해 많은 시스템이 운영되고 있으며, 시스템의 기본인 위험기준을 마련하기 위해 수리 및 수문 분석을 수행하고 있다. 수리, 수문 분석을 위해 해당 유역의 지형정보와 강우사상을 강우-유출 해석 프로그램에 입력하여 계산함에 있어 많은 노력과 시간이 필요하기 때문에 도시지역의 내수침수 위험기준은 방재시설의 설계기준을 적용하는 등 내수침수를 예측하기 위한 연구는 미흡한 실정이다.

도시침수 위험기준을 제시함은 침수분야 영향예보를 위한 것으로 기상현상 예보뿐만 아니라 기상현상으로 인하여 행정구역별 침수발생 여부를 같이 예보하는 것이다. 영향예보는 영국 대홍수(2007년) 이후 처음으로 등장하였으며, 필리핀을 내습한 태풍 '하이옌(2013년)'으로 인해 세계적으로 영향예보의 중요성을 확인하였다. 영국과 필리핀 자연재난의 공통점은 적절한 기상예측으로 인한 특보를 발령했음에도 많은 피해가 발생하였다는 것이다. 이에 기상예측기술의 발전에도 자연재해 피해가 계속된다는 논의가 이루어졌고, 기후변화로 인한 이상기후 외에 기상현상이 미치는 영향의 이해부족으로 피해가 발생한다고 결론 내려졌다. 즉, 극한기상현상이 발생하였을 경우 기상과 각 분야의 전문가가 아닌 사람들은 극한기상으로 인해 발생 가능한 문제를 예측할 수 없기에 기상에 의한 재난을 대비하지 못한다.

세계기상기구(World Meteorological Organization, WMO)에서는 위험기상으로 인한 예보를 정확히 예측하고 적절히 예경보를 했음에도 많은 피해가 발생하는 점에 대한 해결책으로 영향예보를 통한 재해 위험 관리를 제안하였다. 또한 '복합 재해영향기반 예특보서비스에 관한가이드라인'에서 영향예보의 중요성을 강조하고, 기상현상에 대한 위험기준추정 가이드라인을 제시하고 있다(WMO, 2015).

영국과 미국에서는 이미 영향예보를 시행하고 있으며, 기상청(Korea Meteorological Administration, KMA)에서도 영향예보의 생산 및 서비스를 위해 다부처 과제인 '자연재해대응 영향예보 생산기술 개발('18~'22)' 사업을 수행 중에 있다(KMA, 2016).

영향예보의 기준인 침수 위험기준을 설정하기 위한 강우-유출 해석 프로그램의 시간과 비용적 단점을 보완하기 위하여 수문학적 파라미터 분석이 유용한 것으로 입증된 SVM (Support Vector Machine)을 이용하여 침수위험기준을 추정하는 알고리즘을 설계하였다. SVM은 Misra *et al*. (2009)와 Behzad *et al*. (2009)에 의해 유역의 유출해석에 이용되었으며, Lin *et al*. (2009)은 SVM을 이용한 저수지 유입 예측모델을 제안하였다. Kim *et al*. (2012)은 SVM과 신경망 모델을 이용하여 일별 팬 증발량을 예측하였을 경우 신경망모형에 비해 SVM 모델의 성능이 우수성을 제시하였으며, Hipni *et al*. (2013)는 댐의 수위 예측을 위하여 SVM을 이용하는 등 수문학적 예측 및 해석에 SVM은 다양하게 활용되어왔다. 또한 이전 연구에서 도시침수 위험기준 추정을 위하여 SVM, Random Forest, Neuro-Fuzzy 및 Neural Network를 이용한 결과 SVM의 성능이 가장 좋게 나타났다(Cho *et al.*, 2018a; 2018b).

따라서 본 논문에서는 SVM의 커널함수(kernel function) 및 매개변수간의 비교를 중심으로 모델의 적용성을 검토하고자 한다. SVM의 침수 위험기준 추정 가능성과 성능을 평가하기 위하여 유역특성을 입력하여 한계강우량을 예측하는 모델을 설계하였으며, 위험기준인 한계강우량과 읍면동별 유역특성을 학습자료로 구축하였다. 한계강우량은 과거 침수피해이력을 이용하여 설정된 침수 위험기준으로 도시유역의 단기유출 특성을 반영하여 지속시간은 180분 이하로 설정하였으며, 유역특성은 강우-유출과 관련된 인자들을 사용하였다.

한계강우량 산정을 위하여 읍면동 단위의 침수 피해이력 자료를 수집한 후 침수로 인한 피해를 분리하여 피해를 발생시킨 강우를 분석하였다. 피해가 발생된 지역의 지형공간정보를 변환하여 유역특성 자료를 구축하였으며, 한계강우량과 동일 읍면동으로 매칭하여 학습자료를 구축하였다. 구축된 자료를 SVM에 학습시킨 후 학습된 모델에 유역특성을 입력하여 한계강우량을 예측하게 하였다. 예측된 한계강우량과 구축된 학습자료의 한계강우량을 비교하여 모델을 평가하였으며, 모델평가를 위하여 K-fold 교차검증 알고리즘을 활용하였다. 추정에 사용된 SVM 알고리즘은 Python 3.5.2 버전과 scikit-learn 라이브러리 및 MATLAB 인공지능 알고리즘 패키지를 활용하였다.

2. 인공지능 알고리즘의 학습자료 구축

학습자료는 과거 피해이력을 이용하여 산정된 지속시간별 한계강우량과 도시유역의 유출특성을 반영한 유역특성으로 구성되어있다. 여기서, 한계강우량이란 해당지역의 침수를 발생시키는 강우량으로 과거 침수피해이력을 기준으로 설정된 침수 위험기준이다. 한계강우량의 지속시간은 도시유역의 빠른 유출시간을 고려하여 30분, 60분, 180분으로 설정하

였으며, 유역특성은 관거밀도, 유역경사, 빗물받이 밀도 및 불투수면적률 4가지로 이루어져있다.

2.1 피해자료 및 강우자료 수집 및 분석

본 연구는 전국을 대상으로하여 진행되는 것으로 '18년 대상지역은 서울특별시, 5개 광역시(대전, 대구, 부산, 광주, 인천), 세종특별자치시, 경기도 및 강원도 지역으로 123개 시군구, 1,784개 읍면동이다. 과거침수 피해자료는 행정안전부 정보통신과에서 관리하는 국가재난관리시스템 사유시설 피해자료를 활용하였다. 강우자료는 종관기상관측시스템(Automated Synoptic Observing System, ASOS)과 자동기상관측시스템(Automatic Weather System, AWS)의 1분 단위 자료를 수집하였다. 대상지역내 기상관측소는 총 118개(ASOS: 7, AWS: 111)로 각 지점에 대한 강우데이터 파일을 수집하였으며, 피해자료와 강우자료는 비교적 환경적 변화가 빠른 도심지의 유출특성을 고려하여 최근 5년('13~'17)간의 자료를 수집하였다.

수집된 분 단위 강우자료를 이용하여 강우사상에 따른 지속시간별 최대강우량을 산정하였으며, 강우이벤트별 피해발생여부를 확인하여 피해가 발생된 강우와 피해가 발생하지 않은 강우로 구분하였다. 동일유역의 피해 최소강우와 미피해 최대강우의 차이가 20% 이내일 경우 피해 최소강우를 거듭제곱의 추세선으로 작성하여 한계강우량을 산정하였다. 피해최소강우와 미피해 최대강우의 차이가 20% 이상일 경우는 한계강우량의 예상범위 또한 커지게 되며, 최종적으로 한계강우량 추정결과의 정확성에 영향을 주기 때문에 20% 이상의 차이를 가지는 지역은 제외하였다(Fig. 1).

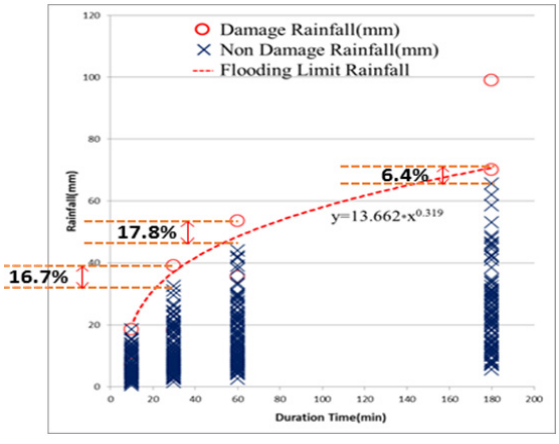


Fig. 1. Example of limit rainfall computation

Table 1. Range of basin characteristics parameter

	Culvert density	Impermeability rate (%)	Basin slope (%)	Inlet density
Maximum	0.08985	100.0	81.26	0.00833
Minimum	0.00000	0.0	0.00	0.00000
Average	0.01784	61.3	10.58	0.00096

2.2 유역특성 자료 수집 및 구축

도시유역의 강우-유출과 관련된 인자에는 유역의 면적, 경사, 관거, 불투수 면적, 빗물받이, 빗물 펌프장 등 다양한 것들이 있다. 이를 인공지능 알고리즘에 학습시키기 위하여 알고리즘이 요구하는 형태로 변환해 주어야 하지만 빗물받이 펌프장의 수방능력과 같은 인자들은 객관적인 수치를 확보할 수 없을 뿐만 아니라, 빗물펌프장 시설의 유무에 따라 한계강우량 예측결과에 영향이 크게 나타나므로 훈련자료로 활용하기 부적합하다고 판단하였다. 따라서 객관적인 수치로 변환할 수 있는 면적, 관거의 길이, 빗물받이 수, 경사, 불투수 면적을 학습자료로 사용하였다. 직경 600 mm 이상인 관거의 총길이, 빗물받이의 개수에 대하여 읍면동의 면적으로 나누어 관거밀도와 빗물받이 밀도로 변환하였으며, 유역의 경사와 불투수 면적률은 읍면동의 평균 경사를 이용하였다.

유역특성자료 구축에 필요한 지형공간정보 자료는 유역경사를 위한 수치지형도, 불투수율 산정을 위한 토지이용도 및 우수관거와 빗물받이의 위치 및 상세정보를 활용하였다. 지형공간정보는 GIS (Geographic Information System) 프로그램을 이용하여 모델 설계에 필요한 데이터를 추출하였다(Table 1).

2.3 학습자료 구성

대상지역의 1,784개 읍면동 중 피해이력기반 침수위험기준 추정 결과 총 671개 읍면동에 침수피해가 발생하였으며, 이 중 127개 읍면동에 대하여 한계강우량 추정이 가능하였다. 총 피해 읍면동 대비 18.9% 수준의 위험기준이 추정된 이유는 과거 피해를 발생시킨 강우량과 동일하거나 많은 강우가 발생하였더라도 개선사업으로 피해가 발생되지 않아 미피해 최대강우량이 피해 최소 강우량을 초과하였기 때문이다. 위험기준이 산정된 읍면동 중 지형공간정보가 전산화 되지 않은 지역을 제외하였을 경우 총 112개의 학습자료를 구축할 수 있었다.

2.4 데이터 전처리

알고리즘에 따라 매개변수 설정과 데이터 스케일에 민감하여 각 특성 값의 범위를 동일하게하거나 분포를 유사하게 변환하는 과정을 거쳐야 한다. 본 논문에서는 Eq. (1)을 이용

Table 2. Learning data formation using normalization

Normalization	District	basin characteristics				limit rainfall (mm)		
		Culvert density	Impermeability rate	Basin slope	Inlet density	30 Min.	60 Min.	180 Min.
Before	A	0.00860	1.20	0.01	0.00060	38.90	52.30	83.70
	B	0.05246	9.30	0.43	0.00208	32.80	47.10	83.20
	C	0.01217	70.57	18.50	0.00212	45.00	63.20	108.10
After	A’	0.01700	0.00	0.00	0.00000	0.29	0.18	0.08
	B’	0.13000	0.08	0.01	0.03700	0.08	0.07	0.08
	C’	0.02630	0.70	0.47	0.03800	0.49	0.38	0.26

하여 변수의 범위를 0~1 사이로 변환시키는 정규화 과정을 적용한 후 학습자료로 사용하였다(Table 2.).

$$X_i = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

3. Support Vector Machine

SVM은 Vladimir Vapnik (Vapnik, 1995)에 의해서 제안된 머신러닝 알고리즘으로 일반화 능력이 뛰어난 분류기로 알려져 있다. 분류(classification), 회귀(regression) 및 특이점 판별(outlier detection)에 사용되는 알고리즘으로 N차원의 공간에 다양한 유형의 집합이 주어지면 마진(margin)을 최대로 하는 N-1 차원의 초평면(hyperplane)을 이용하여 주어진 집합을 여러 그룹으로 분리한다. 여기서 마진이란 학습데이터 중 경계에서 가장 가까운 데이터로부터 분류경계까지의 거리를 말한다. 데이터 분류시 선형적 분류뿐만 아니라 비선형적인 초평면을 찾을 수 있는 고차원 공간으로도 분류가 가능하다(Fig. 2).

최상의 초평면을 계산하는 것은 강제적인 최적화 문제를 제시하는 것이고 이차적 프로그래밍 기술들로 풀어진다. 이러한 초평면은 최적 경계 초평면(Optimal Separating Hyperplane)이라고 불린다(Chang, 2006).

초평면 계산시 선형으로 분리할 수 없는 저차원의 데이터를 고차원의 공간으로 매핑시켜 분류할 수 있지만 계산량이 증가하는 문제가 발생하기 때문에 커널 함수를 적용하여 이를 해결한다. 커널함수는 여러 표본간 유사도를 측정하는 기준으로 선형함수(Linear function), 다항함수(Polynomial function), RBF (Radial basis function) 등이 있다(Table 3.).

위 표에서 선형함수는 가장 단순한 커널함수로 x와 y의 내적과 상수 c의 합으로 이루어진다. 다항함수는 비정상 커널(non-stationary kernel)로 전체 데이터가 정규화(normalized)된 경우 적용성이 좋으며 고차원에서 모델을 학습시킬 수 있

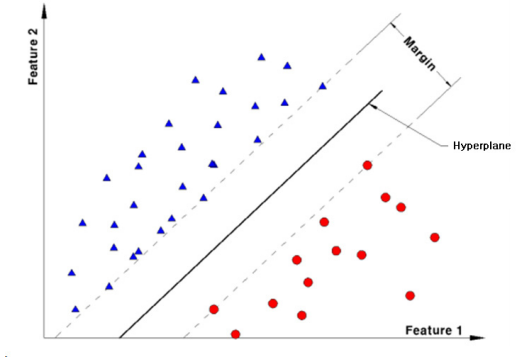


Fig. 2. Classification concept using SVM

Table 3. Governing equation of kernel

Kernel	Equation
Linear	$k(x_1, x_2) = x^T y + c$
Polynomial	$k(x,y) = (ax^T y + c)^d$
RBF	$k(x,y) = \exp(-r\|\|x-y\|\|^2)$

다. 다항함수식에서 a와 c는 기울기와 상수이며 d는 다항식의 차수이다. RBF는 무한차원의 특성공간에 매핑하는 것으로 커널 매개변수(kernel parameter)인 γ에 의한 식으로 표현된다. γ는 SVM의 정칙화 매개변수(regularization parameter)인 cost와 함께 모델의 허용오차를 결정하는 매개변수이다. cost가 커지면 각 데이터들이 모델에 큰 영향을 미쳐 마진의 폭이 좁아지고 결정경계를 비선형적하여 정확하게 분류한다. cost가 작아지면 각 포인트의 영향력이 작아 마진의 폭은 넓어져 제약이 큰 모델을 만든다. 즉, cost는 분류시 다른 클래스에 허용 가능한 데이터의 양을 결정하며, γ는 각 데이터가 영향력을 행사하는 거리이다.

SVM은 분류와 회귀에 적용 가능한 알고리즘으로 분류모델에서는 SVC (Support Vector Classification), 회귀모델에서는 SVR (Support Vector Regression)을 이용하며, 본 논문에서는 SVR을 이용하였다.

4. 교차검증을 이용한 모델평가

학습자료가 결정된 후에는 인공지능 알고리즘이 훈련자료에 대한 성능을 높이고 새로운 자료를 잘 처리해야하므로 전체자료를 훈련자료와 검증자료로 나누어 학습알고리즘에 적용한다. 모델의 성능 검토 시 훈련자료와 검증자료를 한번 나누는 것보다 자료를 여러 형태로 나누어 자료들이 최소한 한 번은 검증 또는 훈련자료로 사용되도록 한 후 모델에 학습시켜 평가하는 방법인 교차검증을 활용하여야 한다. 자료를 무작위로 한번만 나누었을 경우 훈련자료에 예측이 어려운 자료 또는 예측이 쉬운 자료만 남게 되면 모델 성능은 매우 좋거나 나쁜 결과를 제시할 수 있기 때문이다.

본 논문에서는 K-fold와 Leave-One-Out 교차검증 알고리즘을 사용하여 예측결과와 검증자료의 절대평균오차(Mean Absolute Error, MAE), 정확도(Accuracy) 및 표준편차(Standard deviation)를 계산한 후 모델을 평가하였다.

4.1 K-fold

K-fold는 전체자료를 k개의 집합으로 나누어 k-1개의 집합을 훈련자료로 사용하고 나머지 1개의 집합을 검증자료로 사용하는 방법으로 이러한 과정을 k번 반복한다. Fig. 3.과 같이 전체자료를 동일한 크기의 5개 집합으로 나누었을 경우, 첫 번째 경우는 첫 번째 집합을 제외한 나머지 4개의 집합들로 훈련한 후 첫 번째 자료로 검증하여 모델의 성능을 평가한다. 두 번째 경우는 두 번째 집합을 검증자료, 나머지 집합을 훈련자료로 사용하여 학습하는 것으로 이를 총 5번 반복하여 전체 학습자료에 대해 평가를 한다.

학습과 예측을 k번 반복함에 있어 각각의 결과가 서로 상이할 수 있기 때문에 표준편차를 모델의 평가에 적용하였고 표준편차가 작은 모델일수록 안정적인 모델이라 판단할 수 있다. 이러한 방법으로 모델을 평가하게 되면 모든 데이터들이 학습과 검증자료에 사용되어 과적합(overfitting)이 일어날 확률이 낮다는 장점이 있어 k-fold 교차검증 방법은 일반화 성능을 만족시키는 최적의 매개변수를 구하기 위한 모델 튜닝에 사용된다. K-fold 사용시 위 그림과 같이 집합의 개수가 5인 경우는 학습비율이 80.0%인 조건과 동일하며, fold의 수가 3인 경우는 66.7%, 10인 경우는 90.0%으로 볼 수 있다.

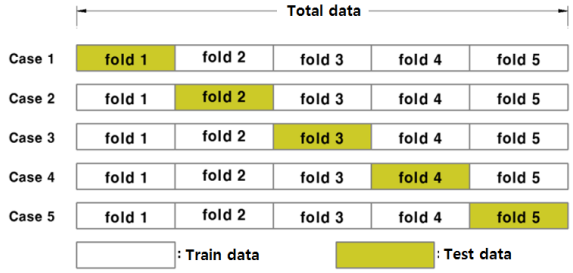


Fig. 3. K-fold cross-validation

4.2 Leave-One-Out

Leave-One-Out(LOO)은 K-fold에서 집합의 개수를 전체자료 개수로 설정한 것과 마찬가지로 전체자료에서 1개의 자료만을 검증자료로 사용하고 나머지를 훈련자료로 사용하는 것으로 전체자료 개수만큼 반복 계산하여 평가하는 방법이다. 자료의 양이 많을 경우 계산시간이 증가되지만 자료의 수가 적을 때 좋은 결과를 제시할 수 있다.

4.3 모델 평가방법

K-fold 교차검증시 학습자료의 순서가 바뀜으로 인하여 예측결과가 상이할 수 있기 때문에 전체 자료의 순서를 무작위로 변경하면서 1,000번 예측하였으며, Eqs. (2) and (3)을 사용하여 예측치와 실측치에 대한 절대평균오차 및 표준편차를 기준으로 성능을 평가하였다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|t_i - p_i| \tag{2}$$

$$Standard\,deviation = \sqrt{\frac{\sum_{i=1}^{n}(MAE_i - \overline{MAE})^2}{n-1}} \tag{3}$$

여기서, n은 자료의 개수, t_i는 실제값, p_i는 예측값, $\overline{MAE}$는 1,000회 반복된 절대평균오차(MAE)의 평균이다. 절대평균오차는 학습자료 선택의 무작위성으로 인하여 예측을 반복할 때마다 크기가 달라진다. 달라지는 오차의 범위가 넓다면 그 모델은 불안정한 결과를 제시한다고 판단할 수 있다. 반대로 절대평균오차의 표준편차가 작다는 것은 반복 계산된 절대평균오차가 유사한 결과를 가진다는 것이고 좁은 범위 안에서 결과들이 도출되어 안정적인 모델이라는 결론을 내릴 수 있다. 따라서 1,000회 반복하여 계산된 절대평균오차의 평균에 1,000개의 절대평균오차에 대한 표준편차를 구하여 두 결과를 더한 값이 해당 모델이 평균적인 또는 일반적인 오차라 판단할 수 있다. 이는 동일한 단위(mm)로 표현된 두 결과에 대하여 평균값과 그 평균값이 가질 수 있는 범위내의 상한선으로 표준편차와 절대평균오차의 합이 최소가 되는 모델을 최적모델로 선정하였다.

5. 모델별 성능평가 결과

SVM 알고리즘을 이용하여 매개변수를 변경하면서 한계강우량을 예측하였다. SVM의 주요 매개변수는 Kernel, cost 및 γ로 Kernel은 Linear, Polynomial 및 RBF를 사용하였으며 Cost와 γ는 10의 등비수열로 0.01~100.0까지 5가지의 경우로 설정하였다. K-fold 알고리즘 사용시 fold의 수는 3, 5, 10을 적용하였다.

5.1 Linear Kernel을 이용한 모델 설계 및 평가

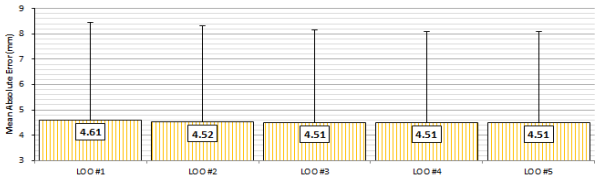
Linear kernel은 γ가 적용되지 않기 때문에 cost 만을 변경하였으며 fold의 수는 3, 5, 10으로 설정하여 LOO의 경우 5개, K-fold는 15개 모델에 대한 성능을 비교하였다. 모델별 매개변수는 Table 4에 제시하였다.

Fig. 4는 Linear kernel과 LOO를 이용하여 한계강우량을 예측한 결과를 나타낸 것으로, 지속시간 30분 한계강우량에 대한 교차검증을 수행하였을 경우 cost 값이 작아질수록 모델의 성능이 좋아지는 것을 확인할 수 있었으며, 60분은 #2, 180분은 #1 모델을 기준으로 cost가 변화함에 따라 오차는 증가 또는 감소하였다.

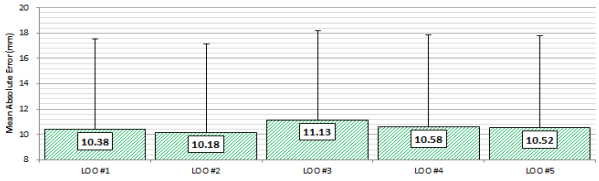
Fig. 5는 K-fold를 이용하여 지속시간별 한계강우량에 대한 교차검증을 수행한 결과로 fold의 수가 많아질수록 절대평균오차는 감소하거나 유사한 반면 표준편차는 커졌으며, 30분의 경우 cost 값이 작아질수록 오차는 감소하는 것으로 나타났다. 60분과 180분의 경우는 30분과 반대로 cost 값이 증가할수록 오차는 증가하였으며, cost에 따른 표준편차의 차이는 크지 않았다.

Table 4. Parameter per model (Linear kernel)

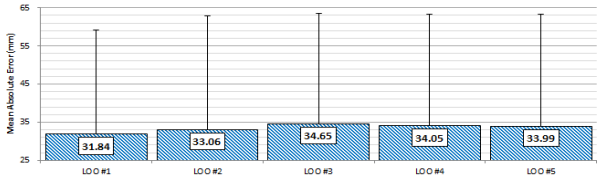
Case	#1~#5	10-1~10-5	5-1~5-5	3-1~3-5
Fold	-	10	5	3
cost	100~0.01	100~0.01	100~0.01	100~0.01



(a) Duration 30 min



(b) Duration 60 min

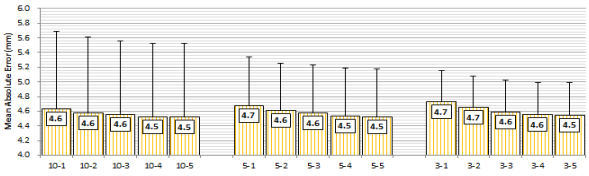


(c) Duration 180 min

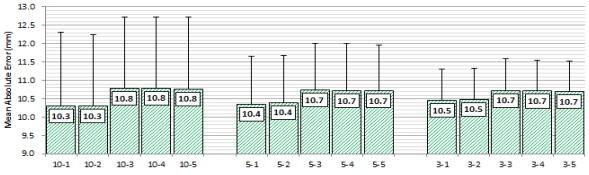
Fig. 4. Estimate result (Linear kernel, LOO)

5.2 Polynomial Kernel을 이용한 모델 설계 및 평가

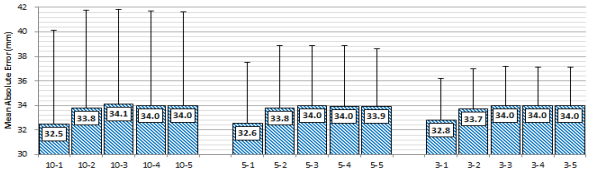
Polynomial kernel은 다항함수을 이용한 것으로 함수의 차수를 설정해 줄 수 있어 2차 함수로 설정하였으며, LOO는 25개, K-fold는 75개 모델의 결과를 비교하였다. 모델별 매개변수는 Table 5에 제시하였다.



(a) Duration 30 min, K-fold



(b) Duration 60 min, K-fold

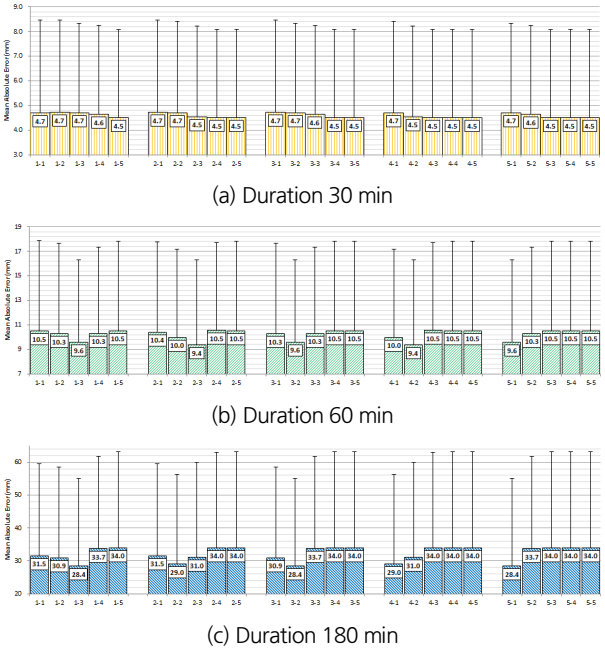


(c) Duration 180 min, K-fold

Fig. 5. Estimate result (Linear kernel, K-fold)

Table 5. Parameter per model (Polynomial, RBF)

Case	cost	γ
1-1~1-5	100	100.0~0.01
2-1~2-5	10	
3-1~3-5	1	
4-1~4-5	0.1	
5-1~5-5	0.01	



(a) Duration 30 min

(b) Duration 60 min

(c) Duration 180 min

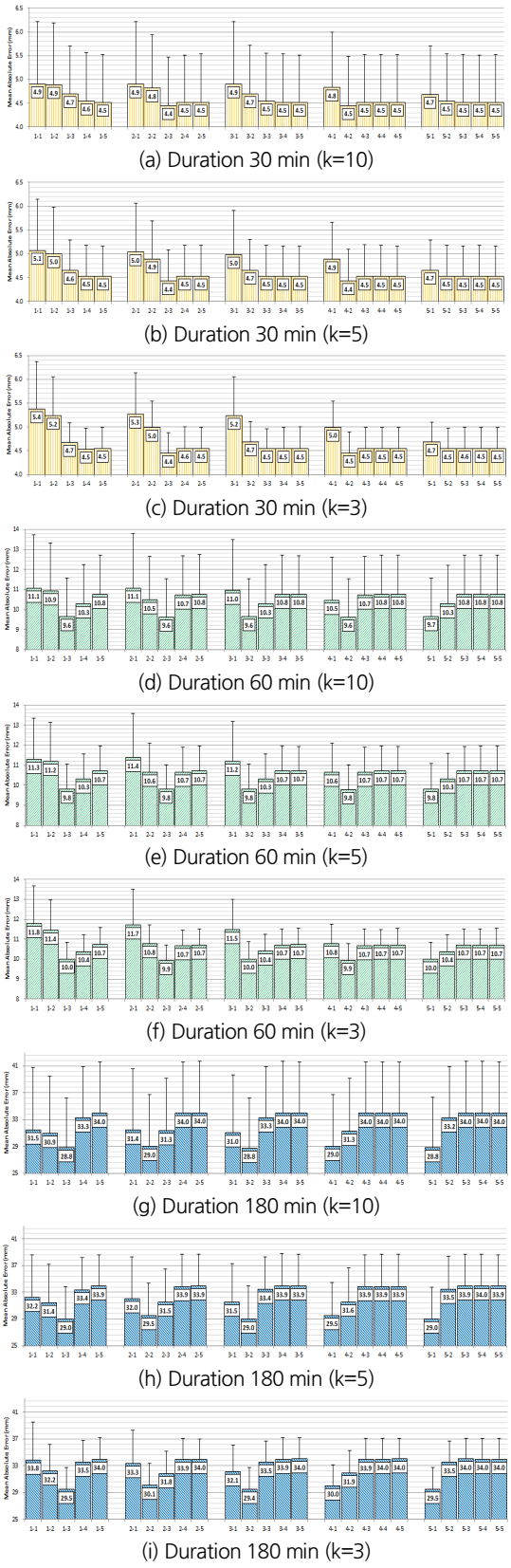
Fig. 6. Estimate result (Polynomial kernel, LOO)

Fig. 6은 Polynomial kernel과 LOO를 이용하여 한계강우량을 예측한 결과를 나타낸 것으로, 지속시간 30분의 경우 γ 값이 감소함에 따라 절대평균오차와 표준편차는 감소하였다. cost에 따른 절대평균오차의 차이는 크지 않은 반면 표준편차는 유지 또는 감소하는 것으로 나타났다. 60분과 180분의 경우 cost가 100.0~10.0에서는 γ값이 1에 가까울수록 좋은 성능을 나타냈으며, γ가 커짐에 따라 cost가 작아질수록 좋은 성능을 나타냈다.

Fig. 7은 Polynomial kernel과 K-fold를 이용하여 한계강우량을 예측한 결과를 나타낸 것으로, 지속시간 30분의 경우 cost와 γ가 작아질수록 표준편차와 절대평균오차는 감소하거나 유사한 것으로 나타났다. 60분과 180분의 경우 cost가 큰 경우 γ가 1일 때 좋은 성능이었지만 cost가 작아짐에 따라 γ가 커질수록 좋은 성능을 나타냈다. 전체적으로 fold의 수가 작아질수록 절대평균오차는 증가한 반면 표준편차는 감소하는 것을 확인 할 수 있다.

5.3 RBF Kernel을 이용한 모델 설계 및 평가

RBF kernel의 매개변수는 함수의 차수 외에 Polynomial kernel과 동일하게 설정하였다. Fig. 8은 RBF kernel과 LOO를 이용하여 한계강우량을 예측한 결과를 나타낸 것으로, 지속시간 30분의 경우 cost가 작아짐에 따라 표준편차와 절대평균오차는 감소하는 것으로 나타났다. 60분과 180분의 경우 cost가 큰 모델일수록 γ가 1.0인 경우 좋은 성능을 나타냈



(a) Duration 30 min (k=10)

(b) Duration 30 min (k=5)

(c) Duration 30 min (k=3)

(d) Duration 60 min (k=10)

(e) Duration 60 min (k=5)

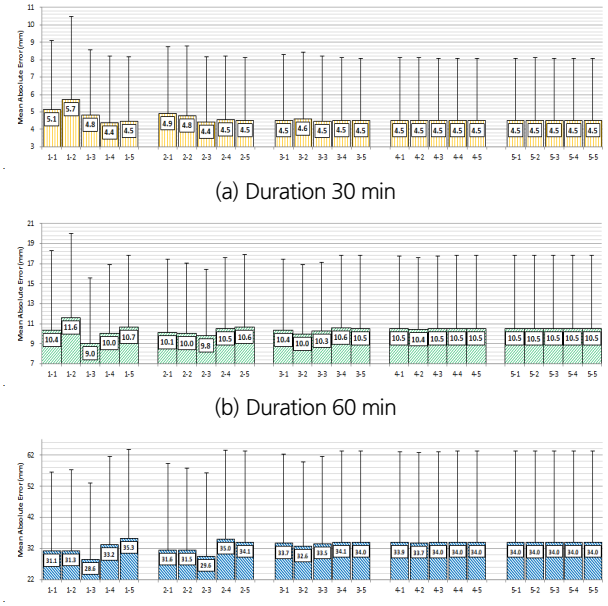
(f) Duration 60 min (k=3)

(g) Duration 180 min (k=10)

(h) Duration 180 min (k=5)

(i) Duration 180 min (k=3)

Fig. 7. Estimate result (Polynomial kernel, k-fold)



(a) Duration 30 min

(b) Duration 60 min

(c) Duration 180 min

Fig. 8. Estimate result (RBF kernel, LOO)

다. 전체적으로 cost가 0.1 이하일 경우 γ에 따른 절대평균오차의 차이는 크지 않았다.

Fig. 9는 RBF kernel과 K-fold를 이용하여 한계강우량을 예측한 결과를 나타낸 것으로, cost가 10.0 이상인 모델에서 성능의 편차가 큰 것으로 나타났으며 cost가 작아질수록 유사한 성능을 나타냈다. 전체적으로 모델별 성능이 유사하게 나타나는 범위보다는 편차가 크게 나타나는 범위에서 좋은 성능을 나타내는 모델을 선택할 수 있었다.

5.4 모델별 결과 비교분석

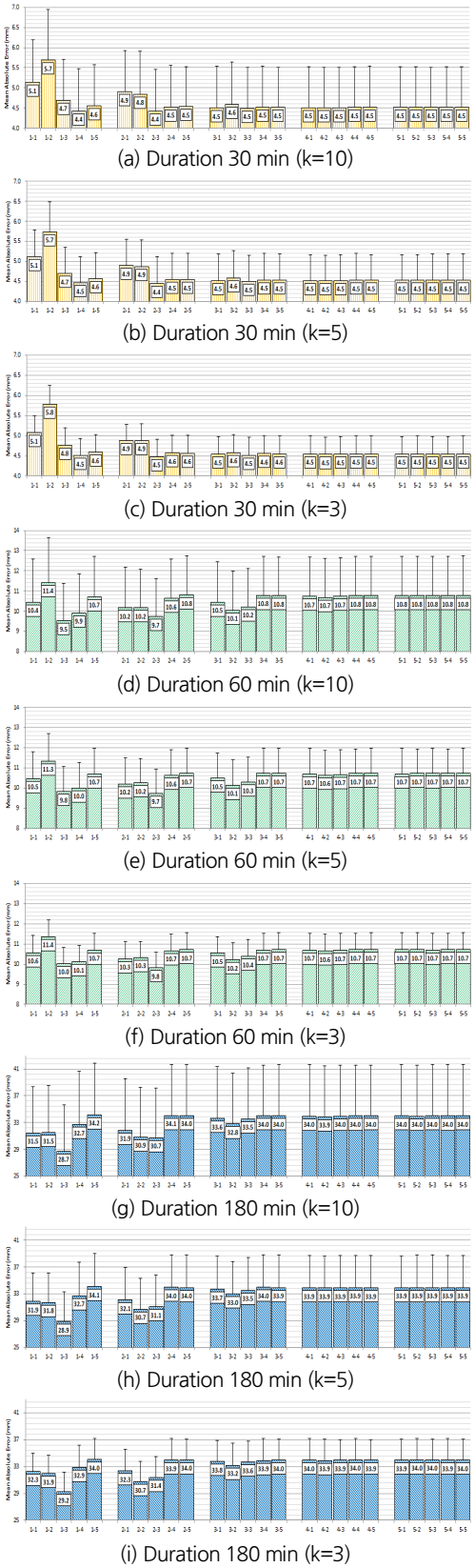
위험기준 예측을 위해 사용된 SVM의 kernel과 매개변수별 예측 결과를 비교하였다. 비교 기준은 표준편차와 절대평균오차의 합이 최소가 되는 모델을 선정하였으며 각 지속시간별 한계강우량의 평균값을 기준으로 오차에 대한 정확도를 Eq. (4)를 이용하여 평가하였다.

$$Accuracy = 1 - \frac{1}{n}\sum_{i=1}^{n}\left|\frac{m_i - e_i}{m_i}\right| \quad (4)$$

여기서, m_i는 지속시간별 한계강우량의 평균, e_i는 선정된 모델의 절대평균오차 또는 절대평균오차와 표준편차의 합이다.

5.4.1 Linear kernel

SVM의 Linear kernel을 이용하여 한계강우량을 예측하였



(a) Duration 30 min (k=10)

(b) Duration 30 min (k=5)

(c) Duration 30 min (k=3)

(d) Duration 60 min (k=10)

(e) Duration 60 min (k=5)

(f) Duration 60 min (k=3)

(g) Duration 180 min (k=10)

(h) Duration 180 min (k=5)

(i) Duration 180 min (k=3)

Fig. 9. Estimate result (RBF kernel, k-fold)

을 때 가장 좋은 성능을 나타낸 모델을 Tables 6 and 7에 제시하였다. K-fold에서는 학습률이 낮은 모델들이 선정되었으며, LOO와 K-fold 모두 지속시간이 길어질수록 cost가 큰 모델의 성능이 좋아지는 것을 확인할 수 있다.

교차검증 방법에 따른 지속시간 60분의 최적모델은 LOO와 K-fold의 Cost값이 서로 다르게 나타났다. 교차검증 알고리즘별 최적모델이 상이하다 하더라도 두 경우 모두 절대평균오차와 표준편차의 합이 두 번째로 작은 모델이 교차되어 최적의 성능을 나타낸바 Case 3-1, 2 모두 우수한 성능을 나타낸 모델이라 판단할 수 있다(Table 8).

SVM의 Linear kernel 알고리즘을 사용하였을 경우 최적모델의 정확도를 평가하였을 때 LOO의 절대평균오차와 표준편차의 합을 제외하고는 평균적으로 목표성능인 정확도 80%이상을 보여주었으며, 절대평균오차의 경우 K-fold 보다 LOO에서 정확도가 높았으며 절대평균오차와 표준편차의 합에서는 LOO의 높은 학습률로 인해 K-fold의 정확도가 높게 나타났다(Table 9).

Table 6. Optimal model result (Linear, LOO)

(Unit: mm)

Duration (Min.)	Case	①	②	①+②
30	5	4.51	3.58	8.09
60	2	10.18	6.97	17.15
180	1	31.84	27.24	59.09

①: MAE, ②: Standard deviation

Table 7. Optimal Model Result (Linear, K-fold)

(Unit: mm)

Duration (Min.)	Case	①	②	①+②
30	3-4	4.56	0.44	5.00
	3-5	4.55	0.45	5.00
60	3-1	10.45	0.86	11.31
180	3-1	32.81	3.41	36.21

①: MAE, ②: Standard deviation

Table 8. Optimal Model Comparison per Cross- validation (Linear)

(Unit: mm)

Duration (Min.)	LOO		K-fold	
	Case	①+②	Case	①+②
60	2	17.15	3-1	11.31
	1	17.53	3-2	11.34

①: MAE, ②: Standard deviation

Table 9. Accuracy per Cross-validation (Linear)

(Unit: %)

Duration (Min.)	LOO		K-fold	
	Ⓐ	Ⓑ	Ⓐ	Ⓑ
30	89.85	81.78	89.73	88.74
60	84.51	73.90	84.10	82.79
180	74.42	52.52	73.64	70.91
AVE.	82.92	69.40	82.49	80.81

①: MAE, ②: MAE+Standard deviation

5.4.2 Polynomial Kernel

SVM의 Polynomial kernel을 이용하여 한계강우량을 예측하였을 때 가장 좋은 성능의 모델을 Tables 10 and 11에 제시하였다. LOO에서는 γ가 1.0에 가까운 모델의 성능이 좋게 나타났으며 K-fold에서는 학습률이 낮고 γ는 1.0, cost는 10.0 이상인 모델의 성능이 좋게 나타났다.

교차검증 방법에 따른 지속시간 30분 최적모델 선정시 K-fold와 LOO의 결과가 다르게 나타났다. 이 경우 매개변수의 범위를 세분화하거나 확대하여 모델을 선정할 필요가 있다.

SVM의 Polynomial kernel 알고리즘을 사용하였을 경우 최적 모델의 정확도를 평가하였을 때 LOO의 표준편차와 절대평균오차의 합을 제외하고는 평균적으로 정확도 80% 이상이었다. 절대평균오차에서 지속시간 30분을 제외하고는

Table 10. Optimal model result (Polynomial, LOO)

(Unit: mm)

Duration (Min.)	Case	①	②	①+②
30	2-4, 4-3	4.51	3.58	8.08
60	2-3, 4-2	9.36	6.93	16.29
180	③	28.45	26.59	55.04

①: MAE, ②: Standard deviation, ③: 1-3, 3-2, 5-1

Table 11. Optimal model result (Polynomial, K-fold)

(Unit: mm)

Duration (Min.)	(Fold) Case	①	②	①+②
30	(3) 2-3	4.45	0.43	4.88
60	(3) 2-3	9.92	0.80	10.72
180	(3) 1-3	29.46	3.24	32.70

①: MAE, ②: Standard deviation

Table 12. Accuracy per cross-validation (Polynomial)

(Unit: %)

Duration (Min.)	LOO		K-fold	
	Ⓐ	Ⓑ	Ⓐ	Ⓑ
30	89.85	81.81	89.98	89.01
60	85.76	75.21	84.90	83.69
180	77.14	55.78	76.33	73.73
AVE.	84.25	70.93	83.74	82.14

①: MAE, ②: MAE+Standard deviation

K-fold 보다 LOO에서 정확도가 높았으며 절대평균오차와 표준편차의 합에서는 K-fold의 정확도가 높게 나왔다(Table 12).

5.4.3 RBF Kernel

SVM의 RBF kernel을 이용하여 한계강우량을 예측하였을 경우 성능이 가장 좋은 모델을 Tables 13 and 14에 제시하였다. LOO에서는 γ가 1.0인 모델들의 성능이 좋게 나타났으며, 60분 이상에서는 cost가 100.0인 모델들의 성능이 좋게 나타났다. K-fold의 결과, LOO와 마찬가지로 γ가 1.0인 모델의 성능이 좋게 나타났으며 학습률이 낮고 cost는 10.0 이상인 모델들이 최적 모델로 선정되었다.

최적모델 선정시 지속시간 30분과 60분에 해당하는 모델은 K-fold의 결과와 LOO의 결과가 다르게 나타났다. 지속시간 60분인 경우 LOO Case 2-3 모델은 1-3 모델 다음으로 성능이 좋은 모델로 지속시간 60분에 해당하는 최적모델은 fold 3의 Case 2-3 모델을 선정할 수 있을 것이라 판단되며 30분의 경우 선정된 모델 사이에 많은 모델들이 있어 매개변수를 세분화하거나 확대하여 최적모델을 선정할 필요가 있다(Table 15).

SVM의 RBF kernel 알고리즘을 사용하였을 경우 최적 모델의 정확도를 평가하였을 때 LOO의 절대평균오차와 표준편차의 합을 제외하고는 평균적으로 목표성능인 정확도 80%이상을 보여주었으며 절대평균오차의 경우 지속시간 30분을 제외하고는 K-fold 보다 LOO에서 정확도가 높았으며 절대평균오차와 표준편차의 합에서는 K-fold의 정확도가 높게 나왔다(Table 16).

Table 13. Optimal model result (RBF, LOO)

(Unit: mm)

Duration (Min.)	Case	①	②	①+②
30	③	4.50	3.58	8.09
60	1-3	9.01	6.52	15.53
180	1-3	28.55	24.44	52.99

①: MAE, ②: Standard deviation, ③: 3-5, 4-3~5, 5-1, 5-3~5

Table 14. Optimal model result (RBF, K-fold)

(Unit: mm)

Duration (Min.)	(Fold) Case	①	②	①+②
30	(3) 2-3	4.48	0.43	4.91
60	(3) 2-3	9.81	0.77	10.57
180	(3) 1-3	29.21	2.96	32.17

①: MAE, ②: Standard deviation

Table 15. Optimal model comparison per cross- validation (RBF)

(Unit: mm)

Duration (Min.)	LOO		K-fold	
	Case	①+②	(Fold) Case	①+②
60	1-3	15.53	(3) 2-3	10.57
	2-3	16.38		

①: MAE, ②: Standard deviation

Table 16. Accuracy per cross-validation (RBF)

(Unit: %)

Duration (Min.)	LOO		K-fold	
	Ⓐ	Ⓑ	Ⓐ	Ⓑ
30	89.87	81.78	89.91	88.94
60	86.29	76.37	85.07	83.91
180	77.06	57.42	76.53	74.15
AVE.	84.41	71.86	83.84	82.34

①: MAE, ②: MAE+Standard deviation

6. 결 론

본 논문은 침수 위험기준 추정시 인공지능 알고리즘의 활용 가능성을 확인하기 위하여 유역특성을 입력하여 한계강우량을 예측할 수 있는 알고리즘을 설계하였다. 먼저 대상지역에 구축된 자료(피해자료, 강우자료)를 이용하여 한계강우량 산정 가능유무를 판단한 후 산정가능 지역에 대해서만 한계강우량을 산출하였다. 한계강우량이 산출된 지역을 기준으로 유역특성을 포함하여 총 112개의 학습자료를 정규화 한 후 SVM 알고리즘에 적용하였으며, 교차검증을 이용하여 절대평균오차와 표준편차를 합한 값을 기준으로 모델의 성능을 평가하였다.

kernel별, 지속시간별로 매개변수에 따른 결과가 일정하게 변하거나 유사한 성능을 나타냈다. LOO에서는 표준편차가 작은 모델이 최적 모델로 선정되었으며, K-fold에서는 fold의 개수가 적은 모델이 많이 선정되었다.

절대평균오차와 표준편차의 합으로 최적모델 결과를 평가하였을 경우, K-fold에서는 지속시간 30분과 60분 결과가 모든 kernel에서 80% 이상의 정확도인 반면, LOO를 이용한 결과는 지속시간 30분에서 80% 이상의 정확도로 나타났다. 절대평균오차만으로 목표성능을 검토하였을 경우 LOO의 지속시간 60분 결과가 추가적으로 정확도 80% 이상의 성능을 만족하였다. LOO의 절대평균오차 기준 정확도는 지속시간별 평균 84.41%로 K-fold와 유사한 결과를 나타냈지만 표준편차를 고려한 오차에서 LOO는 K-fold와 달리 1대 다의 검증이 학습자료의 개수만큼 반복됨에 따라 표준편차가 증가하여 71.86%로 낮은 성능을 확인할 수 있었다.

침수 위험기준 추정을 위하여 SVM 알고리즘을 활용한 결과 교차검증 방법별 최고성능의 평균 정확도가 80%이상의 성능을 나타내어 적용 가능성을 확인할 수 있었다. 지속시간 30분과 60분 대비 180분의 성능이 떨어지는 것으로 나타나 상대적으로 긴 지속시간의 성능을 높일 수 있는 유역특성 인자를 추가하거나 특성간의 상관성을 높일 수 있는 데이터 전처리 방법을 활용하여 모델의 성능을 높여야 될 것으로 판단된다.

감사의 글

본 연구는 행정안전부 '재난안전관리업무지원기술개발' (NDMI-주요-2018-09-01)의 연구비지원에 의해 수행되었습니다.

References

Behzad, M., Asghari, K., Eazi, M., and Palhang, M. (2009). "Generalization performance of support vector machines and neural networks in runoff modeling, expert systems with applications." *An International Journal*, Vol. 36, No. 4, pp. 7624-7629.

Chang, J.G. (2006). "Real-time vehicle recognition mechanism using support vector machines." *Journal of Korea Academia Industrial Cooperation Society*, Vol. 7, No. 6, pp. 1160-1166.

Cho, J.W., Choi, C.W., Kang, H.S., Lee, H.S., Bae, C.Y., Hwang, J.G., and Bae, S.J. (2018b). *Deep learning based urban flood alert critrria estimation model design. NDMI-PR-2018-09-01-01*. National Disaster Management Research Institute, Ulsan.

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

Hipni, A., El-shafie, A., Najah, A., Karim, O.A., Hussain, A., and Mukhlisin, M. (2013). "Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS)." *Water Resources Management*, Vol. 27, No. 10, pp. 3803-3823.

Kim, S., Shiri, J., and Kisi, O. (2012). "Pan evaporation modeling using neural computing approach for different climatic zones." *Water resources management*, Vol. 26, No. 11, pp. 3231-3249.

Korea Meteorological Administration (KMA) (2016). *Planning study on introduction plan of influence forecast*.

Lin, G.F., Chen, G.R., Huang, P.Y., and Chou, Y.C. (2009). "Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods." *Journal of hydrology*, Vol. 372, No. 1-4, pp. 17-29.

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

Misra, D., Oommen, T., Agarwal, A., Mishra, S.K., and Thompson, A.M. (2009). "Application and analysis of support vector machine based simulation for runoff and sediment yield." *Biosystems Engineering, Biosystems Engineering*, Vol. 103, No. 4, pp. 527-535.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer, Verlag New York.

World Meteorological Organization (WMO) (2015). *Guidelines on multi-hazard impact-based forecast and warning services*. No. 1150, Swiss Geneva.