https://doi.org/10.7840/kics.2022.47.10.1704

미기상 데이터의 클래스 불균형을 해결한 서리 예측 모델

문 애 경◆, 김 효 선°

Microclimate-Based Frost Prediction Model Resolving the Class Imbalance

Ae-Kyeung Moon◆, Hyo-Seon Kim°

요 약

미기상은 모든 형태의 시간에 민감한 농업에 영향을 미치고 있다. 신기술 및 구현 기술이 크게 발전함에 따라 방대한 양의 IoT 기반 환경 데이터를 통해 시간에 민감한 농업 서비스에 유용한 정보를 제공함으로써, 부정적인 기후 변화 영향에 대비할 수 있다. 고위험 기상 조건 중 특히 우려되는 예상치 못한 서리 피해는 농업 수확량에 상당한 영향을 미친다. 본 논문은 환경 데이터를 활용한 머신러닝 기반의 적기 서리 예측 모델을 제안한다. 기존 접근 방식은, 서리 발생에 대한 소수의 클래스 라벨링 정보로 인해, 데이터 불균형 문제를 가진다. 그래서 본 논문에서는 예측 서비스가 실시간으로 실행될 때, IoT 스테이션에서 수집된 환경 데이터셋을 활용하여, SMOTE(Synthetic Minority Over-sampling Technique) 방법에 의한 클래스 불균형을 해결한 서리 예측 모델을 제시한다. 실험 결과는 Random Forest 알고리즘이 서리 예측에 가장 적합한 알고리즘으로 선정되었다. 최적화 단계를 통하여 서리 예측 모델의 성능이 평균 4% 정도 향상(f1 기준)되었다. 또한 SMOTE 비율별 성능 평가는 각 성능 지표별 특유의 경향성을 보였고, 이것은 적절한 비율을 사용하는 것이 중요함을 나타낸다.

키워드 : 스마트팜, 서리예측, 농업기상, 미기상, 디지털 농업
Key Words : Smart Farm, Frost Prediction, Agricultural Climate, Microclimate, Digital Agriculture

ABSTRACT

Microclimate has been influencing all forms of time-sensitive agriculture. With substantial advances in emerging and enabling technologies, a vast amount of IoT-based environmental data allows preparation for the adverse impacts by providing helpful information to time-sensitive services. Of particular concern among high-risk weather conditions is nonanticipative frosty damage, affecting agricultural yield significantly. This paper proposes a timely frost prediction model based on machine learning using environmental data. Because of minority information on frost, conventional approaches often suffer from the class imbalance problem with rare labeling data. We address these issues through a frost prediction model using class-balanced data by SMOTE method to environmental datasets collected from IoT stations when predictive service executes. Our experimental results demonstrate that the frost prediction using Random Forest is the most suitable algorithm. With the optimization process, the performance of the frost prediction model was improved by about 4% (Based on f1). Moreover, the performance evaluation by SMOTE ratio shows the importance of an appropriate ratio for data augmentation by unique tendency.

※ 본 연구 논문은 한국전자통신연구원 내부연구과제[22YD1100] 및 안동시 지원사업[22AD1100]의 일환으로 수행되었습니다.
◆ First Author : Electronics and Telecommunications Research Institute, akmoon@etri.re.kr, 정회원
° Corresponding Author : Andong National University Department of Creativity Software, Electronics and Telecommunications Research Institute, etri2022_khs@etri.re.kr, 정회원
논문번호 : 202208-161-0-SE, Received July 13, 2022; Revised August 25, 2022; Accepted September 5, 2022

Ⅰ. 서 론

최근 전쟁, 기후 등의 외부 요인으로 인해 수입에 의존했던 자원에 대한 가격 급등이 발생하여 국민 생활에 불안정을 초래함으로써, 각 국가별 자원의 중요성이 크게 부각되고 있다. 우리나라의 통계자료를 확인해보면, 우리나라의 식량 자급률은 전체 식량을 대상으로 2020년 기준 45.8%로 제시되고 있고, 매년 감소되고 있다[1]. 또한 경상북도 과실별 생산량 조사에서는 사과, 포도, 감 등의 과실 10a당 생산량이 2014년 이후 매년 감소되는 추세로 나타난다[2]. 이는 다양한 이유가 있겠으나, 이중 기후 변화도 큰 원인이라 생각한다. 그래서 최근 많은 과학자들이 농업 분야에 많은 영향을 미치는 것으로 기상관련 현상을 예측하는 데 집중하고 있다[3,4].

농업에서 개화기는 병충해 및 과실 생성의 유무를 결정짓는 중요 요인이고, 서리는 농업에 영향을 미치는 다양한 기후 현상 중에 하나로 만약 개화기에 서리가 발생할 경우, 작물에 심각한 손실을 초래할 수 있다[5]. 그래서 기후 변화의 피해로부터 작물과 농장을 효과적으로 보호하기 위해 활용 가능한 사전 예측 정보는 농부에게 큰 도움이 될 수 있다. 특히 노지 농업이 대부분인 우리나라에서는 더욱 필요한 정보라 할 수 있다. 예를 들면, 기상청의 전역 기상 데이터를 활용하여, 서리 예측 모델을 제시하고, 보다 정확한 서리 예보를 통한 수동 또는 자동(디지털 포함) 조치가 가능함으로써, 농가의 피해를 최소화 시킬 수 있다[6,7].

서리 예측 모델의 근거 데이터로써, 사물 인터넷(IoT) 기반 솔루션을 사용한 데이터 수집·분석 과정은 예측 모델에서 가장 중요하다. 우리나라에서는 기상청, 지역 농업센터 및 기타 수요 기관에서 특정 지역별로 IoT기반 관측소를 통하여 다양한 (빅)데이터를 수집하여 활용하고 있다. 미국의 Sage-platform[8] 및 AgWetherNet[9]과 우리나라의 날씨누리(기상청)와 같은 예측 서비스는 요구자(농업인 포함)의 의사 결정에 중요 요인이 되므로, 디지털 농업에서는 더욱 주목을 받고 있다.

이러한 IoT기반 예측 서비스의 정확도는 과거로부터 수집된 (빅)데이터셋에 크게 영향을 받으므로, 이를 효율적으로 수집 및 활용하기 위해서는 몇 가지 고려사항이 있다. 첫째, IoT기반 관측소는 보다 다양한 환경 데이터를 수집해야 한다. 예를 들면 온도(대기, 지표, 토양), 습도(대기, 지표, 토양), 일사량 및 풍속 등의 미기상(microclimate)을 포함한 환경 요인을 활용하면, 국소적으로 발생되는 서리를 예측하는데 보다 유용할 수 있다. 둘째, 서리는 대게 가을(10월)부터 봄(4월)의 기간중 산발적이고, 특정 조건에서 발생되는 자연현상이므로, 발생 클래스(NonFrost:0, Frost:1)에 따른 데이터 불균형이 매우 심각하다. 따라서 기계학습이 과거 데이터의 학습을 기반으로 하고, 이를 활용한 예측 모델 개발에서는 데이터셋의 클래스 불균형은 반드시 해소되어야 하며, 해소되지 않을시 잘못된 결과로 도출될 확률이 더욱 높아진다. 셋째, 아직까지 실제 IoT기반 관측소에 서리 발생에 대한 자동 관측은 불가하며, 전문가(농업인 포함)의 눈을 통한 직접 관측만 가능하다. 따라서 해당 관측소의 기후 데이터셋에 대한 서리 발생 라벨링 작업에는 오류가 발생하기 쉽고, 상당한 노력이 필요하다. 최근 이를 보완하기 위하여, 영상을 활용한 서리 예측 모델 개발도 제시되고 있다[10].

본 논문에서는 2017년부터 2021년간 수집된 IoT기상 관측소(A ~H: 총 8개)의 데이터셋을 활용하여, 서리 예측 모델을 제안한다(Table 1). 또한 연구의 목적이 서리 발생에 대한 예측이므로, 주요 인자(Frost)는 서리가 발생한 경우(Class: 1)이고, 비주요 인자(NonFrost)는 서리가 발생하지 않은 경우(Class: 0)로 제시한다. 그래서 IR(Imbalance Ratio)은 $\frac{Frost}{NonFrost}$ 로 표기함으로써, 비율에 대한 가독성을 높이고자 한다. Table 1은 관측소별 관측일자 기준 유효 데이터수(Valid Data), 관측일자 기준 서리 발생 건수(Frost), 관측일자 기준 서리 미발생 건수(NonFrost), 서리 발생 여부에 따른 불균형 비율(Imbalance Ratio)로 구성된다. 따라서 관측소별로 평균 765.5일의 관측자료를 보유한다.

서리는 발생 지역의 주변 기후 영향을 많이 받으므

표 1. 기상 관측소별 데이터 분포
Table 1. Data distribution of each weather station

Station	# of valid data	# of NonFrost	# of Frost	Imbalance ratio
A	**905**	**886**	**19**	**0.021**
B	**844**	**818**	**26**	**0.032**
C	664	654	10	0.015
D	843	827	16	0.019
E	784	770	14	0.018
F	753	741	12	0.016
G	486	476	10	0.021
H	**845**	**657**	**188**	**0.286**
mean	765.5	728.6	36.9	0.054

로, 예측에 다소 어려움이 있다. 따라서 가능한 다양한 지역에 대한 많은 정보를 활용하여 서리 예측에 적용하고자 한다. 그래서 8개의 관측소중 불균형 비율이 높고, 유용한 데이터 개수가 많은, 상위 3개(A, B, H)의 관측소를 선택하여 실제 실험에 적용한다(Table 1). 데이터의 불균형에 대한 해소법은 소수의 클래스 데이터에 합성 데이터를 생성하여 데이터 Augmentation을 위하여 오버샘플링의 기법중 하나인 SMOTE(Synthetic Minority Oversampling)[11]를 적용한다. 기계학습 분류 기법중 다양한 알고리즘을 통한 광범위한 평가를 수행한다: Decision Tree(DT)[12], Random Forest(RF)[13], Adaptive Boosting(AdaBoost)[14], Support Vector Machine (SVM)[15], Artificial Neural Network(ANN)[16]. k-겹 교차 검증 방법을 적용하여 알고리즘별 평가 결과를 비교하고, 최적의 알고리즘을 선정하다. 이후 최종 예측 모델은 GridSearchCV를 통해서 최적화를 진행하고, 성능 평가를 제시한다.

Ⅱ. 농업 기상 데이터셋 및 관련연구

AgWeatherNet은 농업 미기상을 수집하는 IoT 솔루션으로써, 워싱턴 주립대학에서 1988년부터 연구가 진행되고 있다. 이와 관련된 IoT 기상 관측소에서는 기온, 상대 습도, 이슬점 온도, 토양 온도, 강우량, 풍속, 풍향, 일사량, 잎 습기 등의 미기상 데이터를 수집한다.

국내 기상청에서는 2021년 5월 1일을 기준으로, 전국 95개의 종관기상관측장비(Automated Synoptic Observing System: ASOS)와 515개의 방재기상관측장비(Automatic Weather System: AWS) 및 농업기상관측(Automated Agricultural Observing System: AAOS)를 이용하여 자동으로 지상기상관측을 하고 있다. 종관기상관측은 지상의 기압, 기온, 습도, 바람, 강수, 일조, 시정 등의 자동화된 관측요소와 함께 유인기상관서의 적설, 구름, 기타 일기현상(계절관측 포함) 등 육안으로 관측한 목측요소를 포함한다. 방재기상관측은 지상의 기압, 기온, 습도, 바람, 강수 등을 자동으로 관측하고 있다. 농업기상관측은 농업기상에 필수적인 기온, 습도, 바람, 수분수지인자(증발량, 지하수위 등), 일조, 복사, 토양온도(지면 및 지중온도), 토양수분 등을 관측하고 있다[17]. 또한 서리 관측 데이터는 1911년부터 현재까지 년도별 시작서리와 끝서리 발생일자를 행정구역별로 제공하고 있다.

Ilseok Noh 외[10]은 한국 기상청 자동기상관측소와 경기도 농업기술원의 기후 데이터를 통한 특정 과수원 지역에 적합한 서리 예측 분류 모델을 제안한다. 클래스별 불균형은 다운샘플링을 적용하고, 총 8개의 기후 요소(기온, 영하 온도 지속시간, 강수량, 풍속, 습도, 강설량, 3h동안의 강설량, 지상온도)를 활용하여, 복잡(8요소) 및 단순(5요소) 모델로 구분하여 단계별 예측 모델의 생성 및 평가를 시행한다.

Ana Laura Diedrichs[18]은 Argentina의 Mendoza 지역의 5개 기상관측소 데이터를 적용하여, 인근 서리 예측 모델을 제시한다. 클래스별 불균형은 SMOTE를 적용하고, 온도와 습도 데이터를 활용한다. 가까운 위치의 측정 데이터를 포함하여 서리 예측 성능을 향상시킬 수 있음을 제시한다.

따라서 본 연구에서는 서리 예측 분석을 수행하기 위해, 작물 주변의 환경 요인을 주기적으로 수집하는 IoT기반의 농업기상 예측 플랫폼을 제안한다(Fig. 1). 해당 플랫폼은 특정 시간 및 장소의 기상 조건에 따른 서리의 발생 가능성을 예측하고, 결과를 농업인에게 제공한다. 따라서 예보에 따라, 농업인은 서리 발생에 대한 사전 대비가 가능하므로, 농가의 피해를 줄일 수 있다. 또한 제안된 플랫폼은 서리 예측에 주안점을 두고 있지만, 이를 활용하여 병해충 예보 및 토양수분 상태 정보를 제공함으로써, 관수, 방제 등의 다양한 서비스에도 적용이 가능하다.

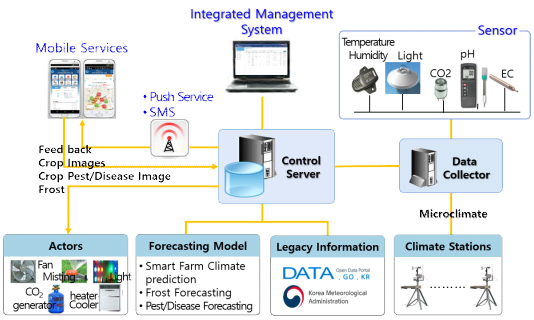

그림 1. IoT 기반 예측 플랫폼의 설계도
Fig. 1. The architecture of our IoT-based predictive platform

2015년부터 영천 지역의 특정 과수원에 IoT 기상관측소를 설치하고, 현재까지 농업 미기상 환경 데이터를 수집하고 있다. Table 2는 2021년 6월까지 수집된 데이터셋의 현황을 보여준다. 데이터셋은 우리가 직접 설치·수집한 미기상 데이터셋으로써, 농업 미기상은 작물·가축 주변의 미세한 기상을 의미하며, 기상청(KMA)에서도 현재 수집하고 있다. Microclimate Category는 관측소에서 분당 수집된 데이터의 총 건수이며, 8개 관측소를 기준으로 21,664,154건이 수집

표 2. 수집된 데이터셋 현황
Table 2. The collected dataset status

Category	Collected data type	# collected data per min
Micro-climate	temperature, grass temperature, rainfall, wind speed, win direction, solar radiation, soil temperature, dew, soil temperature, etc.	21,664,154 × Factors
Frost	# daily Frost Prediction	6,124
	# daily Frost Observation	295

되었다. 따라서 총 건수에 기상 요인(Factors)을 곱한 값은 수집된 데이터셋의 전체 크기를 나타낸다. Frost Category는 서리 실제 예측에 사용된 데이터셋을 의미하며, 총 6,124건 중에 295건의 서리가 관측됐다. 또한 Frost 데이터셋은, 모델의 보다 정확한 예측을 위하여, 2017년중 최초 서리 관측 이전의 데이터와 무상(無霜)기간(영천지역 5~9월)의 데이터를 제외한다[19]. 우리나라는 지역적 특성이 다양(산간지역, 해안가, 평야 등)하고, 4계절 기후가 다르게 나타난다. 이 특징은, 국소적·산발적으로 발생되는 서리의, 보다 정확한 예측에서 매우 중요하다. 따라서 본 연구에서는 행정구역 아래 세분화된 지역별 데이터셋을 적용한다. 그리고 계절별 특성이 반영된 시작서리부터 끝서리까지의 구간내 관측 데이터셋을 적용한다. 이는 기존의 공공데이터셋에서 다룰 수 없는 본 연구만의 중요한 특징이다.

Ⅲ. Feature 설계 및 방법론

Ⅱ장에서는 IoT기반의 농업기상 예측 플랫폼을 제시한다(Fig. 1). 그래서 Fig. 2는 이를 활용한 예측 모델 생성에 대한 모델 전반적인 방법론적 프로세스를 나타낸다. 따라서 Ⅲ 장에서는 예측 모델의 선택 및 최적화 단계별 설계 및 관련 이론에 대한 간략한 설명을 제시하고자 한다.

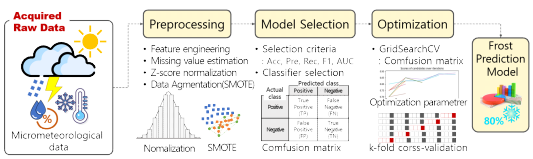

그림 2. 서리 예측을 위한 모델링 및 서비스 프로세스
Fig. 2. Modeling and service process for frost prediction

3.1 Feature 설계

서리는 대게 가을(10월)부터 봄(4월)의 기간중 국소적이면서, 특정 조건에서 발생되는 자연현상이다. 일년중 개화기에 발생하는 봄철 늦서리는 냉해로 인한 과수 및 작물의 꽃을 손상시킴으로써, 그해 심각한 경작 실패를 초래할 수 있다. 서리는 맑고 바람이 없는 날의 야간에 복사냉각이 진행되고, 접지층의 대기가 안정되면서, 기온이 영하로 내려갈 경우, 온도 역전으로 수증기가 승화하여 발생한다[20]. 그래서 서리는 일반적으로 이슬점 이하의 차가운 대기 온도에 영향이 있고, 서리가 내린 날과 내리지 않은 날의 차이가 이러한 특정 패턴을 만든다[18,21-25]. 따라서 온도, 습도, 풍속, 토양 등의 요인은 서리와 매우 관련이 높은 기상 요인이 된다. 온도 역전은 표면의 차가운 공기층이 따뜻한 공기층으로 덮인 것을 의미하는 것으로써, 공기 순환이 어렵기 때문에, 서리가 발생되기 쉽다. 본 논문에서는 지상 10cm에서 수집된 초상 온도와 지상 1.5m에서 수집된 대기 온도의 센서 데이터를 사용하여 온도 역전 데이터를 생성한다. I로 표시되는 온도 역전은 다음과 같이 계산한다(수식 1, 2).

$$\triangle T_t = T_{grass}(t) - T_{air}(t), \quad T_s \le t \le T_e \tag{1}$$

$$I = \sum \triangle T_t \le n(T_{air}), \ \text{if} \triangle T_t < 0 \tag{2}$$

위에서 특정 시간 t에서 초상 온도(T_{grass})와 대기 기온(T_{air})의 차이는 $\triangle T_t$이다. 이러한 온도 변화 측정은 발생 전날 12:00부터 시작하여 23:00(발생 전날의 예측 시간)까지 계속된다.

$$X_{(t)}^{(i)} = \begin{Bmatrix} X_{(1)}^{(1)}, X_{(2)}^{(1)} ..., X_{(n)}^{(1)} \\ X_{(1)}^{(2)}, X_{(2)}^{(2)} ..., X_{(n)}^{(2)} \\ \cdots \\ X_{(1)}^{(p)}, X_{(2)}^{(p)} ..., X_{(n)}^{(p)} \end{Bmatrix} \tag{3}$$

$$\Rightarrow A_{(j)}^{(i)} = \left\{ A_{(1)}^{(i)}, A_{(2)}^{(i)} ..., A_{(fn)}^{(i)}, \theta^{(i)} \right\}$$

이후 수집·계산된 데이터셋은 예측을 위한 구간별 대푯값으로 변환한다(수식 3). 데이터 포인트 $X_{(t)}^{(i)} \in R$이고, 여기서 t는 $1 \le t \le n$이고, n은 센서 데이터에서 측정하는 미기상 요소의 수를 나타낸다. $1 \le i \le p$에서 i는 서리를 예측한 특정구간(본 논문은 1일)이고, $1 \le j \le fn$에서 j는 대푯값 요소의 수를 의미한다. 또한 $A_{(j)}^{(i)}$는 i기간의 데이터에 의해 재계산된 j번째 매개변수이고, $\theta^{(i)}$는 특정 i구간의 서리 발

표 3. 예측을 위한 환경 입력 데이터 세트
Table 3. The environmental input datasets for prediction

Climate factor	Feature	Calculation model	
Temperature Inversion	$I(\triangle T_i)$	Equation 1 and 2	
Dewpoint	T_{dp}	Magnus-Tetens approximation	
Temperature Dew	T_{dew}	between 22:30 and 23:00	
Rainfall	$rain$	between 16:00 and 16:30 between 22:30 and 23:00	
	$solar$	between 12:00 and 19:00	
Air Temperature	T_{air}	-	between 12:00 and 23:00
	$T_{air}(\min)$	minimum	
	$T_{air}(\max)$	maximum	
	$T_{air}(diff)$	$T_{air}(\max)-T_{air}(\min)$	
Wind Speed	W_{speed}	between 21:00 and 23:00	
Grass Temperature	T_{grass}	between 12:00 and 23:00	
	$T_{grass}(\min)$	minimum of T_{grass}	
Soil Temperature	T_{soil}	between 22:30 and 23:00	

생을 특성화하는 레이블(NonFrost:0, Frost:1)이다. 본 논문의 서리 예측 모델은 13개의 특징(Table 3)으로 구성되고, 온도 역전 계수를 포함하여 제안한다. 또한 클래스 불균형을 해결하기 위하여, SMOTE 방식을 통해 $\{A_{(1)}^{(i)}, A_{(2)}^{(i)} ..., A_{(fn)}^{(i)}, \theta^{(i)}\}$의 개수를 증가한다.

3.2 기계학습 분류기

최적의 서리 예측 모델을 위해서는 적절한 학습 알고리즘 선택이 매우 중요하다. 본 논문에서는 DT, RF, AdaBoost, SVM 및 ANN의 5가지 대표적인 분류 알고리즘을 적용하고, 검증 및 비교를 통해 최종 서리 예측 알고리즘을 선택한다.

DT는 다양한 실제 문제를 해결하기 위해 널리 사용되는 지도학습 방법이다. Gini index(수식 4)와 Entropy(수식 5)를 이용하여 주어진 데이터셋의 특정 단말 조건까지 반복적으로 분류한다.

$$Gini = 1 - \sum_{i=1}^{k} p_i^2 \tag{4}$$

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i) \tag{5}$$

여기서 p_i는 주어진 데이터의 요소 수에 대한 i번째 노드의 클래스 k에 속할 비율을 나타낸다. gini는 순도·불순도의 구분이고, entropy는 정보이득의 구분이다. DT는 Entropy와 Gini 값이 낮은 방향으로 노드가 분할되지만, 지나친 노드 분할은 과적합의 경향성이 있고, 훈련 데이터가 약간만 변동시에도 불안정해질 수 있다.

RF는 여러 DT를 결합하여 위에서 언급한 DT 취약점을 해결하고자 하는 의사결정 트리의 앙상블 모델이다. 각 하위 트리는 훈련 데이터의 다른 부트스트랩 샘플을 사용하여 학습된다. 그런 다음 평균, 복수 및 가중 투표를 사용하여 하위 트리의 결과를 집계함으로써, 분류 또는 회귀 문제를 해결한다(수식6).

$$H(x) = \arg\max_{i} \sum_{t=1}^{T} I(h_t(x) = i) \tag{6}$$

여기서 $h_t(x)$는 기본 분류기의 t번째 출력을 나타낸다. RF는 편향을 약간 증가시킬 수 있지만, 작동 원리에 따라 다양한 하위 트리를 결합하여 분산을 줄인다. 그래서 다른 분류기에 비하여 대부분 성능이 좋게 나타난다.

AdaBoost는 DT기반 기계학습 알고리즘의 또다른 개선 사례이다. 기본 가정은 각 분류기의 성능이 약한 경우에도 여러 분류기를 그룹화하면 성능이 향상될 수 있다는 것이다. 하위 트리를 병렬로 훈련하는 RF와 달리, AdaBoost는 잘못 분류된 샘플에 대해 가중치를 조정하여 기본 분류기를 점진적으로 훈련한다. 그 후, 분류 결과는 수식 7과 같이 집계된다.

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{7}$$

여기서 $h_t(x)$와 α_t는 입력 x에 대한 기본 분류기의 t번째 출력과 할당된 가중치를 나타낸다. 이러한 특성 때문에 AdaBoost는 일반화 오류가 적고, 구현하기 쉬우며, 하이퍼 매개변수를 조정할 필요가 적다. 반면에 AdaBoost는 잡음이 있는 데이터와 이상값에 민감하며, 기본 분류기가 너무 복잡할 경우 과적합이 발생된다.

DT, RF 및 AdaBoost와 달리 SVM은, 최적의 데이터 분리 평면을 찾기 위해 데이터를 고차원 또는 무한 차원 기능 공간에 매핑한다. 커널함수를 적용한 SVM은 높은 차원 공간을 쉽게 분리될 수 있기 때문에, 주어진 데이터가 선형으로 분리할 수 없을 때 효과적인 알고리즘이다. 어떤 형태의 함수라도 Mercer의 이론[26]을 만족하면 커널 함수로 사용할 수 있으며, 가장

자주 사용되는 커널 함수는 방사형 기저 함수(수식 8)이다.

$$K(x_1, x_2) = \exp - \frac{\| x_1 - x_2 \|_2^2}{2\sigma^2}, \ \sigma \neq 0 \quad (8)$$

SVM은 차원 수가 샘플 수보다 많을 때, 상대적으로 메모리 활용이 효율적이고 효과적이다. 그러나 수학적으로 복잡하고 계산 비용에 오버헤드가 큰 단점이 있다.

ANN은 입력 레이어, 하나 이상의 은닉 레이어 및 출력 레이어로 구성된 알고리즘이다. 학습 프로세스에는 역전파 및 경사하강법[27]을 사용하여 가중치를 조정함으로써, 오차를 최소화하여 성능을 향상시킨다. 인공신경망 기반 접근인 ANN은 데이터 규모가 클수록 잘 작동하는 것이 일반적이다. 따라서 데이터가 비선형적인 복잡한 분포를 가지면서, 데이터셋에 대한 정보가 충분하지 않을 경우에도 비교적 잘 작동하는 방법이다.

GridSearchCV는 선택된 분류기에 대하여, 예측 성능이 우수한 최적의 하이퍼 매개변수 값을 찾아주는 알고리즘으로써, CV인자 값만큼 계층을 분할하여 교차검증을 시행하고, 지정한 성능 결과값이 가장 우수한 하이퍼 매개변수(best_params)를 반환한다. 일반적으로 많이 활용되는 기계학습 모델 생성 과정은 조작자가 분류기 생성, 데이터셋 분리(train/test), 예측 모델 생성 및 성능 검증의 여러 단계를 진행한다. 하지만 GridSearchCV는 조작자가 한번의 조작으로 이 모든 과정을 수행하는 장점이 있다. 그런데 이런 최적의 하이퍼 매개변수를 찾기위해 모든 경우를 적용하게 되므로, 경우에 따라서는 시간이 매우 많이 소요되는 단점이 있다.

3.3 성능 평가

서리 발생 여부를 예측하기 위한, 서리 예측 알고리즘 및 모델의 성능 평가는 Confusion Matrix를 활용(TN, FP, FN, TP)하고, 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), 조화평균(F1-score) 및 AUC[28]의 총 5가지 성능 지표를 적용한다(수식9~13).

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \quad (9)$$

$$\mathrm{Precision} = \frac{TP}{FP + TP} \quad (10)$$

$$Recall = \frac{TP}{FN + TP} \quad (11)$$

$$F1\,score = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \quad (12)$$

$$AUC = P(X_n \leq X_d) = \int_0^1 ROC(p)dp = \int_0^1 1 - F_d(F_n^{-1}(1-p))dp \quad (13)$$
$$(p = 1 - F_n(x))$$

Ⅳ. 실험 및 검증

본 절에서는 농업기상 예측 플랫폼에서 제안한 서리 예측 모델을 개발하고, 성능 지표를 활용해 검증하고자 한다. 데이터셋은 영천 지역 과수원에 위치한 IoT 기상 관측소로부터 수집한 기상 데이터셋중 2017년부터 2021년까지의 데이터셋만 사용한다. 그리고 보다 정확한 예측을 위하여, 서리 발생 클래스별 불균형 비율이 가장 높은 3개(A,B,H)의 관측소 데이터만 적용한다(Table 1).

4.1 실험 데이터셋 및 입력 변수

영천 지역 과수원에 IoT 기상 관측소를 설치하였으며, 2015년부터 2021년까지 기상 데이터를 수집하였다. 그러나 서리 발생 라벨링 데이터가 2017년도부터 2021년까지 측정되었다. 그래서 최종 실험 데이터셋은 각 관측소별로 서리 발생 라벨링이 시작된 시점(2017년 특정 일자)부터 2021년 6월까지의 기간 데이터로 확정한다. 또한 서리가 발생하지 않는 무상기간(5월~9월)에 대한 기후 데이터는 제외한다.

입력 변수는 총 12개의 일별 기후 대푯값으로 구성한다: 온도(t_air), 최소 온도(t_min), 일일 최대 온도차(t_diff), 초상온도(gtp), 최소 초상온도(gtp_min), 역전층(inverse), 풍속(w_speed), 강수량(rain), 결로 온도(t_dew), 토양 온도(t_soil), 이슬점(dewpoint) 및 일사량(solar). 대푯값 생성방법은, 관측소별 기상 데이터셋에서 기상 요인(온도, 풍속, 습도, 결로, 일사량

표 4. 예측 모델 개발을 위한 입력 변수
Table 4. Input variables for developing predictive models

Division	Input variable	Count
Daily average (by minute)	t_air, gtp, t-soil, w_speed, rain, t-dew, solar, r_hmdt,	8
Specific value (by day)	t_min, t_diff, gtp_min, inverse, dewpoint	5
final feature (by day)	t_air, t_min, t_diff, gtp, gtp_min, inverse, w_speed, rain, t_dew, t_soil, dewpoint, solar	12

등)의 정보를 정규화 변환하고, 이를 활용하여 평균 및 특정값 변환의 방법을 적용해 별도의 매개 변수를 생성한다. 그리고 이 값을 입력 변수로 적용하여 서리 예측 모델을 개발한다(Table 4).

4.2 데이터셋 분포

정규화된 입력 변수에 대한 실험 데이터셋의 통계적인 속성을 사전에 확인한다. 평균(Mean), 표준편차(STD), 왜도(Skewness), 첨도(Kurtosis). 최소값(min), 최대값(max), 사분위수는 입력 변수들의 분포를 나타낸다. 왜도는 분포의 비대칭 정도의 척도로써, 가운데를 중심으로 0은 대칭인 정규 분포, 음수는 왼쪽으로 집중된 분포, 양수는 오른쪽으로 집중된 분포를 의미한다. 첨도는 분포의 중심에서 뾰족함의 척도로써, 정규분포(기준값: 0)를 기준으로, 값이 크면 더 뾰족하고 이상치 발생 확률이 높다.

Table 5는 실험 데이터셋에 대한 통계적 속성을 확인한 결과이다. 평균은 대부분 0 주위로 나타난다. 표준편차는 t_diff, t_dew을 제외하면 1내외로 나타난다. 일부 변수(t_min, t_diff, gtp_min, w_speed, rain, t_dew)에서, 왜도는 좌우 한쪽으로 치우친 분포가 나타나고, 첨도는 이상치 발생 확률이 높을 것으로 예상된다.

표 5. 입력 변수의 통계적 속성
Table 5. Statistical properties of input variables

Input	mean	std	skew	kurto	min	25%	50%	75%	max
solar	0.010	0.995	0.409	0.025	-2.237	-0.642	-0.102	0.590	3.681
t_air	0.030	0.990	-0.090	-0.541	-2.968	-0.739	0.036	0.790	2.855
t_min	0.029	0.908	**-0.753**	**9.083**	-10.778	-0.603	0.049	0.631	2.800
t_diff	0.012	**0.870**	**1.692**	**30.193**	-2.569	-0.508	0.016	0.554	18.214
gtp	0.029	0.998	-0.014	-0.731	-2.900	-0.796	0.045	0.816	2.665
gtp_min	0.020	0.907	**-0.698**	**8.664**	-10.081	-0.604	0.017	0.618	2.895
w_speed	-0.011	1.021	**1.365**	**1.304**	-1.287	-0.737	-0.422	0.497	5.144
rain	-0.008	0.974	**16.373**	**352.279**	-0.155	-0.143	-0.132	-0.126	28.284
t_dew	-0.071	**0.706**	**2.950**	**9.920**	-0.672	-0.526	-0.176	-0.133	4.384
t_soil	0.052	1.035	-0.057	0.336	-3.125	-0.698	0.152	0.771	8.262
dewpoint	-0.009	0.977	-0.147	-0.366	-3.219	-0.709	0.068	0.675	2.800

4.3 합성 데이터 생성(SMOTE)

SMOTE는 클래스별 균형을 맞추기 위한 데이터 Augmentation(Over-Sampling) 기법중의 하나로써, 일정 구간내 소수 클래스 데이터를, 특정 비율만큼 임의 생성하는 방법이다. 데이터 Augmentation은 Under-Sampling에 비하여 기존 데이터의 유실을 막으면서 클래스별 불균형을 해소 가능하므로, 예측 정확도를 높일 수 있는 유용한 방법으로 많이 활용되는 기법이다. Fig. 3은 Station A에 대한 SMOTE 전후의 기계학습 분류기를 적용한 성능 결과이다(좌측부터 DT, RF, AdaBoost, SVM, KNN, ANN의 순으로 F1-score와 AUC 결과 제시). SMOTE 전의 평균값은 0.29이며, 후의 평균값은 0.85로써, SMOTE를 통한 Augmentation이 불균형 데이터의 성능 향상에 영향이 있음을 나타내고 있다.

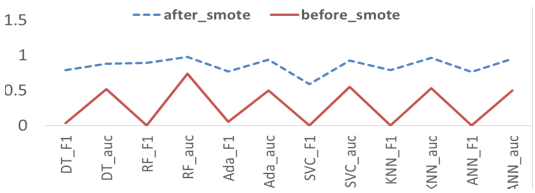

그림 3. smote(station: A) 전후 예측 성능
Fig. 3. Predictive performance before and after smote(station:A)

Table 5에 따르면, 입력 변수는 정규분포와 유사하게 평균을 중심으로 몰려있는 분포로 유추가 되고, 일부 변수는 이상치 발생의 확률이 높게 나타나므로, 입력 변수의 값에 다소 민감하게 작용할 수 있음을 예측할 수 있다. 따라서 Table 1에서 IR>0.2인 데이터는 IR<0.1 데이터보다 넓게 분포되어 있고, 적정 Augmentation 비율을 적용하지 않을시, 서리 미발생 데이터 수를 초과하는 경우가 발생될 수 있다. 또한 기존 데이터수에 비하여 지나친 Augmentation 비율은 오히려 잘못된 입력값을 제공하는 경우가 될 수 있으므로, 적정 Augmentation 비율 산정은 중요하다.

따라서 본 연구에서는 서리 발생 여부에 대한 불균형 데이터의 Frost(class: 1) 비율을 최소 20% 정도로 자체 설정하고, 해당 비율 산정시 활용한다. 따라서 본 연구에서는 관측소별 서리 발생 데이터에 대한, 불균형 비율의 차이를 감안하여, 차등 비율(A와 B는 10배, H는 1.5배)의 SMOTE 기법을 적용한다(Table 6). SMOTE 적용 전후의 데이터 분포에 대한 변화 예시로써, 관측소 A에 대한 t-min(x축)과 dewpoint(y축)간 그래프를 제시한다(Fig. 4).

또한 데이터 Augmentation 비율은 모델 예측시 과적합 예방에 매우 중요하다. 따라서 최적 예측 모델 생성 후, 소수 데이터 생성 비율을 달리하여, 성능 지표를 재점검 하고자 한다.

표 6. SMOTE 전후 비교
Table 6. Comparison before and after SMOTE

Station		A		B		H	
Occurrence		Non Frost	Frost	Non Frost	Frost	Non Frost	Frost
SMOTE	Before	886	19	818	26	657	188
	After	886	190	818	260	657	282
SMOTE rate		Frost×10		Frost×10		Frost×1.5	

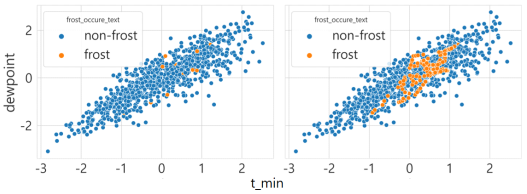

그림 4. SMOTE 전(좌)과 후(우) 산점도 비교(Station_id: A, x: t-min, y: dewpoint)
Fig. 4. Comparison of scatter plots before(left) and after(right) SMOTE(Station_id: A, x: t-min, y: dewpoint)

4.4 최적의 알고리즘 선택 및 모델 최적화

본 절에서는 SMOTE를 적용한 관측소별 데이터셋을 활용하여, 기계학습의 다양한 알고리즘에 적용하고, 가장 적합한 알고리즘을 선택하고자 한다. 본 연구는 서리 발생의 유(Class:1)와 무(Class:0)를 예측하기 위함이므로, 기계학습의 분류 기법에 해당하는 알고리즘을 다양하게 적용하고, k-겹 교차검증(cv=5)을 통해 각 분류기의 성능을 평가한다: DT, RF, AdaBoost, SVM 및 ANN. 성능 평가 지표는 앞서 제시한 정확도(Acc), 정밀도(Pre), 재현률(Rec), 조화평균(F1) 및 AUC의 총 5가지를 기준으로 적용한다. 이후 최종 선택된 알고리즘을 활용하여, GridSearchCV를 적용함으로써, 최적의 예측 모델을 제시한다.

Table 7은 기계학습 알고리즘별 k-겹 교차검증을 통한 성능 결과를 보여준다(Fig 5). RF는 5개의 모든 지표에서 다른 알고리즘보다 우수한 성능을 보여준다. DT와 SVM은 다른 알고리즘에 비해 값이 낮게 나타난다. 그리고 SMOTE 생성 비율이 높은 A와 B 관측소는 성능이 높게 나타나고 있지만, 생성 비율이 낮은 H 관측소의 성능은 다소 낮게 나타나고 있다. 따라서 앞서 예측한 대로 SMOTE 생성 비율은 모델의 성능에 매우 깊은 영향을 줄 수 있는 또 하나의 매개 변수가 될 수 있음을 예측할 수 있다.

이로써 5개의 기계학습 알고리즘중 RF가 서리 예측에 가장 적합한 알고리즘이라는 결론을 내릴 수 있다. 그래서 최종 선택된 알고리즘(RF)은 예측율을 높

표 7. 기계 학습 분류자별 성능 측정항목(cross-val 점수 평균)
Table 7. Performance metrics by machine learning classifier(mean of cross-val_scored)

ID	Classifier	Acc	Pre	Rec	f1	AUC	Average
A	DT	0.924	0.805	0.777	0.788	0.877	0.83
	RF	**0.965**	**0.842**	**0.954**	**0.893**	**0.979**	**0.93**
	AdaBoost	0.919	0.774	0.778	0.772	0.938	0.84
	SVM	0.878	0.5	0.73	0.592	0.924	0.72
	ANN	0.917	0.763	0.77	0.766	0.953	0.83
B	DT	0.9	0.796	0.791	0.79	0.864	0.83
	RF	**0.943**	**0.846**	**0.913**	**0.877**	**0.98**	**0.91**
	AdaBoost	0.882	0.742	0.763	0.752	0.92	0.81
	SVM	0.87	0.627	0.788	0.696	0.935	0.78
	ANN	0.9	0.773	0.803	0.787	0.95	0.84
H	DT	0.675	0.514	0.463	0.486	0.629	0.55
	RF	**0.783**	**0.5**	**0.697**	**0.577**	**0.787**	**0.67**
	AdaBoost	0.731	0.44	0.566	0.493	0.73	0.59
	SVM	0.733	0.287	0.631	0.393	0.738	0.56
	ANN	0.755	0.433	0.639	0.513	0.759	0.62

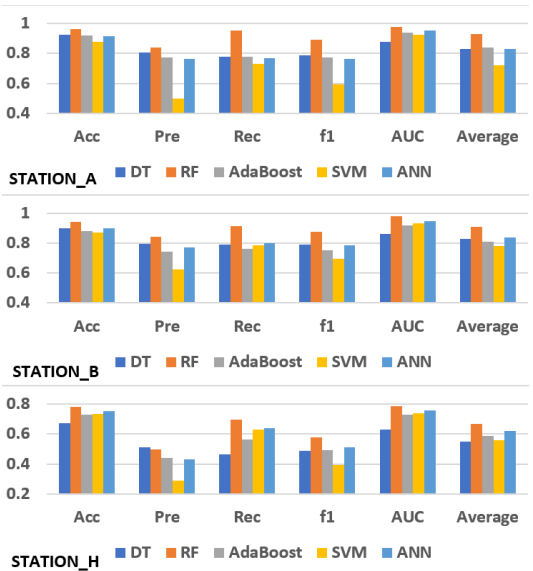

그림 5. 기계 학습 분류기별 성능 측정항목
Fig. 5. Performance metrics by machine learning classifier

이기 위한 훈련 및 검증을 통한 최적화 과정이 반드시 필요하다. 본 연구에서는 GridSearchCV(cv=5) 기법을 적용하고, 검증 비교지표인 조화평균(F1)을 기준으로 적용하여, 관측소별 예측 모델에 대한 최적의 하이퍼 파라미터를 찾는다(Table 8). 이때 하이퍼 파라미터는

표 8. 스테이션별 최적화된 RF 하이퍼파라미터
Table 8. Optimized RF hyper_parameters for each station

Station	Criterion	Maximum Depth	#of subtrees
A	Entropy	17	112
B	Gini	27	73
H	Gini	14	104
common	Gini	24	101

서브트리의 수(65~120), 각 서브트리의 최대 깊이(10~30), 분할 기준(Gini 및 엔트로피)[29]을 적용한다.

GridSearchCV는 적용시마다, 출력되는 하이퍼 파라미터에 다소 변동이 있다. 그러나 제시한 검증 지표값이 가장 높은 하이퍼 파라미터를 제시하므로, 모델의 성능에는 문제가 되지 않는다.

4.5 최적화 모델 검증

최적화 하이퍼 파라미터를 적용한, 관측소별 최적의 서리 발생 예측 모델에 대하여 검증 결과는 다음과 같다(Table 9, Fig. 6).

최적화 모델의 성능 지표는 기계학습 알고리즘 성능(Table 7)과 유사한 양상이 나타난다. SMOTE 생성 비율이 낮은 H 모델이 A와 B 모델에 비해 상대적으로 성능이 낮게 나타나고 있다. 그러나 최적화 과정을 통해 이전 성능에 비해서는 다소 조정이 된 것으로 나타난다. A 모델은 Pre가 높아지면서, F1값도 같이 높게 나타난다. B와 H 모델은 Pre가 높아졌지만, Rec이 낮아지면서 F1은 약간 높게 나타난다. A, B 및 H 관측소 데이터셋을 모두 합하여 생성한 common 모델은 최적화 모델에서만 검증하였고, 모든 성능 지표값에서 안정적으로 나타난다. 따라서 실제 서리 발생 예측 서비스에서는, 생성된 관측소별 모델이 다소 불안정할 경우, 해당 관측소 모델과 common 모델을 비교하여, 보다 성능이 높은 모델로 예보하는 방법도 제시해 본다.

이제는 4.3단원에서 제시한 데이터 Augmentation 비율에 따른 최적화 모델의 성능 결과를 비교하고자 한다(Fig 7, Table 10). 대상 관측소와 데이터셋은 최적화 과정과 동일하게 적용한다.

SMOTE 비율은 $Frost \times 1.1$부터 $\frac{Frost}{NonFrost} = 1$ 까지의 구간으로 적용한다. 총 4개 모델의 클래스간 비율 차이로 인해, 생성 데이터 수의 차이가 크므로, SMOTE 비율 구간은 0.4간격으로 생성한다. A, B, H 및 common 모델의 SMOTE 구간별 데이터 크기는 114, 76, 6 및 23으로 다양하게 나타난다. Table 10은 SMOTE 비율별 검증 결과의 예시로써, common 모

표 9. 최적화 모델에 대한 스테이션별 검증 결과
Table 9. Station-specific validation result for optimization models

Station	Acc	Pre	Rec	f1	AUC	Average
A	0.981	0.964	0.93	0.946	0.961	0.956
B	0.951	0.942	0.844	0.89	0.914	0.908
H	0.787	0.657	0.561	0.605	0.72	0.666
common	0.904	0.831	0.734	0.779	0.844	0.818

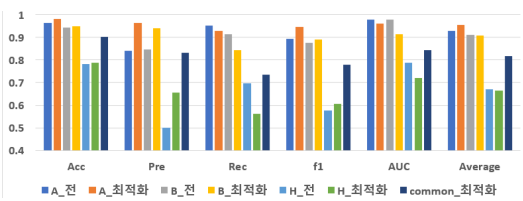

그림 6. 최적화 모델 검증 그래프
Fig. 6. Optimization model validation graph

표 10. SMOTE 비율에 따른 서리 예측 모델(comm) 검증 결과
Table 10. Frost prediction model(comm) verification result by SMOTE ratio

Rate	Acc	Pre	Rec	F1	AUC	IR
1.1	0.97	0.867	0.816	0.841	0.901	0.108
3.1	0.93	0.944	0.747	0.834	0.866	0.306
5.1	0.895	0.964	0.714	0.82	0.85	0.503
7.1	0.874	0.974	0.713	0.823	0.85	0.701
9.1	0.846	0.979	0.689	0.808	0.838	0.898
9.9	0.842	0.98	0.695	0.813	0.841	0.977

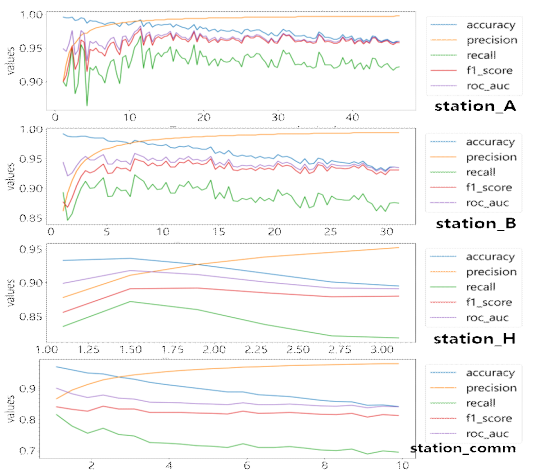

그림 7. 모델 사이즈별 검증 지표 현황: A(114), B(76), H(6), comm(23)
Fig. 7. Status of verification indicators by model size: A(114), B(76), H(6), comm(23)

델의 검증 결과에 대해 2단위 간격의 rate를 발췌한 결과이다. 또한 SMOTE 비율별 검증 결과의 전체적인 흐름을 확인하고자, 그래프를 제시한다(Fig. 7). 4가지 모델은 모두 각 지표별 유한한 양상으로 나타난다. 정확도, 조화평균, AUC는 점차 감소하면서 일정 값으로 수렴되는 경향을 나타내고, 재현률은 좀더 많이 감소하는 경향을 나타낸다. 정밀도는 상승하다가 일정값으로 수렴되는 경향을 나타낸다. 또한 당초 클래스간 격차가 많은 모델은 그렇지 않은 모델에 비해, 정확도·재현률·F1·AUC의 감소 기울기가 작게 나타난다. 따라서 이런 특징을 활용하면, 보다 유용한 데이터셋 생성이 가능하다고 사료된다.

V. 결과 및 향후 계획

서리는 가을부터 봄까지, 특정 기후조건에서 국소적 지역에 발생되는 자연환경이다. 따라서 작물의 개화기인 늦봄에 발생되는 서리는 그 해 생산량과도 직결되므로, 이를 예방하기 위한 사전 예보 서비스는 반드시 필요하다. 그래서 본 연구에서는 IoT기반의 농업기상 예측 플랫폼을 제안한다(Fig. 1).

데이터셋은 2017년부터 2021년간 영천 지역의 과수원에 설치된 IoT 관측소와 전문가(농업인)의 목측을 통하여 수집된 기상 데이터(미기상 포함)를 적용한다. 또한 서리 발생에 대한 클래스별 불균형이 발생되므로, SMOTE 기법을 적용하여 부족한 클래스의 데이터를 보완한다. 따라서 본 논문에서는 2017년도 이후의 데이터셋(무상기간 제외)중 클래스 불균형 비율이 높고, 유요한 데이터 개수가 많은 상위 데이터셋만 활용하여 서리 예측 모델을 생성한다. 기계학습 분류 기법중 5개의 알고리즘의 성능을 평가하여 최적의 알고리즘(RF)을 선택하고, 해당 알고리즘에 최적화 훈련 및 검증과정을 적용하여 최종 서리 예측 모델을 생성한다. 최종 서리 예측 모델의 성능은 F1값이 0.61에서 0.95까지로 나타났으며, SMOTE 생성 비율이 높은 모델의 예측률이 더 높게 나타난다. 이후 시행한 SMOTE 생성 비율에 따른 모델 성능 지표 변화는 정밀도는 상승 수렴, 나머지는 감소 수렴의 추세를 나타내고 있다.

본 논문에서는 데이터셋의 클래스간 비율 차이로 인해, 합성 데이터의 생성 비율에서도 차이가 발생하기 때문에 관측소간 모델 비교는 적합하지 않다고 생각되어 시행하지 않는다. 그러나 향후 연구에서는 SMOTE 기법의 적정 비율을 활용하여, 관측소 데이터셋의 동질성 검증 및 데이터 보완을 통한 관측소별 서리 예측 모델간의 성능 비교를 진행해보고자 한다. 그리고 시계열 데이터셋인 기후 데이터를 딥러닝 알고리즘에 적용하여, 관측소별 최적 모델 생성을 통한 서리 발생 예측 모델을 개발하고자 한다.

References

[1] Ministry of Agriculture, "*Food and Rural Affairs _ Grain self-sufficiency rate*(2022)," Retrieved Jul. 10, 2022, from https://www.data.go.kr/data/15101563/fileData.do

[2] Gyeongsangbuk-do, "*Production status by fruit*(2021)," Retrieved Jul. 10, 2022, from https://www.data.go.kr/data/15004871/fileData.do

[3] J. R. Rozante, E. R. Gutierrez, P. L. da Silva Dias, A. de Almeida Fernandes, D. S. Alvim, and V. M. Silva, "Development of an index for frost prediction: Technique and validation," *Meteorological Applications*, vol. 27, Issue. 1, Jan. 2020.

[4] A. Moon, J. Kim, J. Zhang, and S. W. Son, "Evaluating fidelity of lossy compression on spatiotemporal data from an iot enabled smart farm," *Comput. and Electr. in Agric.*, vol. 154, pp. 304-313, Nov. 2018. (http://dx.doi.org/10.1016/j.compag.2018.08.045)

[5] P. Matzneller, K.-P. Gotz, and F.-M. Chmielewski, "Spring frost vulnerability of sweet cherries under controlled conditions," *Int. J. Biometeorology*, vol. 60, no. 1, pp. 123-130, Jan. 2016. (https://doi.org/10.1007/s00484-015-1010-1)

[6] U. Chung, H. C. Seo, and J. I. Yun, "Site-specific frost warning based on topoclimatic estimation of daily minimum temperature," *Korean J. Agric. and Forest Meteorology*, vol. 6, no. 3, pp. 164-169, 2004.

[7] R. Snyder and J. de Melo-Abreu, "Frost protection: Fundamentals, practice, and economics - Volume 1," Rome: *Food and Agriculture Organization of the United Nations*, pp. 1-240, 2005.

[8] Sage, Retrieved July 10, 2022, from

https://sagecontinuum.org/.

[9] *AgWeatherNet Current Conditions Map*, Retrieved July 10, 2022, from https://weather.wsu.edu/.

[10] I. Noh, H.-W. Doh, S.-O. Kim, S.-H Kim, S. Shin, and S.-J. Lee, "Machine learning-based hourly frost-prediction system optimized for orchards using automatic weather station and digital camera image data," *MDPI Atmosphere 2021*, vol. 12, no. 7, 846, 2021. (https://doi.org/10.3390/atmos12070846)

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *J. Artificial Intell. Res.*, vol. 16, pp. 321-357, 2002. (https://doi.org/10.48550/arXiv.1106.1813)

[12] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers-a survey," *IEEE Trans. Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pp. 476-487, Oct. 2005. (https://doi.org/10.1109/TSMCC.2004.843247)

[13] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, Oct. 2001.

[14] R. E. Schapire and Y. Freund, "Boosting: Foundations and algorithms," *Kybernetes*, vol. 42, no. 1, pp. 164-166, 2013. (https://doi.org/10.1108/03684921311295547)

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.

[16] M. H. Hassoun, et al., *Fundamentals of artificial neural networks*, MIT press, 1995.

[17] Korea Meteorological Administration, *Climate Statistical Guidelines* (2021.5), pp. 1-2, Retrieved Jul. 10, 2022, from https://book.kma.go.kr/viewer/MediaViewer.ax?cid=37367&rid=5&moi=5513

[18] A. L. Diedrichs, F. Bromberg, D. Dujovne, K. Brun-Laguna, and T. Watteyne, "Prediction of frost events using machine learning and IoT sensing devices," *IEEE Internet of Things J.*, vol. 5, no. 6, pp. 4589-4597, 2018. (https://doi.org/10.1109/JIOT.2018.2867333)

[19] M. K. Shim, *Frost occurrence date and distribution status of free period by region, Agricultural technology information Post*, p. 6, Retrieved Jul. 10, 2022, from http://reurl.kr/39014A4ABUU

[20] Y. A. Kwon, H. S. Lee, W. T. Kwon, and K. O. Boo, "The weather characteristics of frost occurrence days for protecting crops against frost damage," *J. Korean Geographical Soc.*, vol. 43, no. 6, pp. 824-842, 2008.

[21] P. F. Verdes, P. M. Granitto, H. D. Navone, and H. A. Ceccatto, "Frost prediction with machine learning techniques," in *VI Congreso Argentino de Ciencias de la Computacion*, 2000.

[22] P. Sallis, M. Jarur, and M. Trujillo, "Frost prediction characteristics and classification using computational neural networks," in *Advances in Neuro-Inf. Process. ICONIP 2008*, pp. 1211-1220, Auckland, New Zealand, Nov. 2008. (https://doi.org/10.1007/978-3-642-02490-0_147)

[23] Y. Tamura, L. Ding, K. Noborio, and K. Shibuya, "Frost prediction for vineyard using machine learning," in *2020 Joint 11th Int. Conf. SCIS-ISIS, IEEE*, pp. 1-4, Hachijo Island, Japan, Dec. 2020. (https://doi.org/10.1109/SCISISIS50064.2020.9322770)

[24] I. Zhou, J. Lipman, M. Abolhasan, N. Shariati, and D. W. Lamb, "Frost monitoring cyber–physical system: A survey on prediction and active protection methods," *IEEE Internet of Things J.*, vol. 7, no. 7, pp. 6514-6527, Jul. 2020. (https://doi.org/10.1109/JIOT.2020.2972936)

[25] E. S. Diniz, A. S. Lorenzon, N. L. M. de Castro, G. E. Marcatti, O. P. Santos, J. C. de Deus Junior, R. B. L. Cavalcante, E. I. Fernandes-Filho, and C. H. Amaral, "Forecasting frost risk in forest plantations by the combination of spatial data and machine learning algorithms," *Agricultural and Forest Meteorology*, vol. 306, Aug. 2021. (https://doi.org/10.1016/j.agrformet.2021.108450)

[26] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization,* Cambridge university press, 2004.
[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning,* MIT press, Nov. 2016.
[28] D. H. Jang, *Partial AUC using the sensitivity and specificity lines*, Sungkyunkwan University, 2021.
[29] L. Rokach and O. Z. Maimon, *Data mining with decision trees: theory and applications*(vol. 69), World Scientific Publishing Co. Pte. Ltd, 2007. (https://doi.org/10.1142/6604)

문 애 경 (Ae-Kyeung Moon)

1992년 2월 : 영남대학교 전산공학과 졸업
2000년 2월 : 영남대학교 컴퓨터공학과 박사
2000년 4월~현재 : 한국전자통신연구원 책임연구원

<관심분야> IoT 데이터분석, 빅데이터 압축/이상감지 기술, 스마트 팜

김 효 선 (Hyo-Seon Kim)

2002년 2월 : 안동대학교 통계학과 졸업
2019년 2월 : 안동대학교 컴퓨터교육전공 석사
2020년 9월~현재 : 안동대학교 창의소프트웨어전공 박사과정
2022년 1월~현재 : 한국전자통신연구원 위촉연구원

<관심분야> 빅데이터, 이상탐지, 머신러닝, 정보(SW·AI) 교육, 스마트 팜