

기상 IoT 데이터의 전처리를 통한 서리 예측 성능 향상

김효선, 문애경*, 송희현**

ETRI/안동대학교, ETRI*, 안동대학교**

etri2022_khs@etri.re.kr, *akmoon@etri.re.kr, **hhsong@anu.ac.kr

Improved frost prediction performance through pre-processing of weather IoT data

Kim Hyo Seon, *Moon Ae Kyeong, **Hee Heon Song

요약

본 논문은 기상 IoT 데이터를 활용한 서리 발생 예측의 성능 향상을 위하여, 데이터 전처리에 대한 연구를 제시한다. 서리 발생 여부는 아직까지 목측이 유일하고, 이상 탐지 분석에서는 학습데이터의 신뢰성을 높이는 것이 분석에서 가장 중요할 것이다. 따라서 본 논문에서는 동일 지역의 기상청 데이터를 추출하여 자체 데이터셋과 분포 구간을 비교하고, 제외할 이상치 구간을 설정한다. t-검정으로 유사성 유무 그룹을 도출하고, 해당 그룹 및 IR 비율별 예측 모델에 대한 성능을 평가한다. 결과에서는 유사성이 높고, 불균형 비율이 낮은 그룹에서는 데이터 증강시 예측 성능 향상에 유리하고, 서리 예측시 대부분의 그룹에서 “solar, gtp, t-soil, Tdp”이 높은 중요도로, “rain”은 최저 중요도로 나타난다.

I. 서론

서리는 지면 접지층의 대기가 함께 안정되면서 냉각되고, 기온이 0°C이하의 이슬점으로 내려가면서, 포화상태의 공기(수증기)가 승화하여 발생하는 자연현상이다[1]. 특히 초봄에 발생하는 서리 현상은, 농작물에 열매가 생성되기 위한 초기 단계인 개화기에 발생되므로, 그 해 농작물의 수확량에 큰 영향을 미치고, 농민에게 경제적 어려움을 줄 수 있다. 농가에서는 송풍법, 연무법, 살수법 등을 활용하여 서리 피해를 최소화 하고자 노력하고 있다[2]. 따라서 농가에서는 서리 피해를 사전에 예방하고자, 보다 정확한 서리 발생에 대한 예보를 요구하고 있지만, 서리의 관측은 아직까지 농업 전문가의 목측에 의한 방법이 유일하므로, 수집된 기후 데이터에 대한 서리 발생 유무를 판별한 데이터의 수집은 매우 어렵다. 따라서 본 논문에서는 기상 IoT 센서를 통해 수집된 빅데이터를 바탕으로, 서리 예측의 보다 정확한 성능 향상을 위한 전처리 과정을 제시하고자 한다.

II. 기상데이터 수집

본 연구에서는 서리 예측을 위하여, 경상북도 영천지역의 일부 과수원내에서 총 12개의 기상 IoT 장치를 통하여, 2015년부터 2021년까지 기상데이터셋을 수집하였다. 이중 농업인의 목측에 의한 서리 발생 여부 측정은 2017년부터 2021년간 수집하였다. 그리고 해당 관측소 데이터셋에 서리발생 클래스별로 라벨링(frost:1, nonfrost:0)을 처리하고 데이터셋으로 활용한다. 그러나 서리는 발생 기간중 생성 조건이 맞을시, 국소적으로 발생되므로써, 서리 미발생 데이터 크기 대비 서리 발생 데이터 크기가 매우 소수이다. 따라서 이를 보완하고자, 기상청(Korea Meteorological Administration, KMA)[3]에서 제공하는 종관기상관측(Automated Synoptic Observing System, ASOS), 농업기상관측(Automated Agricultural Observing System, AAOS) 및 계절관측 데이터를 활용하고, 이중 동일한 영천지역내 서리 발생 및 미발생 데이터를 추가로 확보하여, 서리 발생에 따른 기후적 요소에 대한 전처리시 활용하고자 한다.

자체 관측소(2017~2021년)와 기상청(2000~2008년)을 통해 수집한 데이터셋의 기후 요인은, 기존 연구[4]를 바탕으로 대기온도(TP), 초상온도(GTP), 토양온도(NR_9), 강수량(RF), 풍속(WeS), 일사량(ASR), 이슬점

온도(Tdp) 및 상대습도(R_HMDT)로 총 8종만 활용한다. 기상청 데이터셋에는 시작과 끝서리 일자만 제공하므로, 해당 일자의 일주일 전후의 자료를 서리가 내리지 않은 데이터셋으로 활용한다. 기상청 데이터셋의 기후 요인별 분포는 봄가을의 특정 일자만을 포함하므로, 자체 관측소 데이터셋에 비해 큰 변화폭을 나타내는 데이터는 없었고, 자체 관측소의 데이터셋의 분포범위 이내로 나타나고 있었다. 또한 Fig. 1에서와 같이 일부 요인은 서리 발생일과 그렇지 않은 날의 기후 차이가 뚜렷히 구분 되었다.

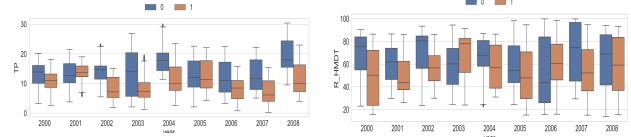


Fig. 1. KMA Weather data distribution(TP, R_HMDT)

III. 기상데이터 검증

두 데이터셋의 분포를 바탕으로, 해당 지역의 통상적인 기후 분포를 판단하고, 예측 방해를 위해 제거가 필요한 이상치의 기준 값을 Table 1과 같이 제시한다. 또한 이상치 기준 설정은, 서리 발생 기간(일자)을 모두 포함하는 구간으로 설정한다. 또한 본 데이터셋을 바탕으로 해당 기후 요인을 통한 별도의 일자별 대푯값(일사량(solar), 일교차(t_diff), 최저기온(t_min), 상대습도(gtp), 최저 상대습도(gtp_min), 강수량(rain), 풍속(w_speed), 토양온도(t-soil), 이슬점(Tdp): 총 9종)을 생성하고, 이 대표값을 기계학습의 입력값으로 활용한다.

Table 1. Distribution of IoT station's collection dataset

Climate \ Stat	TP	GTP	WeS	RF	ASR	R_HMDT	NR_9	Tdp
mean	12.178	12.388	1.22	1.65	6.13	70.28	25.97	5.98
std	10.66	12.21	1.31	9.32	6.95	23.50	11.15	11.72
min	-100	-100	0	0	0	0.1	0	-124.694
max	104.7	104.7	102.3	3276.7	327.67	102.3	6553.5	105.3438
skew	-0.081	-0.899	2.329	72.737	1.476	-0.472	287.201	-0.223
kurt	-0.777	7.527	80.893	23228.1	12.131	-0.934	1668546	-0.837
out_min	-21.95	-24.4	0	0	0	0	0	-31.6
out_max	46.45	50	20	470	77	102.3	60	45

Table 2. Comparison t-test between stations(mean)

STATION- STATION	3001- 8001	3001- 9001	4001- 6001	4001- 9001	4001- 9100	4001- 9300	6001- 9100	8001- 9001
p-value	0.019	0.116	0.004	0.086	0.108	0.011	0.012	0.025

관측소별 수집 데이터셋의 일자별 대표값을 추출하고, T-검정을 시행하였다. Table 2는 전체 t-검정 결과중 특이값만 추출한 결과이다. 이상치를 제거한 데이터셋이므로, 대부분의 p-value는 0에 가깝게 측정되었다. 그러나 일부 모델(4001, 6001, 8001, 9001, 9100, 9300)에서 0.01이상의 값이 나타났고, 그중 9001은 대부분의 모델 비교에서 값이 높게 나타났다. 그리고 기후 요인 간 비교에서도 대부분이 0에 가깝고, 일부 요인(RF)에서 최대 0.02 이하로 가장 많은 빈도로 차이가 나타났다.

따라서 본 연구에서는 t-검정의 p-value가 0에 가까운 그룹(A)과 그렇지 않은 그룹(B)으로 구분하여 서리 발생 예측에 대한 성능 평가를 시행한다. A(1001, 2001, 5001, 7001, 9200)는 고연관성 그룹, B(3001, 4001, 6001, 8001, 9001, 9100, 9300)는 저연관성 그룹, C는 전체 관측소 그룹으로 한다. 또한 그룹별 무상기간 포함 유(1*)·무(2**)와 클래스간 불균형 해소(SMOTE)를 적용하여 재분류한다. SMOTE 적용시,

$$IR = \frac{Minority(frost:1)}{Majority(nonfrost:0)} \text{은 } 0.2 \text{와 } 0.4 \text{ 비율로 자체 설정후,}$$

RandomForest Classifier 예측을 시행하고, 성능 지표(Accuracy, Recall, Precision, F1, AUC)에 대해 비교(F1)를 시행한다(Table 3, Fig 2).

Table 3. Performance evaluation table according to t-test test result group and IR vision

Group	IR	accuracy	precision	recall	auc	f1	f1_sub	
							Aft-Bef	2*-1**
A1*	0.04	0.960	0.709	0.529	0.529	0.544		
	0.2	0.939	0.925	0.845	0.845	0.878	0.334	
	0.4	0.954	0.942	0.942	0.942	0.942	0.063	
A2**	0.06	0.940	0.845	0.519	0.519	0.521		-0.023
	0.2	0.920	0.906	0.786	0.786	0.830	0.309	-0.048
	0.4	0.911	0.903	0.878	0.878	0.889	0.059	-0.053
B1*	0.01	0.992	0.496	0.500	0.500	0.498		
	0.2	0.985	0.983	0.957	0.957	0.97	0.472	
	0.4	0.987	0.982	0.984	0.984	0.983	0.014	
B2**	0.01	0.989	0.495	0.500	0.500	0.497		-0.001
	0.2	0.977	0.974	0.942	0.942	0.957	0.460	-0.012
	0.4	0.978	0.980	0.967	0.967	0.973	0.016	-0.010
C1*	0.02	0.964	0.958	0.954	0.954	0.956		
	0.2	0.962	0.948	0.912	0.912	0.929	-0.027	
	0.4	0.964	0.958	0.954	0.954	0.956	0.027	
C2**	0.04	0.968	0.484	0.500	0.500	0.492		-0.464
	0.2	0.943	0.934	0.843	0.843	0.881	0.389	-0.048
	0.4	0.947	0.938	0.927	0.927	0.933	0.052	-0.023

* 1: Free frost period(May-September) included

** 2: excluding free frost period(May-September)

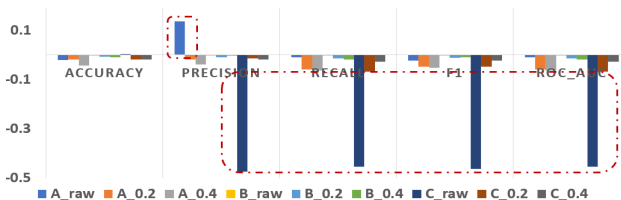


Fig. 2. Comparison of prediction performance between models(2*-1*)
(Accuracy, Precision, Recall, F1, AUC)

모든 그룹에서 예측 성능은 대부분 0.9 이상으로 높게 나타난다. raw dataset 그룹간 비교에서는 그룹간 유사성과 예측 성능이 비례한다(A1>A2>B1>B2>C2). SMOTE 적용 그룹간 비교에서는 C1을 제외한 나머지 그룹에서 IR 비율과 예측 성능은 비례하고, raw dataset의 IR 비율

이 낮은 그룹(B, C)은 높은 그룹(A)에 비해 SMOTE 이후 성능이 오히려 높게 나타난다(B1>B2>C1>C2>A1>A2). 이는 raw dataset의 소수 class 이 너무 작을시에는 SMOTE에 따른 과적합이 나타남을 예상할 수 있다. 또한 대부분의 모델에서 무상기간을 포함시 예측 성능이 향상되었다.

서리 예측 모델 생성에 영향력이 높은 요인을 알아보기 위하여, Table 3의 그룹 및 IR 비율별 모델에 대한 RandomForest Classifier(RF)의 Feature importance(by Scikit-learn)을 확인한다. 유사성 그룹 A, B, C는 SMOTE에 상관없이 동일 그룹간 유사한 분포를 나타냈다. Fig 3는 각 유형(무상기간 포함 및 미포함)별 RF 입력값(9종)에 대한 중요도 예시 그래프이다. 전체 그룹 및 세부 유형에서 공통적으로 높은 요인은 A·C 그룹은 “solar, gtp, t-soil”이고, B 그룹은 “solar, gtp, Tdp”이다. “rain”은 모든 그룹에서 가장 낮은 중요도로 나타난다.

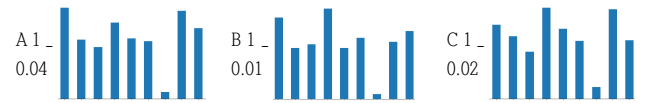


Fig. 3. RF of feature_importance(Up: Free period included, Down: Excluding free period)
(x: solar, t_min, t_diff, gtp, gtp_min, w_speed, rain, t-soil, Tdp / y: Importance)

IV. 결론

본 논문에서는 기상 IoT 데이터를 활용한 서리 발생 예측의 성능 향상을 위하여 데이터 전처리에 대한 연구를 제시한다. 데이터셋은 경북 영천 지역내 과수원에서 관측소(IoT 센서)를 통해 수집되었고, 해당 지역내 전 문가(농민 포함)를 통한 목측 관측으로 서리발생 여부를 수집한다. 서리 발생 유무에 따른 클래스 불균형이 필수적으로 나타나므로, 기상청의 공공데이터를 활용하여, 자체 데이터셋의 분포와 추가로 비교하고, 제외할 이상치 구간을 확정한다. t-검정을 통한 관측소별 유사성을 확인하였고, 연관성에 다른 A그룹(5개)과 B그룹(7개)으로 구분함으로써, 데이터간 결합을 통한 소수 데이터의 확보와 데이터 증강의 효율성을 향상 시키고자 한다. 또한 서리가 발생하지 않는 무상기간의 데이터를 제외한 그룹과 포함한 그룹으로도 구분한다. 성능 평가는 RandomForest Classifier를 활용한 F1-score를 적용한다. 결과에서는 유사성이 높고, 불균형 비율이 낮은 그룹에서는 데이터 증강시 예측 성능 향상에 유리하고, 서리 예측시 대부분의 그룹에서 “Solar, gtp, t-soil, Tdp”이 중요하게 나타난다.

ACKNOWLEDGMENT

본 연구 논문은 한국전자통신연구원 내부연구과제[22YD1100] 및 안동시 지원사업[22AD1100]의 일환으로 수행되었음

참 고 문 헌

- [1] Young-Ah Kwon et al., "The Weather Characteristics of Frost Occurrence Days for Protecting Crops against Frost Damage," Journal of the Korean Geographical Society, 43(6), pp. 824-842, 2008
- [2] Nongsaro(Agricultural Information Portal), "Frost occurrence date and free period distribution by region," August 23, 2022, (<http://www.nongsaro.go.kr/>)
- [3] Korea Meteorological Agency meteorological data open portal, August 23, 2022, (<https://data.kma.go.kr/cmmn/main.do>)
- [4] Yongseok Kim et al., "Study on Improvement of Frost Occurrence Prediction Accuracy," Korean Journal of Agricultural and Forest Meteorology, 23(4), pp. 295-305, 2021