2022년도 한국해양과학기술협의회 공동학술대회
6월 2일(목)~4일(토) 제주 국제컨벤션센터 및 부영호텔

기계 학습을 이용한 항로표지 기상 자료의 보간에 관한 연구

정성훈* · 마준익** · 조성현** · 임기륜** · 이준우** · † 한준희

*부산대학교 대학원 석 · 박사통합과정생, **동아대학교 학부연구생, † 부산대학교 산업공학과 교수

Study on Weather Data Interpolation of a Buoy Based on Machine Learning Techniques

*Seong-Hun Jeong** · *Jun-Ik Ma*** · *Seong-Hyun Jo*** · *Gi-Ryun Lim*** · *Jun-Woo Lee*** · † *Jun-Hee Han*

**Student, Graduate School of Pusan National University, Pusan 46241, Korea*
***Student, UnderGraduate School of Donga University, Pusan 49315, Korea*
† *Professor, Department of Industrial & Management Systems Engineering, Pusan National University, Pusan 46241, Korea*

요 약 : 해상에 설치된 항로표지 부표의 발달로 다양한 자료가 수집된다. 그러나 원시 관측자료는 기계 결함 및 기상환경에 따라 결측과 이상치를 포함한 오류로 인하여 곧바로 사용되기 어렵다. 따라서 본 연구에서는 항로표지에서 수집된 미흡한 기상 관측 자료를 기계학습이 가능하도록 누락된 시각의 자료를 추가하여 선형 보간을 실시했다. 이후 XGBoost기법과 KNN-regressor을 이용하여, 오류가 발생한 시점의 자료를 보간하는 기법을 연구하고자 한다.

핵심용어 : 기계학습, 기상자료, 결측, 이상치, 보간

Abstract *: Several types of data are collected from buoy due to the development of hardware technology.. However, the collected data are difficult to use due to errors including missing values and outliers depending on mechanical faults and meteorological environment. Therefore, in this study, linear interpolation is performed by adding the missing time data to enable machine learning to the insufficient meteorological data. After the linear interpolation, XGBoost and KNN-regressor, are used to forecast error data and suggested model is evaluated by using real-world data of a buoy.*

Key words *: machine learning, meteorological data, missing value, outlier, interpolation*

Acknowledgement : This research was a part of the project titled 'Marine digital AtoN information management and service system development (2/5) (20210650)', funded by the Ministry of Oceans and Fisheries, Korea. 이 논문은 2021년 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (해양 디지털 항로표지 정보협력시스템 개발(2/5) (20210650))

1. 서 론

해양수산부에서 관리하는 해양 부표들은 연안 및 근해에 설치되어 목적에 따른 역할을 수행하며, 인근의 기상자료를 실시간으로 수집한다. 수집된 관측 정보는 실시간으로 해상 상황을 통하여 항해중인 선박의 항해환경 파악, 해양 기상 환경 관리, 정보를 통한 어플리케이션에 활용된다. 무선으로 통신되는 원시 관측자료는 기상 환경에 따른 통신장애와 기계 오류로 인하여 이상치 발생과 정보의 누락이 발생한다. 이러한 관측 오류 정보는 기상 정보를 활용하는 지점에서 오판단 및 정보의 부재를 야기할 수 있다.

기상청에서는 데이터의 이상치와 정보의 누락 등, 데이터의 품질을 관리하고자 "기상관측자료 실시간 품질관리시스템 (Real-time Quality control system for Meteorological Observation Data; RQMOD)"(오승준 외, 2006)를 개발하였다. 그러나 해양 부표의 경우 해양 교통에 관련한 표지의 기능에 집중되어 있으며 데이터 관리에 대한 연구는 미비하다.(해양수산부, 2021) 따라서 본 연구에서는 XGBoost, KNN을 통한 머신러닝 기법을 이용하여 오류 자료의 개선 및 보간하는 기법을 연구하고자 한다.

2.목표 자료 선정 및 전처리

해상 부표는 MMSI라는 고유 식별자로 구분된다. 따라서 본 연구에서는 비교적 온전하게 자료가 수집된 MMSI 440101007 부표의 2019년 01월 02일 ~ 2019년 02월 28일 기간의 습득 자료를 학습에 사용하였다.

해상 부표는 활용 목적, 관리 지자체에 따라서 각기 다른 센서를 부착하고 있다. 본 연구에서는 자료가 거의 존재하지 않는 부표를 제외하고 모든 부표에 사용하며 학습을 위한 데이터가

† 교신저자 : junhan@pusan.ac.kr
* skek4915@pusan.ac.kr

충분한 기압, 기온, 습도, 풍속의 4가지 인자를 유효 인자로 선정하였다.

수집되는 자료는 통신 환경에 따라서 다양한 간격으로 수집된다. 따라서 해당 연구에서는 분 단위로 하루를 1440개로 나누어 [0, 1439]구간의 day_min 인자를 추가하였으며, 각 해당 시간에 자료가 존재하지 않을 경우 결측값으로 자료를 추가하였다.

Table 1 선정 인자 유효구간과 수치 재조정

인자	기압	기온	습도	풍속
유효 구간	[900, 1100]	[-50, 70]	[0, 100]	[0, 80]
입력 변형	(x-900)/200	(x+50)/120	x/100	x/80

수집된 자료는 Table 1에서 표기된 대로 유효 구간 밖의 정보는 이상치로 분류하여 결측 처리하였으며, 절대적 수치 차이가 기계학습의 성능에 영향을 미치지 않도록 최소 값을 0 최대 값을 1로 갖는 수치로 재조정 하였다. 또한 결측이 발생할 경우 학습이 불가능한 기계학습 모델이 학습이 불가능한 경우가 발생하므로 결측값에 대해서는 존재하는 데이터간의 선형 보간을 실시하였다.

3. 기계학습 모델과 학습절차

기계학습 모델은 XGBoost, KNN의 2개의 회귀모델로 선정하였다. 학습의 입력값은 목적 자료 포함한 4개의 인자에 하루에 대한 시간 정보인 day_min을 추가한 5개를 입력값으로 갖는다.(Table 2) 목적 인자는 입력 인자에 대해서 1분 후의 자료가 사용되었다.

Table 2 입출력 인자 정의

입력 인자	예측 인자
day_min, 기압, 기온, 습도, 풍속	기압(1분 뒤)
day_min, 기압, 기온, 습도, 풍속	기온(1분 뒤)
day_min, 기압, 기온, 습도, 풍속	습도(1분 뒤)
day_min, 기압, 기온, 습도, 풍속	풍속(1분 뒤)

학습 모델은 scikit learn을 활용하여 제작했으며 실시간으로 많은 분석을 진행해야 한다는 조건에 맞추어 경량화하여 설계했다. XGBoost는 n_estneimators=500, max_depth=5의 조건에서 실행했으며, KNN은 10000개의 구간 분리를 설정하여 이웃 개수를 1개에서 100개 까지 적용하여 가장 잘 설명하는 것을 선정하였다.

마지막 일자인 2019년 02월 28일을 검증 구간으로 설정하고, 그 외의 모든 일자는 학습 구간으로 사용하였다. 학습 구조상 학습 순서에 대한 영향이 있지는 않지만 학습 결과가 하루에 대해서 특정 시간에 잘못 작동하는 경향을 파악하기 위해서 다음과 같이 설정하였다.

4. 결과분석

Table 3 학습 모델의 검증 구간에 대한 평균절대비오차

	기압	기온	습도	풍속
XGBoost	0.1270 %	0.2442 %	2.2105 %	7.3978 %
KNN	13.4774 %	6.3133 %	5.0708 %	15.4571 %

학습의 오차는 MAPE(mean absolute percentage error)로 XGBoost는 KNN에 대해서 모든 영역에서 정확도가 우수하였으며, 특히 기온과 기압에 있어서 99% 이상의 정확도를 도달하였다.(Table 3)

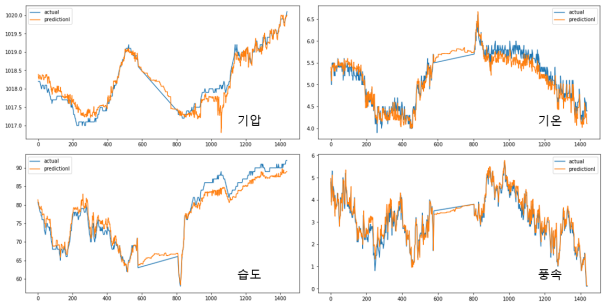

Fig1 XGBoost를 이용한 비교 결과

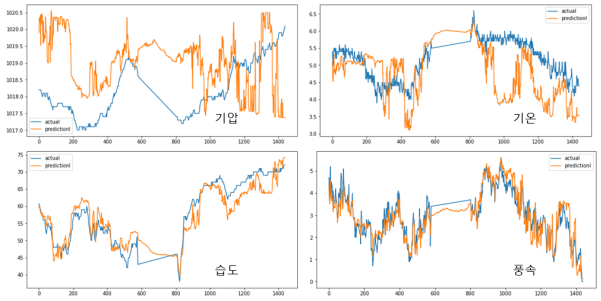

Fig2 KNN을 이용한 비교 결과

실측값과 예측값을 시각적으로 비교한 결과에서도 XGBoost는 모든 인자에서 적합하게 나타나지만 KNN은 기압과 기온에서 낮은 적합형태를 보여준다. (Fig 1, Fig 2)

5. 결 론

본 연구에서는 XGBoost와 KNN을 활용하여 해상 부표의 기상 자료의 보간에 머신러닝이 적합함을 보였으며, XGBoost가 평균 오차 2.5%로 높은 정확도 수준을 보여주었다. 그러나 해상 부표의 결측은 연속적으로 발생하는 경향을 보여주기에 보간값을 입력값으로 받는 점진적인 예측이 진행되어도 높은 적합도를 보여주기 위하여 정확도를 더욱 높여야 필요가 있다.

Acknowledgment

이 논문은 2022년 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (해양 디지털 항로표지 정보협력시스템 개발(2/5) (20210650))

참 고 문 헌

[1] 오승준, 이상우, 이종혁, 허복행, 유동봉, 이진아 & 박남철.(2006), “기상관측자료 실시간 품질관리시스템 Ⅰ (RQMOD Ⅰ) 구축 개발”, *한국기상학회 학술대회 논문집*., pp. 312-313.

[2] 해양수산부(2021). “2021년 항로표지 시행계획”