

항로표지 수집정보의 관리 기술 개발

양진홍* · 한준희** · 장준혁*** · 오세웅***

한윤석**** · 이예경**** · 정제한**** · 김민규* · 김호준* · 정성훈** · † 신상문

*인제대학교 의료·헬스케어IT공학과, **부산대학교 산업공학과,
선박해양플랜트연구소, *동아대학교 산업경영공학과 † 동아대학교 산업경영공학과 교수

요 약 : 항로 표지 수집정보의 관리에 있어서 데이터의 품질을 높이고 진단하는 것은 중요하다. 본 연구에서는 디지털 항로표지에서 수집되는 정보의 품질을 향상시키고 진단하기 위해 다양한 데이터 알고리즘을 비교 분석하였으며, 공정능력지수를 이용하여 데이터 품질진단 지수를 개발하였다.

핵심용어 : 데이터 평활, 선형 회귀, 자기 회귀, NG Boost, 공정능력지수

1. 서론

연구 배경

스마트항로표지로의 전환

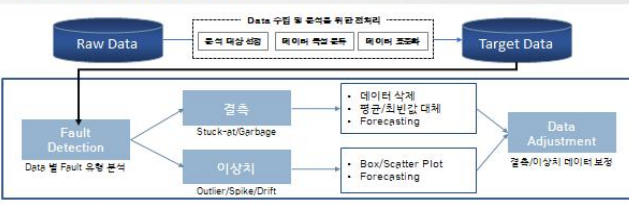
- 주요 국제기구와 선진국에서는 4차산업혁명 및 미래 해상항로 변화에 대응하기 위해 항로표지 기반 정보를 종합하여 디지털 표준화 개발 중
- 국내에서도 디지털 인프라 구축과 국가기반시설(SOC) 디지털화를 추진 중이며, 해양수산 분야에서는 자율운항선박, 스마트항만, 해상 환경 감시 등 해상 디지털 산업을 지칭하기 위한 인프라 구축 계획을 수립함
- '스마트 항로표지'는 자율운항선, 스마트항만 등 해상 디지털 산업의 활성화와 해상환경 변화에 능동적인 대응이 가능한 미래형 항로표지



<스마트 항로표지 개념도>

2. 연구 방법

데이터 품질 관리 기술 개발 분석



Raw Data → Data 수집 및 분석을 위한 전처리 (목적 대상 선정, 데이터 특성 분석, 데이터 조종) → Target Data

Fault Detection (Data 불, Fault 유형 분석) → 결측 (Stuck-at/Garbage), 이상치 (Outlier/Spike/Drift) → Data Adjustment (결측치/이상치 데이터 보정)

• 데이터 삭제
• 평균/치분값 대체
• Forecasting
• Box/Scatter Plot
• Forecasting

데이터 품질지수 개발

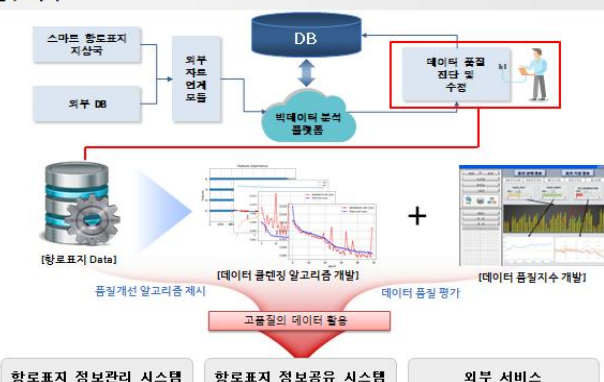
[다양한 시 모델을 활용한 데이터 결빙] → $C_p(i) = \frac{USL_i - LSL_i}{k\sigma}$ [Data 기반 품질 지수 산출]

품질 지수 기준	품질 지수 수준
1.67 ≤ Cp	품질 지수 매우 양호
1.33 ≤ Cp < 1.67	품질 지수 양호
1.00 ≤ Cp < 1.33	품질 지수 개선 필요
0.67 ≤ Cp < 1.00	품질 지수 부적
Cp < 0.67	품질 지수 매우 부적

[장성적 평가]

1. 서론

연구 목적



스마트 항로표지 자산과 외부 자원 연계 모듈, 외부 DB, DB, 데이터 품질 진단 및 수집, 빅데이터 분석 플랫폼, 데이터 클렌징 알고리즘 개발, 데이터 품질 평가, 데이터 품질지수 개발, 품질개선 알고리즘 제시, 고품질의 데이터 출생

항로표지 정보관리 시스템, 항로표지 정보공유 시스템, 외부 서비스

3. 연구 결과

타겟 데이터 선정

▶ 데이터 수집 현황 분석

- 수집중인 정보의 경우 총 62개 열명 중 19개 열명에 대해서만 데이터가 수집됨
- 수집중인 항목(열명)의 경우에도 누락 데이터의 개수가 확인건을 넘는 항목이 13개 존재
- 항로표지장치 모델도 데이터 수집 주기와 수집일이 일정하지 않음



데이터 불, 데이터 유

TB_STATUS 상세 정보 데이터: 총 데이터 속성 = 62개

TB_STATUS 기상 정보 데이터: 총 데이터 속성 = 55개

항로표지수집 현황

구분	신항 기간	속성
TB_STATUS	2017.09.15	MAIN_VOLT_STATUS
TB_WEATHER_101	2019.01.01	WIND_SPEED AIR_TEMPERATURE

Data 설명 기준: 데이터 연속성 문제, 데이터 누락, 데이터 값의 변동성

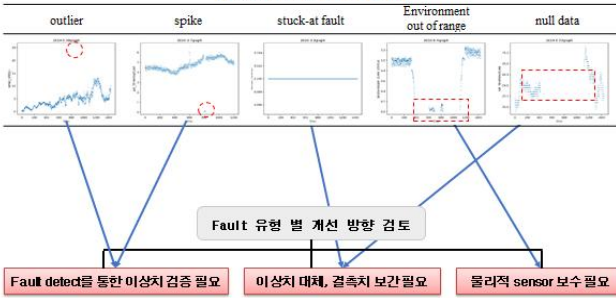
† 교신저자 : 신상문, sshin@dau.ac.kr

3. 연구 결과

데이터 이상치 구분 (fault classification)

> Data 별 Fault 유형 분석

[참조표지 데이터의 Fault type 별 사례]



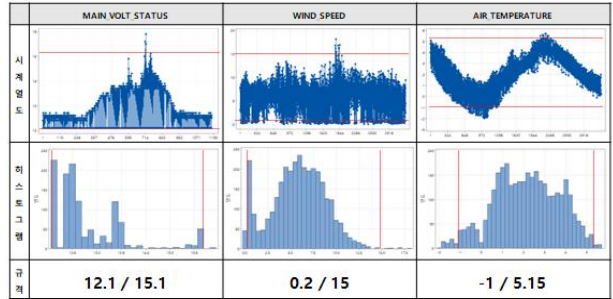
-7-

3. 연구 결과

데이터 품질지수 개발

> 전체 데이터에 대한 분포 파악

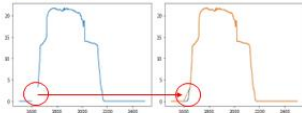
- 데이터 속성 별로 특정 구간마다 변동성이 존재 → 1일 동안의 데이터만 분석 수행



3. 연구 결과

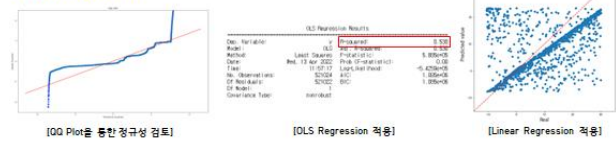
데이터 결측치 보간 알고리즘 개발

> Smoothing



- 임의로 생성한 결측값을 Smoothing 기법으로 보간했을 때(주황색), 같은 데이터와 유사한 추세로 데이터 보간되었지만 **단순히 선형적으로 보간됨**

> Linear regression



[Q-Q Plot을 통한 정규성 검토] [OLS Regression 적용] [Linear Regression 적용]

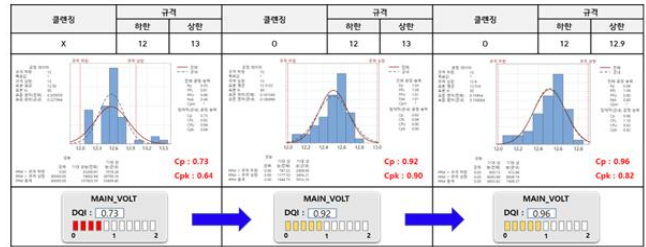
- Linear Regression을 통해 결측치를 예측했을 때, R-squared 점수와 비슷한 수치의 데이터가 정상적으로 예측됨

-8-

3. 연구 결과

데이터 품질지수 산출 결과 비교

> MAIN_VOLT_STATUS



-11-

3. 연구 결과

데이터 이상치 보간 알고리즘 개발

> Autoregression



- 일정한 패턴이 반복되다가 데이터가 크게 튀는 부분(이상치)을 대상으로 Autoregression 기법 적용
- 큰 차이가 나면 이상치로 패턴의 평균값으로 보간하였지만 **패턴까지는 예측하지 못함**

> NG Boost



- NGBoosting 기법을 활용해 기존 분산치를 이용한 데이터 결측치 아닌 데이터 변화에 따른 이상치 검정
- **부분적으로 이상치 결측치 적용되었음**

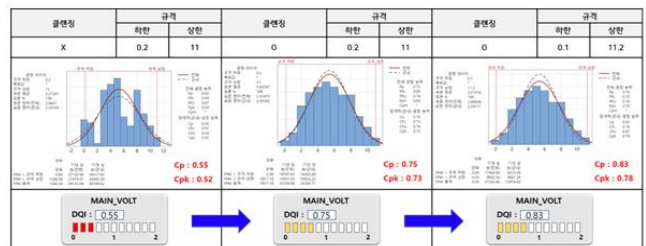
*NGBoosting을 이용한 유효구간 선정
- 초특성 경 : 양분 데이터 값
- 특관선 : NGBoost가 예측한 데이터 값
- 회색구간 : NGBoost가 예측한 데이터 유효구간

-9-

3. 연구 결과

데이터 품질지수 산출 결과 비교

> WIND_SPEED

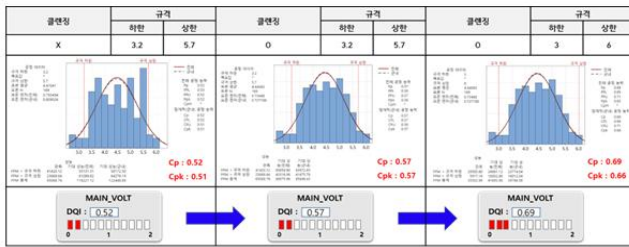


-12-

3. 연구 결과

데이터 품질지수 산출 결과 비교

▶ AIR_TEMPERATURE



-13-

4. 결론

결론 및 향후 진행방향

Conclusion

- 본 연구에서는 항로표지 데이터에 대하여 데이터 속성에 맞는 데이터 클렌징 알고리즘을 개발하기 위해 다양한 데이터 클렌징 분석 기법을 비교 및 평가하였음
- 또한, 데이터 품질에 대한 지수를 산출하였으며, 클렌징 전/후와 상/하한규격을 재조정된 데이터를 품질 지수를 비교함
- 비교 분석 결과, 데이터 속성에 따라 AI 모델의 분석 성능이 달랐으며 Linear regression 및 NG Boost가 가장 좋은 성능을 나타냄
- 또한, 기존의 Raw 데이터보다 클렌징 데이터의 품질지수가 더 높아짐에 따라 신뢰성 있는 품질지수를 확보할 수 있었음

Further study

- 데이터 속성마다의 특성이 존재하므로 각 속성의 특성에 맞는 데이터 클렌징 알고리즘에 대한 추가 연구 필요
- 데이터 기초 연구결과 각 항로표지 상태정보, 기상정보의 속성마다 상관성 및 고려해야 할 특성 존재 → 전문가와의 데이터 특성 및 적절한 상하한 규격 설정 등에 대한 논의 필요

-14-