다중 센서를 활용한 회귀 모델 진행 시 효율적인 데이터 전처리 기법

□ 최선, 안종현 / 가천대학교

요 약

다양한 센서의 사용이 가능해짐에 따라 센서 데이터 처리 및 선정의 중요성이 극대화되고 있다. 본고에서는 다중 센서를 통해 습득한 데이터를 회귀 모델에 사용하기 위한 다양한 전처리 기법을 실험하고 그 결과를 비교해 보고자 한다. 데이터 타입에 적합한 전처리 기법과, 노이즈를 줄임으로써 성능 향상을 달성하는 과정을 설명한다.

I. 서 론

데이터 처리에서 전처리는 매우 중요한 파트이다. 그 중 다양한 센서를 활용한 데이터는 그 특징과 형태가 다양하기에 더욱 다이나믹한 전처리 기법의 활용이 요구된다. 수많은 센서 데이터의 형태를 특정 기준에 맞추고, 원하는 결과가 적절히 도출될 수 있도록 해야 한다. 이 과정에는 정답이 없으며, 어떤 전처리 기법을 사용하느냐에 따라 데이터 분석의 결과가 달라지기에 다양한 실험이 필요하다. 모든 방식을 시도해 보는 것은 현실적으로 불가능하지만 제한된 데이터를 가지고 최대한 합리적인 결과물을 도출할 수 있도록 하기 위해 다양한 전처리 기법들이 제안되었다[1]. 또한, 입력되는 모든 센서 데이터를 사용하는 것은 제한된 컴퓨팅 자원에 치명적일 수 있기에 feature selection 과정을 추가하여 효율성 및 정확성을 모두 향상시키고자 한다.

본고에서는 여러 개의 센서 데이터가 존재할 때 효율적인 회귀 모델을 위한 과정을 제안하고자 한다. 이를 데이터 확인, 데이터 가공 및 특이값 제거, 데이터 전처리, STL decomposition noise reduction, feature selection 과정

을 단계별로 설명한다. 센서가 가지는 특성을 최대한 고려함으로써 특정 도메인과 필요한 작업에 가장 적합한 회귀 모델을 표현하는 것을 목표로 한다. 다양한 전처리 기법과 feature selection 기법, 그리고 회귀모델 종류를 실험하고 이에 대한 결과를 통해 센서 데이터 분석의 효율성과 정확성을 높이는 데 기여할 수 있을 것이다.

II. 데이터 처리 과정

다중 센서 데이터를 처리하기 위한 과정을 상세히 설명하며, 각 단계가 어떤 영향을 미치는지에 대해 살펴본다. 데이터를 정성적, 정량적으로 살펴본 후 어떠한 특징을 가지고 있는지 파악하는데 목적이 있다. 데이터의 형태 및 길이 등과 같은 특징에 따라 필요한 데이터 처리 기법들을 선별한다.

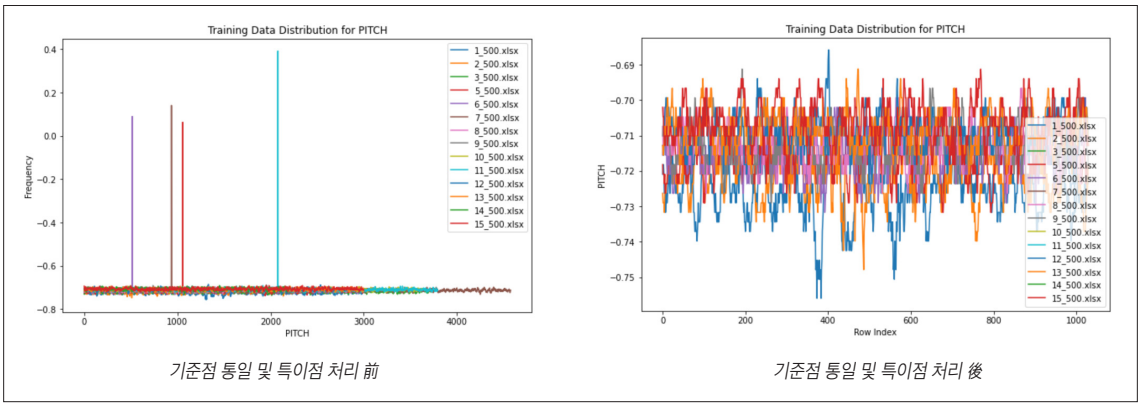
1. 데이터 확인

먼저 데이터 확인 단계에서는 데이터가 어떤 형태인지, 총 몇 개의 feature를 갖고 있으며, 그중 실제로 사용할 수 있는 feature가 무엇인지 결정하기 전에 시각적으로 확인한다. 이 단계에서는 데이터의 기초 통계량을 확인하고, 데이터 분포 및 결측치를 탐색하여 이후 단계를 위한 준비를 한다. 구체적으로, 데이터의 평균, 표준편차, 최솟값, 최댓값 등을 분석하고, 시각화한 데이터의 분포를 확인한다. 이를 통해 데이터의 형태를 직관적으로 확인할 수 있어, 데이터의 전반적인 특성을 파악할 수 있다. 데이터 특성 파악을 통해 데이터 가공 및 전처리 과정에서 고려해야 할 사항들을 도출할 수 있다.

2. 데이터 가공 및 특이값 제거

다음으로 진행되는 데이터 가공 단계에서는 사용하고자 하는 데이터의 기본적인 형태를 통일시키며 특이값을 제거한다. 서로 다른 시점에서 취득한 센서 데이터는 그 길이와 양이 다르기 때문에 이를 맞춰 줌으로써 실제로 사용할 수 있는 데이터 형태를 만들어 주는 과정이다. 먼저 본고에서는 전체 데이터의 유효한 데이터 길이를 맞추기 위해 타겟 시점을 기준으로 데이터를 잘라 사용했다.

센서 데이터 특성 상 튀는 값, 즉 특이값이 들어올 수 있다. 이러한 비정상적인 값들이 분석에 영향을 미치는 것을 방지하기 위해 미리 제거해 주어야 한다. 이러한 특이값을 제거하는 방식에는 quartiles를 기준으로 특정 범위를 지정하여 제거하는 IQR(Interquartile Range) 방식과 mean을 기준으로 특정 standard deviation 이상을



기준점 통일 및 특이점 처리 前

기준점 통일 및 특이점 처리 後

<그림 1> 기준점 통일 및 Outlier 처리 전/후 비교

초과하는 데이터들을 제거하는 z-score 방식 등이 있다. Z-score 방식은 각 데이터 포인트가 평균으로부터 얼마나 떨어져 있는지를 표준편차 단위로 측정하여, 일정 임계값을 초과하는 데이터를 제거하는 방식이다. 해당 과정에서는 특정 범위에 해당하는 값들이 아닌, 센서의 오류로 인해 존재하는 특정 값들만 제외하는 것이 목표이기에 z-score 방식을 사용했다. 제외한 특이값 데이터는 전/후 데이터의 평균값으로 대체함으로써 원본 데이터의 의미를 최대한 보존했다. 이러한 특이값 제거는 데이터의 품질을 향상시키며, 모델의 성능에도 직접적인 영향을 미친다. 실제로 해당 과정을 진행하기 전과 후의 차이는 시각적으로 봤을 때 매우 크다는 것을 알 수 있다.

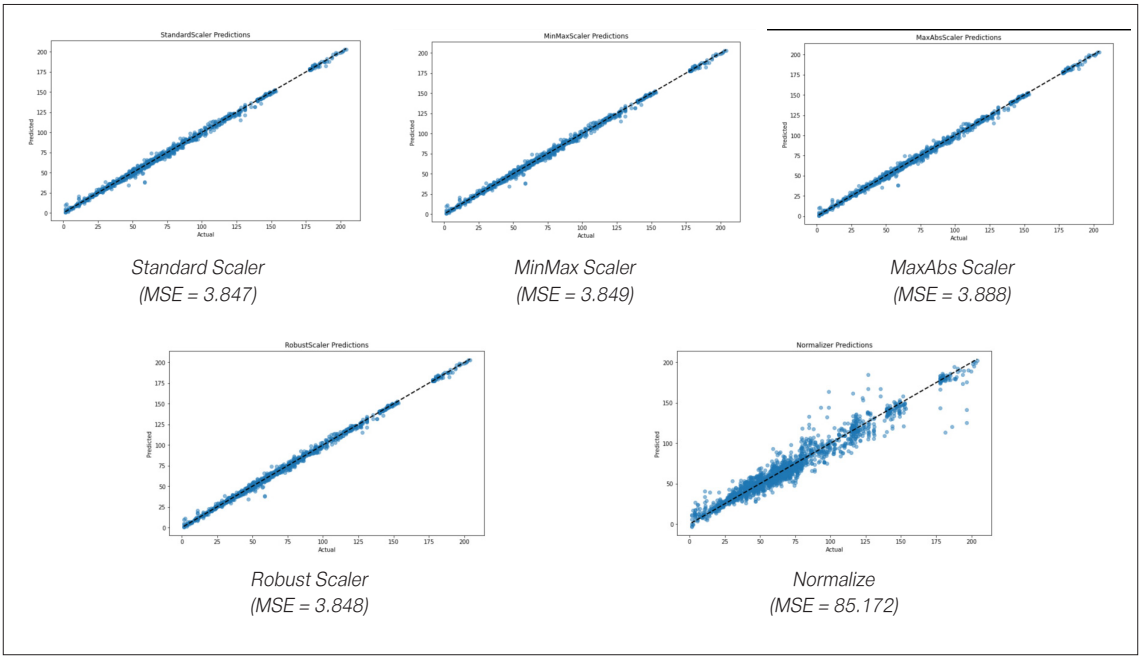
3. 데이터 전처리

데이터 전처리 단계에서는 데이터를 모델에 사용할 수 있도록 scaling 및 normalization을 진행한다. 이 단계에서는 데이터를 가장 잘 표현할 수 있는 전처리 기법을 탐색하는 것이 중요하다.

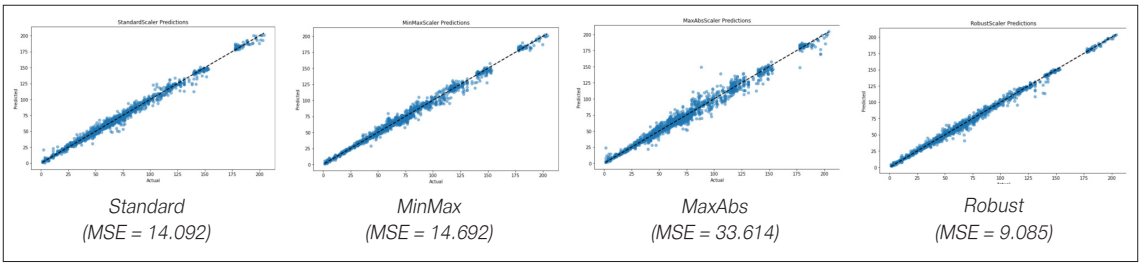
Scaling 기법은 데이터를 일정한 범위로 변환하는 것이 목적이다. 주로 변수 간의 크기 차이를 줄이기 위해 사용되며, 일반적으로 데이터의 최솟값과 최댓값을 0과 1 사이로 맞추는 방법이 많이 사용된다. 하나의 feature에 대해 독립적으로 적용되며, 특정 feature 내의 데이터들의 크기 차이를 줄인다[2].

Normalize 기법은 데이터의 크기를 조정하여 각 데이터가 동일한 크기(scale)를 가지도록 하는 과정으로 크기 1의 벡터로 변환하는 것이 목적이다. 하나의 데이터에 대해 독립적으로 적용되며, 특정 데이터의 각 feature 간의 벡터 차이를 줄인다[3].

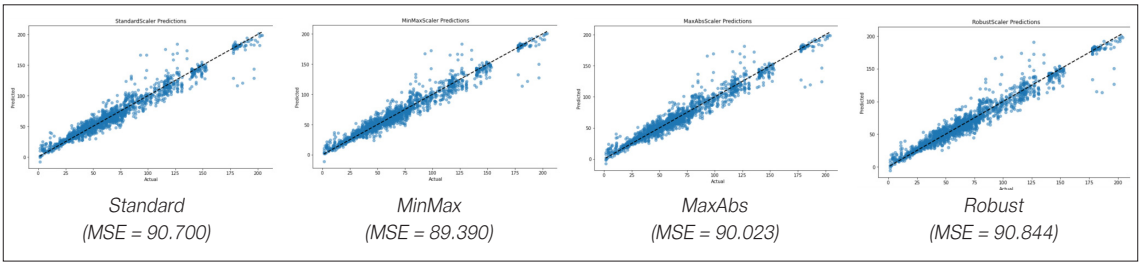
본고에서는 데이터셋에 가장 적합한 전처리 기법을 찾기 위해 총 4개의 scaler와 1개의 normalizer를 개별적으로 진행하는 실험과, normalize와 scaler를 둘 다 사용하지만 순서만 바꿔서 진행하는 실험을 통해 최종적으로 사용할 전처리 기법을 결정한다. 각 전처리 기법을 평가하기 위한 regressor로는 데이터의 non-linear한 형태를 잘



<그림 2> Scaling과 Normalize의 단독 실험 결과



<그림 3> Scaling 후 Normalize 실험 결과



<그림 4> Normalize 후 Scaling 실험 결과

표현할 수 있는 gradient boosting regressor를 사용했다. 〈그림 2〉는 scaler와 normalize를 단독으로 실행했을 때의 결과를 시각화와 함께 보여준다.

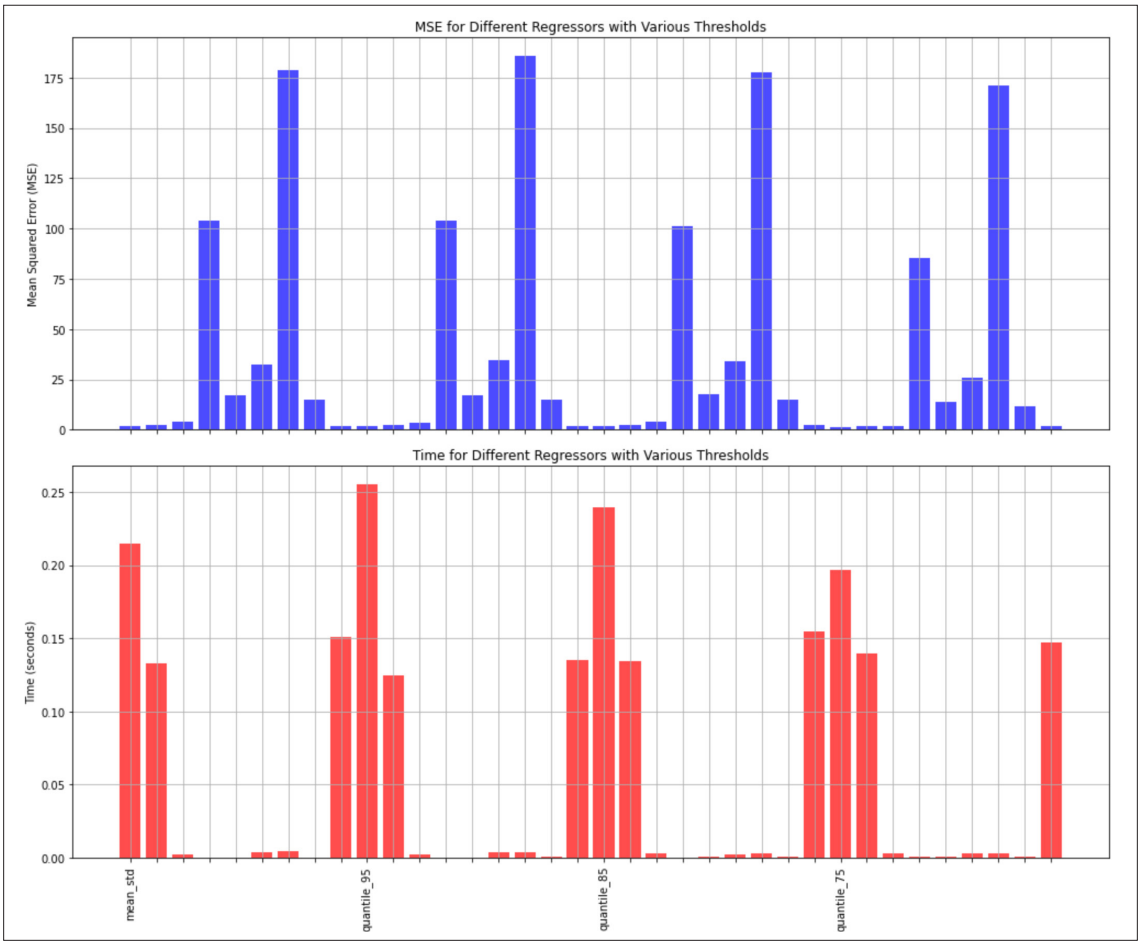
〈그림 3〉은 scaling을 먼저 진행한 후 normalize를 추가적으로 진행했을 때의 결과이며, 〈그림 4〉는 반대로 normalize를 먼저 진행한 후 scaling을 진행했을 때의 결과이다.

실험 결과, scaling을 단독으로 진행할 때 가장 낮은 MSE(Mean Square Error)가 나오는 것을 볼 수 있다. Scaling은 데이터의 범위를 제한하고 그 속에서 데이터의 형태를 잘 보존하기 때문에 다음과 같은 결과가 도출되었다고 예상할 수 있다.

4. STL 적용

기본적인 데이터 전처리가 끝났으니 본격적인 feature selection을 진행하기 전에 센서 데이터가 가지는 노이즈를 최대한 제거하고자 한다. 본 과정에서는 노이즈 제거를 위한 방법으로 STL(Seasonal and Trend decomposition using Loess) decomposition 기법을 활용한다. STL 분해 기법은 시계열 데이터를 세 가지 구성 요소인 계절성(Seasonal), 추세(Trend), 그리고 잔차(Residual)로 분해하는 방식을 사용한다. 특히 비선형성과 불규칙한 패턴을 처리하는 데 유용하며, 시계열 데이터의 복잡한 변동을 이해하고 예측하는 데 유용하다[4]. 본 실험에서는 센서 데이터가 시간의 흐름에 따라 값이 들어온다는 점에 집중해 해당 기법을 선택했다. 먼저 노이즈 제거를 위해 총 4가지 threshold를 설정했다. Threshold는 mean값을 중심으로 quantile 비율과 std를 활용했으며, 이를 STL 분해를 통해 얻어낸 잔차에 적용시킨다. Quantile 비율을 95/5, 85/15, 75/25로 지정하여 사용했다. Threshold를 초과하는 노이즈에 대해서는 제한을 걸어 값을 안정시키고 다시 재결합하는 recomposition 과정을 통해 데이터의 노이즈 제거 효과를 준다.

STL 분해를 통한 노이즈 제거 효과의 결과는 다음과 같다. 파란색 그래프는 MSE값을, 빨간색 그래프는 Time값을 나타낸다. 총 4가지의 threshold에 대한 9가지의 regressor 결과이다. MSE와 Time 모두 낮을수록



<그림 5> STL 노이즈 제거 후 전체 feature에 대한 회귀모델 결과

좋은 결과를 뜻한다. 각 그래프는 threshold를 기준으로 시각화하였고, threshold 그룹 내에서 regressor의 순서는 다음과 같다: RandomForestRegressor, GradientBoostingRegressor, DecisionTreeRegressor, KNeighborsRegressor, Ridge, Lasso, SVR, LinearRegression, XGBRegressor.

〈표 1〉은 〈그림 6〉의 결과를 바탕으로 가중치를 통해 MSE와 Time 결과를 적절히 반영한 랭킹 결과이다. 상위 5개의 조합을 보여준다.

〈표 2〉는 STL 분해 기법을 통한 노이즈 제거 과정을 거

<표 1> <그림 6>의 결과를 바탕으로 최적의 조합 결과

Rank	Feature Selection	Regressor	MSE	Time
1	Quantile_75	DecisionTreeRgressor	2.0341	0.0030
2	Quantile_95	DecisionTreeRgressor	3.6470	0.0033
3	Mean_std	DecisionTreeRgressor	4.1541	0.0028
4	Quantile_85	DecisionTreeRgressor	4.2704	0.0030
5	Quantile_75	LinearRegression	11.6487	0.0000

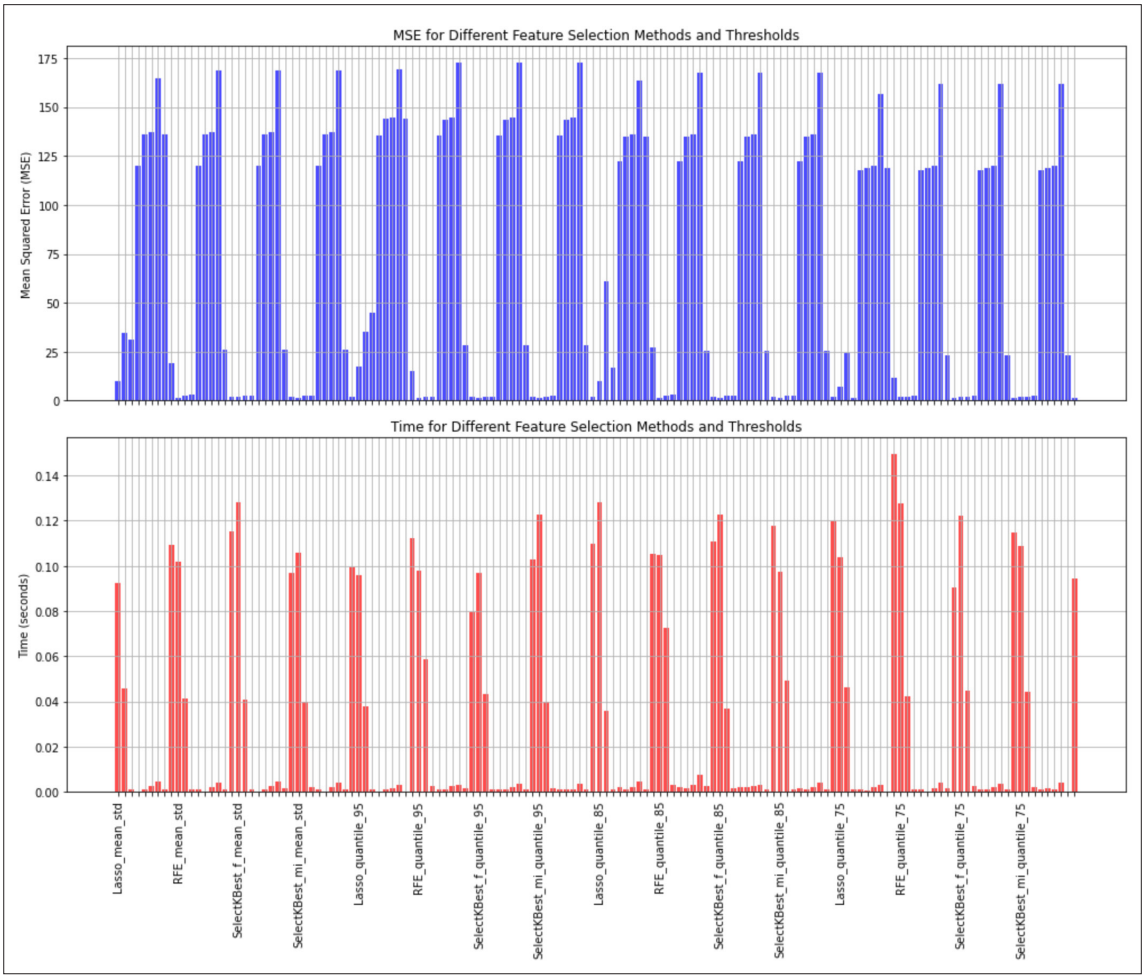
<표 2> STL 없이 진행한 전체 feature에 대한 회귀모델 결과

Rank	Regressor	MSE	Time
1	DecisionTreeRgressor	8.3062	0.1112
2	XGBRegressor	5.2619	0.2215
3	KNeighborsRegressor	372.8478	0.2109
4	GradientBoostingRegressor	26.0856	2.4533
5	LinearRegression	867.2516	0.0065

치지 않고 바로 전체 feature에 대해 회귀모델을 진행한 결과이다. 마찬가지로 가중치를 통해 MSE와 Time 결과를 반영하여 상위 5개의 조합을 보여준다. 〈표 1〉과 비교했을 때 MSE와 Time값이 평균적으로 훨씬 높은 것을 확인할 수 있다. 이를 통해 STL 분해 기법을 활용한 노이즈 제거가 효과적이라는 것을 증명할 수 있다.

5. Feature Selection 적용

마지막으로 feature selection을 적용하여 결과가 어떻게 달라지는지 살펴보고자 한다. Feature selection은 사용되는 feature의 개수를 물리적으로 줄여주어 연산 시간 감소에 효과적이다[5]. 뿐만 아니라 방해가 되는 feature를 제거함으로써 성능 향상에도 도움을 줄 수 있는 중요한 과정이다. 본고에서 사용한 feature selection 기법



<그림 6> STL 노이즈 제거 후 Feature Selection 기법에 대한 회귀모델 결과

<표 3> STL denoising 후 feature selection에 대한 회귀모델 결과

Rank	Feature Selection	Regressor	MSE	Time
1	SelectKBest_mi_mean_std	DecisionTreeRgressor	2.1867	0.0010
2	Lasso_quantile_75	DecisionTreeRgressor	1.9548	0.0019
3	SelectKBest_f_quantile_95	DecisionTreeRgressor	2.0428	0.0020
4	SelectKBest_mi_quantile_75	DecisionTreeRgressor	2.1320	0.0019
5	SelectKBest_f_quantile_75	DecisionTreeRgressor	2.3105	0.0019

은 총 4가지로 다음과 같다: Lasso, RFE, SelectKBest(f-regression), SelectKBest(mutual-info-regression)[6-8]. 〈그림 5〉의 그래프는 4가지의 feature selection 기법과 앞선 실험과 동일한 4가지의 Threshold 조합을 기준으로 9가지의 regressor를 시각화했다.

〈표 3〉은 〈그림 6〉의 결과를 바탕으로 가중치를 통해 MSE와 Time 결과를 적절히 반영한 랭킹 결과이다. 상위 5개의 조합을 보여준다.

〈표 1〉과 비교했을 때 〈표 3〉의 MSE와 Time값이 평균적으로 대폭 낮아진 것을 확인할 수 있다. 이를 통해 feature selection이 결과 향상에 효과적이라는 것을 증명할 수 있다.

III. 결 론

지금까지 다중 센서 데이터를 회귀모델에 활용하기 위한 데이터 전처리 및 노이즈 제거의 과정을 살펴보았다. 이 과정은 데이터 확인, 데이터 가공 및 특이값 제거, 데이터 전처리, STL 분해를 통한 노이즈 감소, 그리고 feature selection을 포함하는 단계별 접근법으로 구성된다. 각 단계는 센서 데이터의 특성을 최대한 고려하여 설계되었으며, 특정 도메인과 필요한 작업에 가장 적합한 회귀 모델을 표현하는 것을 목표로 했다.

4가지의 노이즈 제거를 위한 threshold, 4가지의 feature selection 기법, 9가지의 regressor을 활용한 총 144개의 실험 결과, 회귀모델에서의 STL 분해를 통한 노이즈 제거와 feature selection의 효과를 입증하였다. 이는 다중 센서 데이터 활용 시 불필요한 데이터를 제거하고 중요한 특징만을 선별하는 과정이 회귀모델 결과에 얼마나 큰 영향을 미치는지 잘 보여주는 실험으로, 효율적인 데이터 처리가 중요함을 강조한다. 이러한 결과는 다양한 도메인에서 센서 데이터를 활용한 회귀 모델 구축에 있어 실질적인 가이드라인을 제공할 수 있을 것이다.

참 고 문 헌

[1] García, Salvador, et al. “Big data preprocessing: methods and prospects.” Big data analytics 1 (2016): 1-22.
[2] Ahsan, Md Manjurul, et al. “Effect of data scaling methods on machine learning algorithms and model performance.” Technologies 9.3 (2021): 52.
[3] Patro, S. G. O. P. A. L., and Kishore Kumar Sahu. “Normalization: A preprocessing stage.” arXiv preprint arXiv:1503.06462 (2015).
[4] Cleveland, Robert B., et al. “STL: A seasonal-trend decomposition.” J. Off. Stat 6.1 (1990): 3-73.
[5] Y. Guo, W. Wang and X. Wang, “A Robust Linear Regression Feature Selection Method for Data Sets With Unknown Noise,” in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 31-44, 1 Jan. 2023.
[6] Venkatesh, B., and J. Anuradha. “A review of feature selection and its methods.” Cybernetics and information technologies 19.1 (2019): 3-26.
[7] Li, Jundong, et al. “Feature selection: A data perspective.” ACM computing surveys (CSUR) 50.6 (2017): 1-45.
[8] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. “A survey of feature selection and feature extraction techniques in machine learning.” 2014 science and information conference. IEEE, 2014.

저 자 소 개

최 선

- 2023년 : 가천대학교 AI/소프트웨어학과 학사
- 2023년 ~ 현재 : 가천대학교 AI/소프트웨어학과 석사 과정
- 주관심분야 : 딥러닝, 인공지능, 컴퓨터 비전, 3차원 센서, semantic segmentation

안 종 현

- 2020년 : 연세대학교 전기전자공학과 박사
- 2020년 ~ 2022년 : 국방과학연구소(ADD) Senior Researcher
- 2022년 ~ 현재 : 가천대학교 AI/소프트웨어학부 조교수
- 주관심분야 : 자율주행시스템, 정보융합, 머신러닝, 딥러닝, 다중객체 detection 및 tracking, Laser scanner 기반 인지 기술