KETI 한국전자기술연구원
Korea Electronics Technology Institute

Rev.0

모터 데이터 분석 - Feature Importance

한국전자기술연구원

2024.04

지능메카트로닉스연구센터
이정환 연구원

Analysis Pipeline - 1

◇ Data Preparation

- 30 vibration datasets (X, Y, Z axes + Timestamp)
- Each dataset has a different length

  → truncated to fixed length (**99,000 samples**)
  → 1 Chunk = **55 seconds (1,800 Hz × 55 sec)**

◇ Time Domain

- Extracted features (per-axis statistics): **Mean, Std, Variance, Skewness**
- X/Y/Z 3 axes × 4 features = **12 Features**
- Result: (30, 12) Feature Matrix

◇ Frequency Domain

- Sampling frequency: 1,800 Hz
- Operating frequency: 900 Hz
- FFT-based Harmonic Feature extraction (0~35 → total 36 features)
- Use **Feature Importance** to select top 12 features only -> prevent overfitting

Analysis Pipeline - 2

◇ Model Training

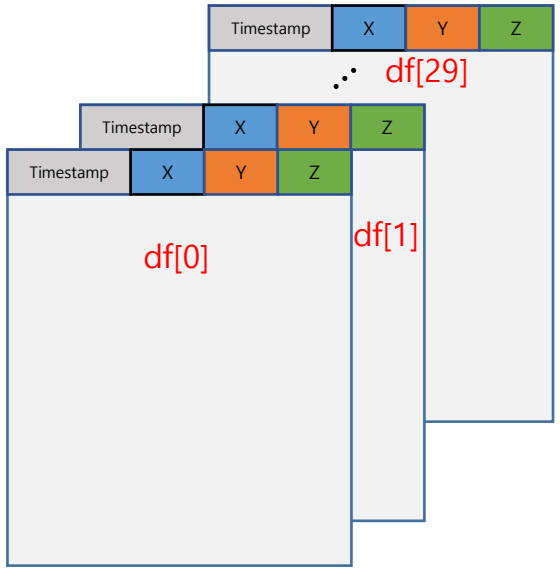
- **RandomForestClassifier (n=100)**
- Data split: Train/Test = 7:3
- Cross-validation (cv=5) → Accuracy evaluation

◇ Overall Process

1. Data loading & fixed-length truncation
2. Time-domain **feature extraction** (12 features)
3. Frequency-domain **feature extraction** (36 → top 12 selected)
4. Construction of Feature Matrix (X_data, y_data)
5. Random Forest training & performance evaluation

Time domain

30개의 진동데이터

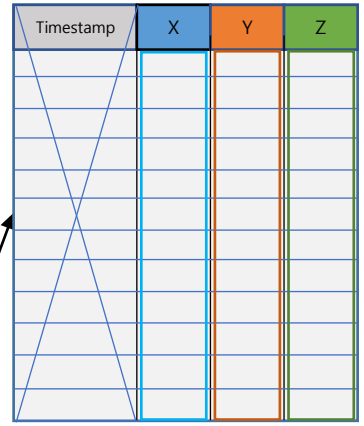


df[0].shape = (107346, 4)
df[1].shape = (102909, 4)
df[2].shape = (117343, 4)
.
.
.
df[29].shape = (101405, 4)

df.iloc[:99000, ::]

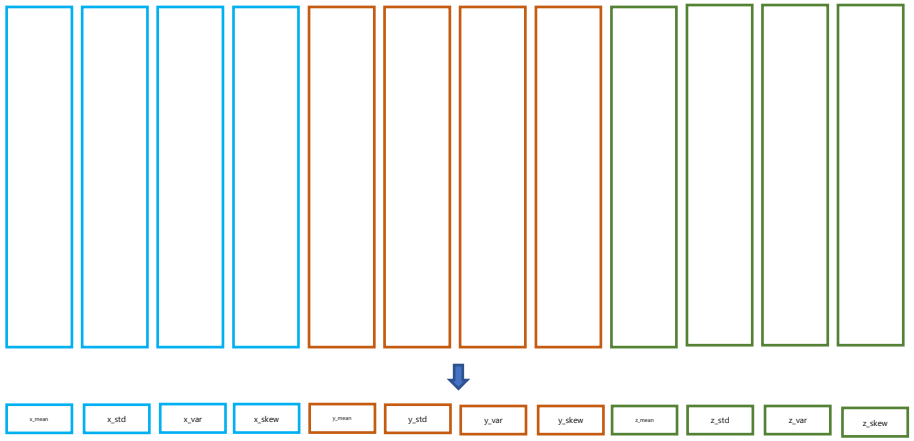
chunk_size로 나누기

df[0].shape = (99000, 4)
df[1].shape = (99000, 4)
df[2].shape = (99000, 4)
.
.
.
df[29].shape = (99000, 4)



1, 99000, 3

1, 99000, 3 -> 1 ,12



X 30개

1 ,12 -> 30, 12

Statistical Features 추출
(mean, std, var, skew)

X_data.shape = 30, 12

Y_data.shape = 30, 1

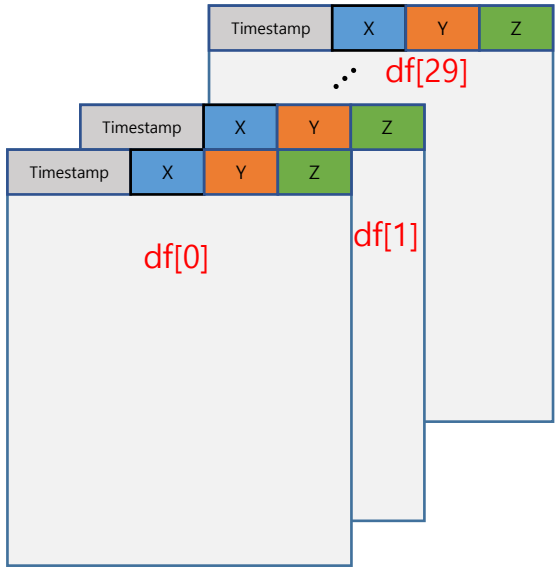
```
rf = RandomForestClassifier(n_estimators = 100)

y_pred_cv = cross_val_predict(rf, X_train, y_train, cv =5)
y_pred = np.argmax(y_pred_cv, axis = 1)

accuracy = accuracy_score(y_test, y_pred)
```

X_train, X_test, y_train, y_test = train_test_split(X_data, y_data, test_size=0.3)

chunk_size = 99000 (1800 (sampling_frequency)*55개)

Frequency domain

30개의 진동데이터



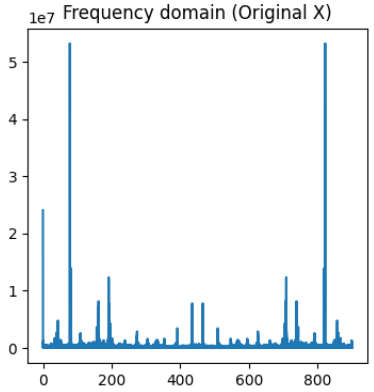
df[0].shape = (107346, 4)
df[1].shape = (102909, 4)
df[2].shape = (117343, 4)
.
.
.
df[29].shape = (101405, 4)

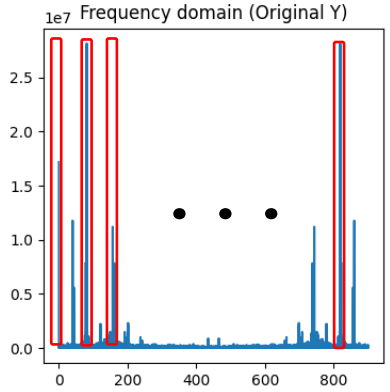
df.iloc[:99000, ::]

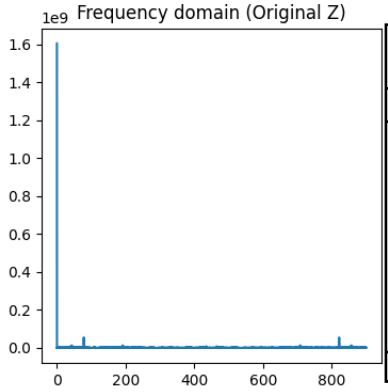
chunk_size로 나누기

df[0].shape = (99000, 4)
df[1].shape = (99000, 4)
df[2].shape = (99000, 4)
.
.
.
df[29].shape = (99000, 4)

Chunk_size = 99000 (1800Hz (sampling_frequency)*55개)





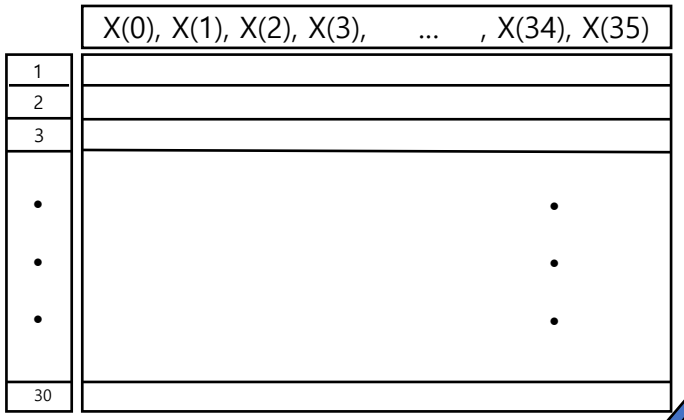


운전주파수 900 Hz
샘플링 주파수 1800 Hz

Ex) f(0) = 50Hz(**Harmonic Feature**)

X(0), X(1), … , X(35)

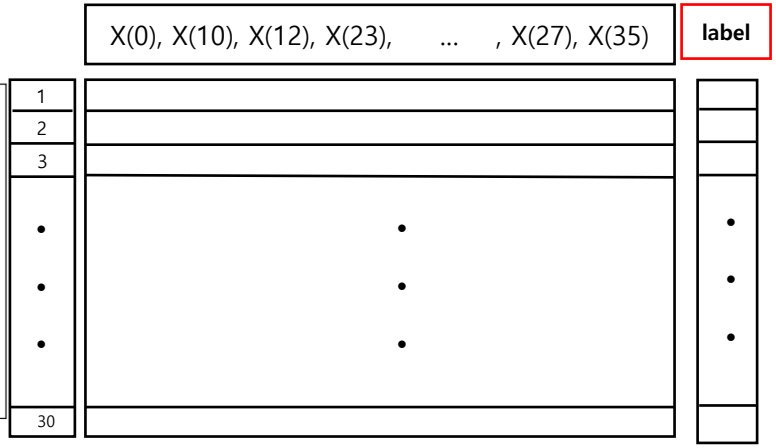
총 36개 특성



X(0), X(1), X(2), X(3), … , X(34), X(35)

너무 많은 특성을 사용하면 과적합 발생
과적합을 방지하기 위하여 **feature importance**을
사용하여 상위 중요도 12개만 사용

총 12개 특성



X_data.shape = 30, 12

Y_data.shape = 30, 1

X_train, X_test, y_train, y_test =
train_test_split(X_data, y_data, test_size=0.3)

rf = RandomForestClassifier(n_estimators = 100)

y_pred_cv = cross_val_predict(rf, X_train, y_train, cv =5)
y_pred = np.argmax(y_pred_cv, axis = 1)

accuracy = accuracy_score(y_test, y_pred)

KETI
embedded

Thank you

기업과 함께 성장하는 최고의 파트너

전자·IT분야 글로벌 전문생산연구기관