PhD Studentship in Production Technology - AI-Pipeline for Industrial Sensor Data Assimilation with ref no: 2024/97

Inas AL-Kamachy

July 15, 2024

1 Data Analysis

1.1 Load and Read the Data

First, I define a function, `read_and_clean_svmlight`, to read data files related to SVMlight format and perform basic cleaning by removing semicolons from the index. I perform this operation by looping over the 10 batch files, then combining all the loaded features (sensor readings) into a single DataFrame `X` and target labels (gas types) into a Series `Y` then concat the features and the target in one dataframe.

1.2 Identify the Outlier Using IQR

Using the Interquartile Range (IQR) method, the outlier was identified and replaced with the column median.

1.3 Data Normalization

To scale the data in the range (-1, 1), as recommended in the Gas Sensor Array Drift Dataset.

1.4 Feature Engineering

Add additional features named (month) which contain the Data volume for different sample gases in 10 batches.

1.5 Feature Selection

The main `data.csv` shape is $(13910, 129)$, so to reduce the dimensionality by selecting the most relevant features, trains an XGBoost classifier (`xgb.XGBClassifier`) for classification tasks, evaluating performance on training and testing data (`eval_set`). The `early_stopping_rounds` parameter aids in the automatic termination of training based on the validation set performance, guiding effective feature selection. The final `data.csv` shape was $(13910, 27)$.

2 ML Algorithms

We use four different ML algorithms:

- MLP Regression
- ElasticNet
- SVR
- Random Forest Regressor

2.1 MLP Regression

Complex Non-linear Relationships: MLP (Multi-layer Perceptron) Regression is well-suited for calibration tasks involving AI because it can model complex non-linear relationships between sensor inputs and outputs. This capability is crucial in AI-driven calibration where the relationship between sensor readings and actual values may be intricate and non-linear.

2.2 ElasticNet

Balanced Regularization: ElasticNet strikes a balance between L1 (Lasso) and L2 (Ridge) regularization, making it suitable for AI-driven calibration tasks. It effectively handles multicollinearity and feature selection, ensuring that only relevant sensor inputs influence the calibration process.

2.3 Support Vector Machine (SVM)

Non-linear Mapping: SVR is chosen for AI-driven calibration tasks due to its capability to map non-linear relationships between sensor inputs and calibration outputs effectively. This is crucial in scenarios where sensor drift introduces complex variations that need precise calibration adjustments.

2.4 RandomForestRegressor

Ensemble Learning Benefits: Random Forest Regressor benefits AI-driven calibration because it leverages ensemble learning to aggregate predictions from

multiple decision trees. This approach improves generalization and resilience against overfitting, which is crucial in handling diverse sensor data affected by drift, and besides, it is robust to Noisy Data: Each decision tree in Random Forest focuses on different subsets of data and features, making it robust against noisy sensor measurements. This capability ensures that AI-driven calibration remains accurate despite fluctuations caused by sensor drift.

3 Results

3.1 Visualization of RandomForestRegressor

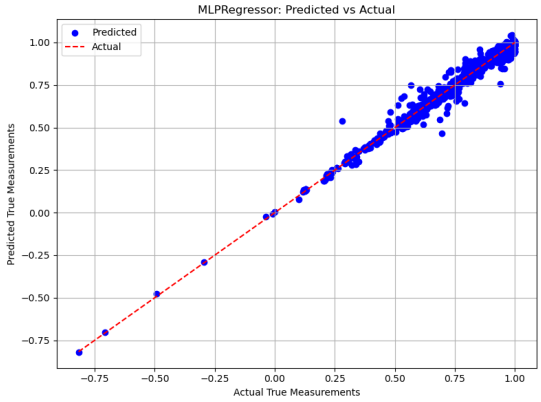

Figure 1: Best Model Visualization

3.2 Models Evaluations

Model	MSE	R^2	Training Time (s)	Prediction Time (s)	**Predicted True Measurements**
MLP Regression	0.0004	0.9886	0.671	0.00226	0.97314647
ElasticNet	0.0008	0.9748	0.0826	0.00054	0.95499576
SVR	0.0013	0.9611	0.258	0.00884	0.91607862
RandomForestRegre	0.0002	0.9943	3.0521	0.07438	0.96357482

Table 1: Performance Metrics of Various Regression Models

4 Conclusions

As we see from the result:

- MLP Regression achieved the lowest MSE (0.0004) and highest R^2 (0.9886), indicating very accurate predictions and a good model fit.
- RandomForestRegressor also performed well with very low MSE (0.0002) and high R^2 (0.9943), albeit with longer training and prediction times compared to other models.
- ElasticNet and SVR show slightly higher MSE and lower R^2 compared to MLP and RandomForest, suggesting they may be slightly less accurate for this particular calibration task.

  Finaly, Models like MLP Regression and RandomForestRegressor demonstrate high accuracy (low MSE) and precision (R^2 close to 1), making them suitable for precise calibration tasks where accurate measurement retrieval from sensor data is essential

5 Github

https://github.com/InasALKamachy/ML-for-calibration/tree/C