Task - PhD studentship in Production Technology - AI-Pipeline for Industrial Sensor Data Assimilation

Shinorina Shahrin Shaon

14 July 2024

1 Introduction

This work is about calibration using AI. Here, I designed 2 machine learning algorithms for the proposed UCI data. At 1st I loaded the data in my environment, performed some data prepossessing. Then, I applied PCA on the data, performed KNN and SGD algorithms and calculated accuracy for each algorithm. I applied the algorithm in each batch data separately. To complete this task I used Jupyter Notebook, ML libraries, Windows 11 with 11th Gen processor and 8.00 GB RAM.

2 Data Analysis

This data is industrial data and there are 10 different batches. From the site I came to know that there is no missing values. I extracted all the .dat file from the archive folder and loaded them on my environment. But there was some syntax error and I removed all the error from the .dat files manually. Then, the all .dat file was transformed to data frame by using pandas library. I called describe() and info() to see the details about the data frame. All the values are float values, and there were differences between all batch datasets. I used StandardScaler() to transform all the datasets and specially the train dataset values were fitted. I applied the algorithms on these high dimensional datasets. But the performance was very poor. To improve the performance, I used PCA() on all high dimensional datasets to transform them to lower dimensional data. This time I got better result than previous.

3 ML Algorithms

KNN is a supervised learning algorithm and it classifies the data points. The knn algorithm is more basic and common algorithm in machine learning. This

algorithm is one of the best algorithm for classification problems.
Another one is Stochastic Gradient Descent algorithm. This algorithm performs best to fit the models, optimizes the models and loss functions are used as hyper parameter. This algorithm is also performed in an iterative way and it works on individual data points. As, here, I have 10 different batch data.
I also tried several supervised machine learning algorithms - SVM gave me the almost similar results as KNN, Random Forest provided lower results than others, Logistic Regression is mainly used for binary classification.

4 Results

In this result section, all the Figures and Table are generated after performing PCA. I used Batch1 data as train data and other batches data as test data. I fitted the pca transformation on train data. Here, the number of principal component is 95%. I applied KNN and SGD algorithms on the lower dimensional datasets. If we look at Figure 2, the accuracy table, we can find that Batch2

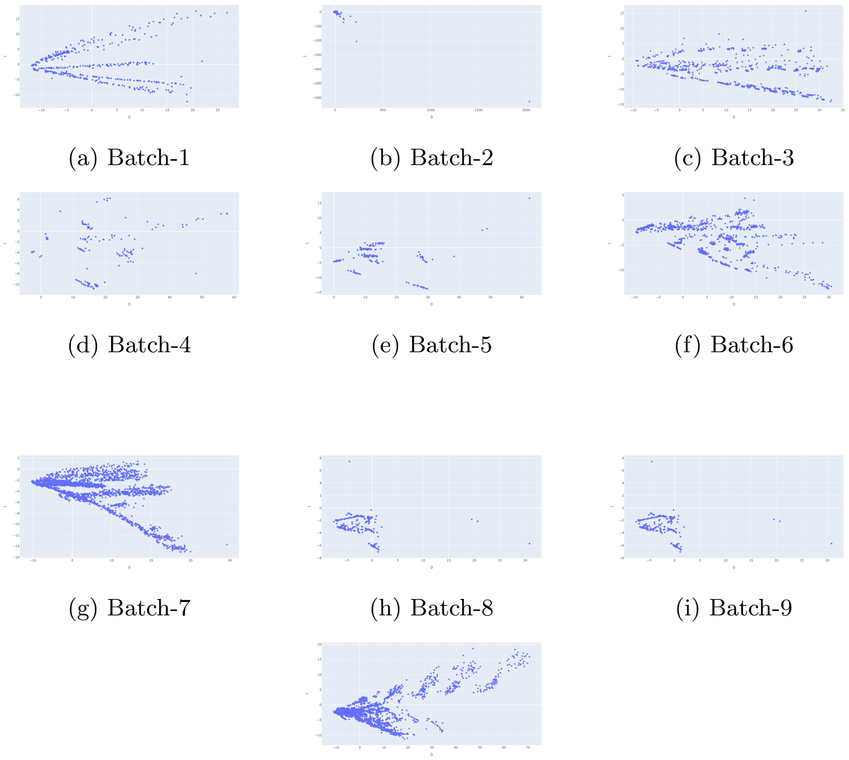
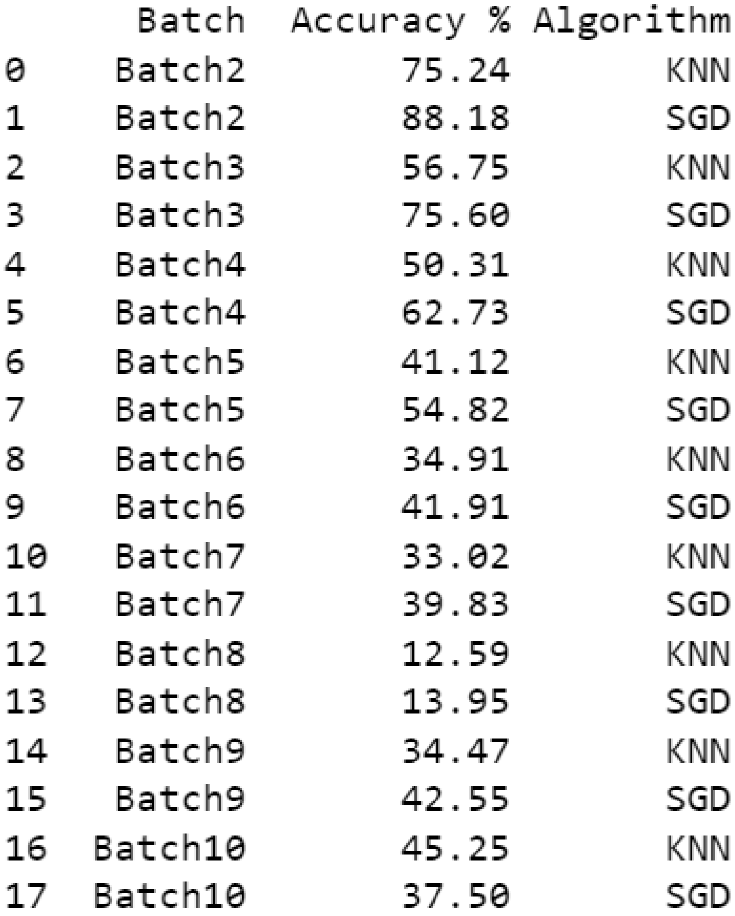

(a) Batch-1
(b) Batch-2
(c) Batch-3
(d) Batch-4
(e) Batch-5
(f) Batch-6
(g) Batch-7
(h) Batch-8
(i) Batch-9
(j) Batch-10

Figure 1: PCA on Batch Data

has higher accuracy rate, and it's about 88% for SGD. The lowest accuracy is about 12.59% for KNN algorithm at Batch8.



```
      Batch  Accuracy % Algorithm
0    Batch2       75.24       KNN
1    Batch2       88.18       SGD
2    Batch3       56.75       KNN
3    Batch3       75.60       SGD
4    Batch4       50.31       KNN
5    Batch4       62.73       SGD
6    Batch5       41.12       KNN
7    Batch5       54.82       SGD
8    Batch6       34.91       KNN
9    Batch6       41.91       SGD
10   Batch7       33.02       KNN
11   Batch7       39.83       SGD
12   Batch8       12.59       KNN
13   Batch8       13.95       SGD
14   Batch9       34.47       KNN
15   Batch9       42.55       SGD
16  Batch10       45.25       KNN
17  Batch10       37.50       SGD
```

Figure 2: Accuracy Table

From Figure 3 bar plot, we also can see that the only Batch2 data has higher accuracy than others for both KNN and SGD algorithm.

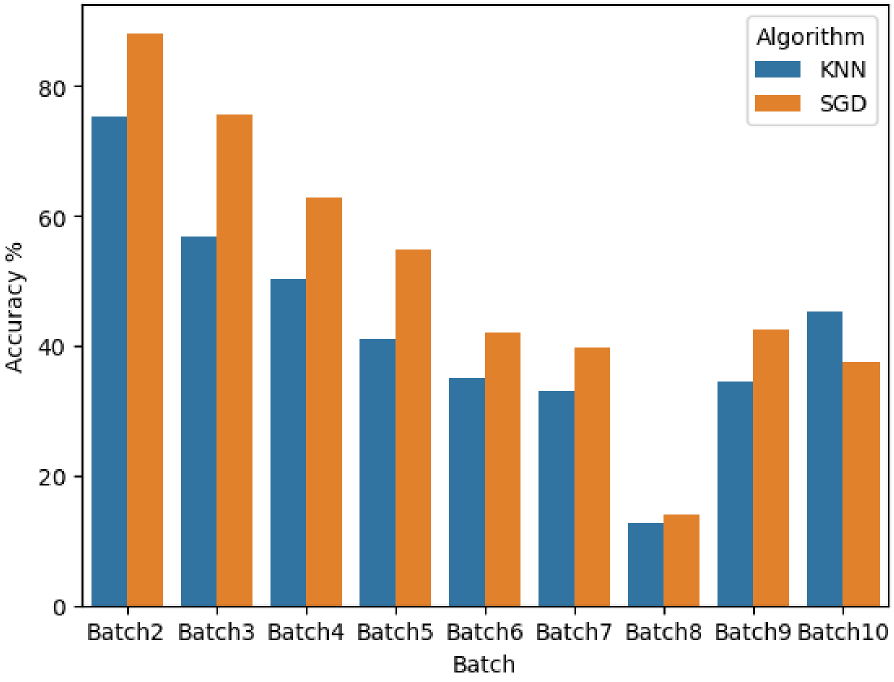


Figure 3: Bar Plot

5 Conclusions

From the table data and bar plot I can conclude that SGD is better than KNN and without PCA the results will be very poor. But I expected higher accuracy for all batch data. At the end I want to propose that, one can use SGD algorithm to classify sensor drift data.

6 References

1. https://www.mdpi.com/2076-3417/12/19/9529
2. https://www.mdpi.com/1424-8220/19/18/3844
3. https://github.com/miltongneto/Gas-Sensor-Array-Drift/blob/master/notebooks/analise.ipynb
4. https://medium.com/analytics-vidhya/calibration-in-machine-learning-e7972ac93555
5. https://medium.com/data-science-at-microsoft/model-calibration-for-classification-tasks-using-python-1a7093b57a46