Project 1 – Machine learning applied to fiber-optic sensors

Predict sensor response and evaluate sensor performance

1.1 General background

What is a fiber-optic sensor?

A fiber-optic sensor is a device that is produced by inducing mechanical or optical changes in an optic fiber. The sensor that has been used to explore machine leaning in this project is called a Mach-Zehnder interferometer. A schematic of the sensor is shown in Figure 1.

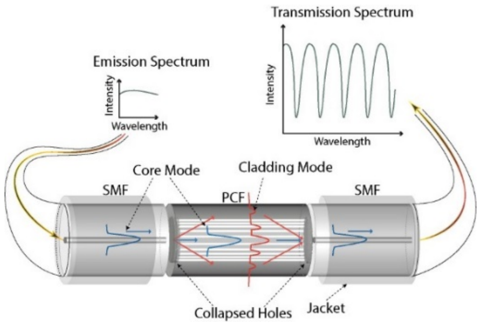

Figure 1. Fiber-optic sensor based on a Mach-Zehnder interferometer

The details of the fabrication process and the working mechanism of this sensors are described in a scientific article in which I am first author[1]. The output of the sensor, which have a sinusoidal pattern (see transmission spectrum in Figure 1), is dependent on three physical parameters: temperature, strain and the refractive index that surrounds the sensor. Changes in any of these physical parameters cause horizontal shifts in the transmission spectrum, and can be quantified by measuring the wavelength shift.

1.2 Compelling case

What are we measuring?

This sensor was investigated in my research to measure CO_2 based on the **refractive index** of the media that surrounds the sensor. In the case of the data presented here, the sensor was applied to differentiate between water (refractive index = 1.33) and liquid CO_2 (refractive index = 1.18) at high pressure. Regarding machine learning, this is double class classification problem, *i.e.* CO_2 versus water.

Current challenge

The approach that have been used in the literature to take measurements of chemical species with fiber-optic sensors can be summarized in two steps. Step 1 - The sensor is calibrated with commercially available solutions with known refractive indices and a calibration curve of wavelength as a function of refractive index is calculated. Step 2 – The sensor is exposed to the testing solutions (*i.e.* CO_2 or water) and the refractive index of each solution is inferred based on the wavelength shift and the results from the calibration curve. However, sensor calibration is time consuming and increases the risk of damaging the sensor.

1.3 Objective

The overall objective of this project is to develop an approach based on machine learning which uses the data of the CO_2/water experiment to train the sensor and predict new data. The main goal is to avoid the need to use the calibration step (*i.e.* step 1 described in Section 1.2). The second objective is to identify sensor damage using machine learning algorithms.

[1] L. Melo, *et al.*, Sensors and Actuators B: Chemical, 236, 537–545, 2016

1.4 Project methodology and approach

A new machine learning approach is investigated which combines **regression** and **classification**. Regression is used as a **quality measure** whereas classification is used as a **predictive measure**. The selected algorithm for regression is the **simple linear regression** and the selected algorithm for classification is the ***K*-Nearest Neighbor**, with $K = 5$. The project is developed in **python** and the machine learning algorithms are applied using the open source library **sciKit-learn**.

The algorithms are applied to two datasets. The first dataset corresponds to an experiment in which the sensor was stable (*i.e.* **high-quality dataset**). The second dataset corresponds to an experiment in which the sensor performance decreased over time (*i.e.* **low-quality dataset**). In both dataset, the dependent variable is defined by setting the **value of water to 0** and the **value of CO_2 to 1.** A schematic of the split of the training and testing data is shown in Figure 2.

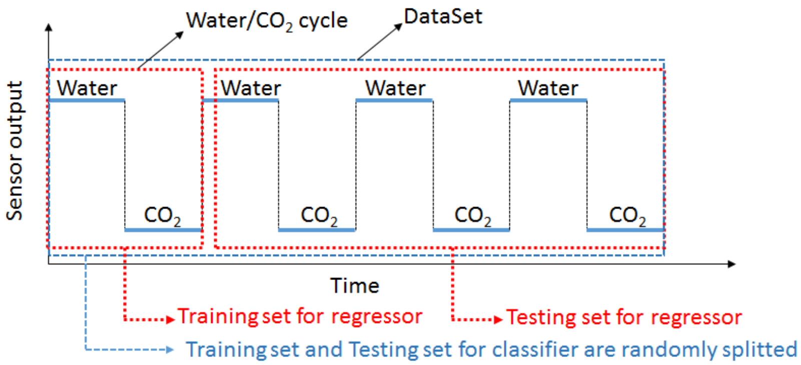
Water/CO_2 cycle
DataSet
Sensor output
Water
Water
Water
Water
CO_2
CO_2
CO_2
CO_2
Time
Training set for regressor
Testing set for regressor
Training set and Testing set for classifier are randomly splitted

Figure 2. Split of train and test data for regression and classification

Dataset	Total features for each dataset			Regression		Classification	
	Number of cycles (i.e. water/CO_2)	Number of Data points	Skewness	Training Set (number)	Testing Set (number)	Training set (number)	Testing set (number)
High quality	10	12159	-0.03	1197	10962	9119	10962
Low quality	4	12237	-0.4	4900	7337	9177	3060

1.4 Results

<u>High-quality dataset</u>

The experimental data obtained with the sensor exposed to 10 cycles of water and CO_2 is shown in Figure 3. It can be observed that response at each level is repeatable which proves the high quality of the dataset.

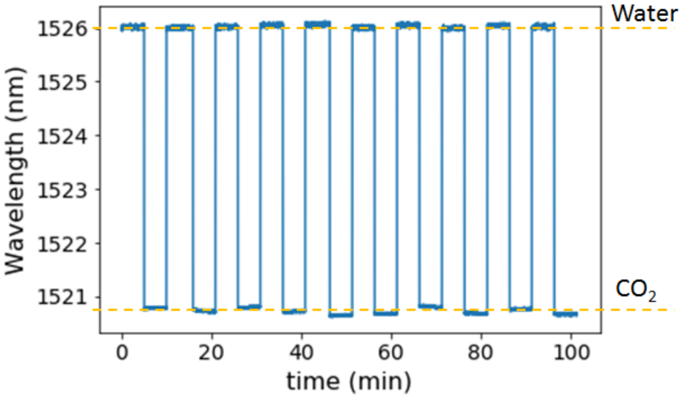

Figure 3. Experimental data obtained with a Mach-Zehnder interferometer sensor exposes to cycles of water and CO_2

Figure 4 (a), shows the results of the fitting curves calculated by simple linear regression using the train dataset (blue marks) and test dataset (red marks). This figure shows that the predicted values are approximately equal to the water and CO_2 values, *i.e.* 0 and 1, respectively. The coefficient of determination (r^2) for the train and test dataset is 0.99996, and 0.99943, respectively. The similarity of r^2 between the train and the test dataset confirms that the response of the sensor is repeatable. Therefore, r^2 in this case is an indicator of the high performance of the sensor.

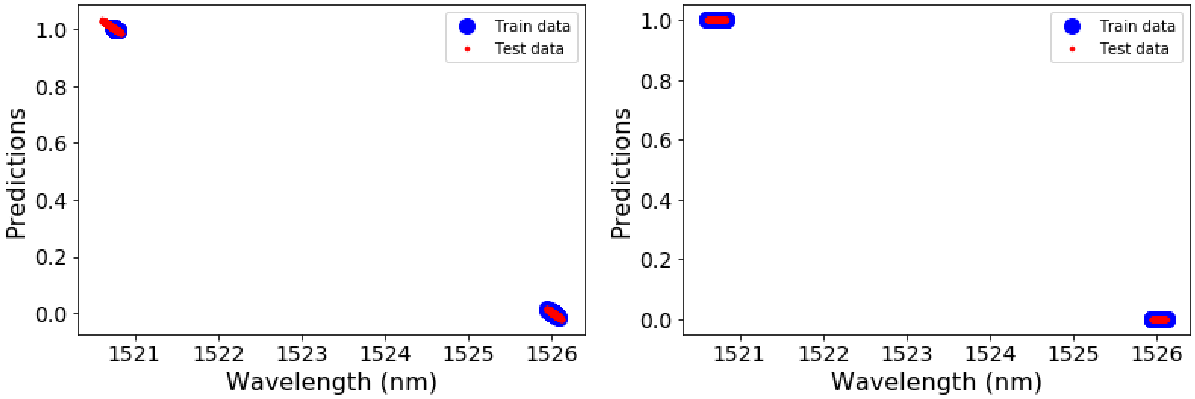

Figure 4 (a). Results of simple linear regression and (b) K-Nearest Neighbor (K=5) applied to the high-quality dataset

Figures 4 (b) shows the results of the *K*-Nearest Neighbor classifier applied to the train and test data. This figure shows predicted values of 0 at wavelength = 1526 nm and 1 at wavelength = 1521 nm, for both train and test data. These results show that *K*-Nearest Neighbor classifier provides accurate predictions of sensor exposure to water or CO_2.

Low-quality dataset

Figure 5 shows the experimental results obtained with a Mach-Zehnder interferometer for the low-quality dataset. The low performance of the sensor is visible by the data drift.

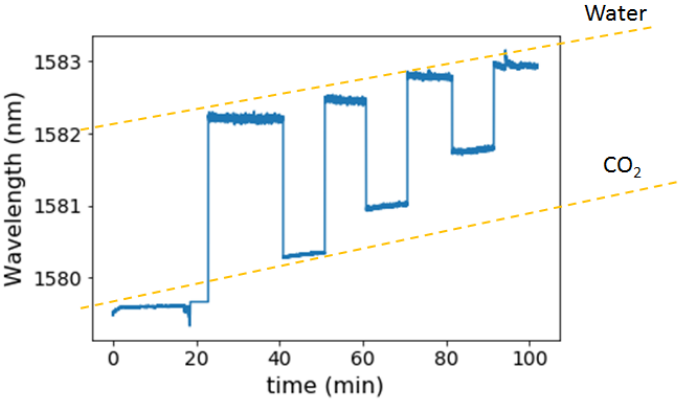

Figure 5. Low quality dataset obtained with a Mach-Zehnder interferometer exposed to cycles of water and CO_2

Figures 6 (a) shows the results of the regression applied to the train and test data. This figure shows that the train data have values approximately equal to 0 or 1 (see blue marks). This is because the wavelength is virtually constant for CO_2 and water in the first cycle, and the dependent variable was set to 0 or 1. For the test data, Figure 6 (a) shows that the data is more disperse. This occurs because the wavelength of each subsequent cycle is not consistent with the first cycle, and thus, the training fit does not perform well with the test data. The r^2 for the train set is 0.9992 and for the test set is 0.2199. This result confirms that the parameter r^2 can be used as a quality measure to evaluate sensor performance.

Figure 6 (b) shows the results of the classifier applied to the train and test data. The results are identical for both data sets. The classifier is able to predict the correct solution to which the sensor is exposed. This result is achieved due to the large number of data points used in this project, and because the full set of data points was randomly divided into train and test set for the *K*-Nearest Neighbor.

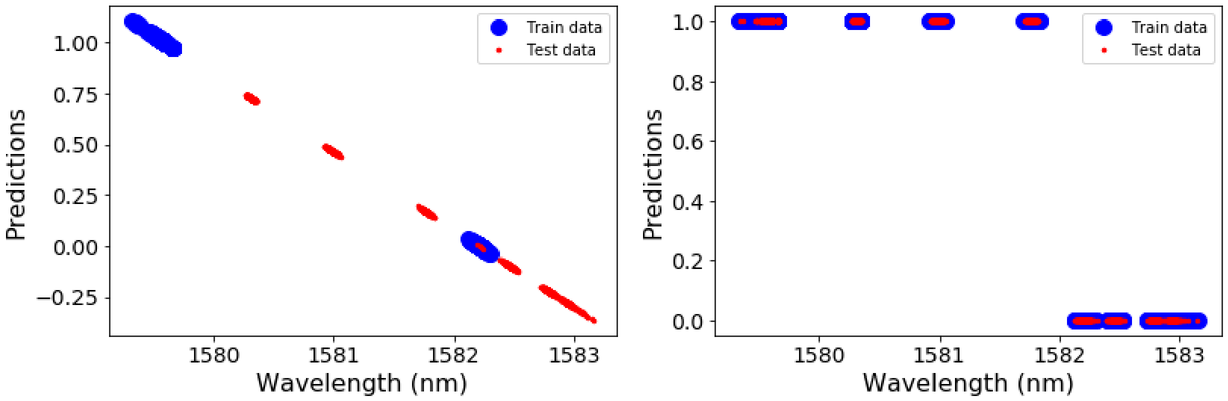

Figure 6. (a) Results of the simple linear regression and (b) K-Nearest Neighbor (K=5) applied to the low-quality dataset

1.5 Final remarks

- Simple linear regression is applied to the dataset and is used has a quality measure using the parameter r^2. If r^2 is close to 1, the sensor is showing the expected behavior, if r^2 decreases for each cycle to which the fitting is applied, the sensor is contaminated or damaged;
- *K*-Nearest Neighbor classifier is applied to the full dataset randomly split into train and test set. This classifier predicts if the sensor is being exposed to CO_2 or water even when the sensor is not operating at optimum conditions.