Forecasting the Unseen: detecting weather sensor faults in wind farms through unsupervised learning

Eng. Máximo Iaconis
Facultad de Ingeniería
Universidad de Palermo
Buenos Aires, Argentina
iaconis@live.com

Tec. Félix Cuello
Facultad de Ingeniería
Universidad de Palermo
Buenos Aires, Argentina
felix.cuello@gmail.com

Keywords—wind energy, sustainable energy, wind turbines, fault detection, anomaly detection, time series analysis, SCADA, machine learning, kmeans, sensor data analysis, weather station monitoring, renewable energy, wind farms

I. Introduction

We live in an era where humans consume more and more energy every year; although fossil fuels have been around for a very long time to satisfy the demand, this causes a big impact on the environment and, sometimes, on human health. Thus, there's an urge to move into clean energy sources.

In the search for sustainable energy solutions, wind energy has become a fundamental resource, which requires advanced technologies for its efficient use and utilization. Even though wind farms provide a cleaner alternative there are also many challenges that we need to overcome with this, such as noise, impact on wildlife, intermittency, and land and resource use.

Optimizing the performance of wind turbines is closely linked to the precision and reliability of meteorological measurements obtained from meteorological masts. These structures are equipped with a series of sensors that monitor the atmospheric conditions, providing critical data for the evaluation of wind potential and operational control of the turbines.

The harsh and often remote environments in which these instruments operate make them susceptible to failure. Therefore, fault detection in weather sensors is of utmost importance to maintain data integrity and ensure the continuous operation of wind energy systems.

In this paper we delve into methodologies to identify failures in an automated way.

II. Wind Farms Objective

Wind farms' primary objective is to harness wind energy and it's imperative that alongside this energy generation, we constantly monitor weather conditions around these wind turbines.

For this purpose, masts equipped with weather sensor devices are strategically located around the turbines, recording meteorological parameters such as: air pressure, air direction, air speed and other measurements non related to weather to monitor the health of the whole mast).

This data is recorded in time series.

III. Time Series

The data for these sensors are captured in time series. time, which is nothing more than a sequence of data points collected in time at uniform intervals in order to monitor the equipment and conditions.

They play a critical role in maintaining efficiency, making smarter decisions, and communicating system issues to help mitigate downtime.

To gather and make these time series available through networks, Supervisory Control and Data Acquisition (SCADA) is often used.

Over time we can identify:

- **Forecast Events**: By analyzing historical patterns you can possibly predict how values are going to move in the future
- **Seasonality**: By analyzing patterns, we can associate data with events making it easier to predict.
- **Trend Identification**: By monitoring the historical values it's possible to predict and get trends of data to identify how the data is going to move next.
- **Monitor and control**: The time series could help to take real-time decisions when the data is showing a definitive failure.

IV. Monitor and Control

Since these devices are located in remote locations the data gathered over time is stored in the SCADA and can be accessed remotely by operators that download the data in different file formats.

However, downloading and analyzing the data from each weather mast requires human intervention. Which often requires analyzing a huge amount of data, which is tedious and slow.

There are already systems that can do health check based on metrics, but we wanted to develop something general that can suit our needs to analyze the data we gathered from our systems.

V. Data Analysis

The first step before digging into the problem is to analyze some representative time series data that can show anomalies and let us create a baseline. This will help us to identify any unusual patterns in the data that may be indicative of a problem.

In this case, the time series was collected from anemometers, which are devices that measure wind speed over time (roughly every 10 minutes).

The data shows (**Fig 1**) that there were some clear outliers in which the data points are significantly different from the rest of the data caused by faulty sensors. These sensors are usually damaged as a result of severe weather conditions or just malfunction. The anomalies are often caused by external factors and show that a particular value is not behaving properly like the others.

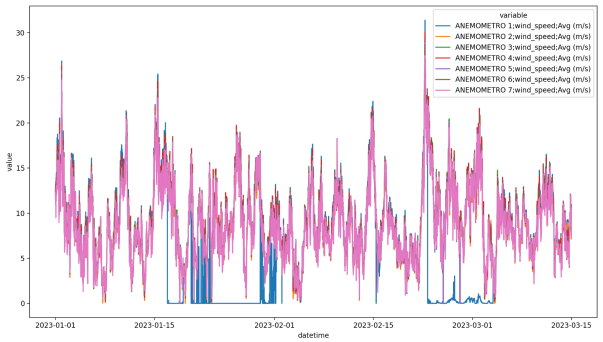

Fig. 1. Plot of 3 months of anemometers data

Anemometer #1 (**blue**) suffered two major outages in this time frame which become evident when you plot the data in a time series graphic.

VI. Possible Algorithms

When using machine learning we can use either supervised or unsupervised models.

Given the good amount of data, we were in a good position to use an unsupervised model to detect outliers in data that could tell us that we have a faulty sensor.

VII. Unsupervised Algorithms

A. *Isolation Forest*

Isolation forest is a very popular unsupervised machine learning algorithm to detect anomalies in data.

This method selects a feature and makes a random split in the data between minimum and maximum values creating decision trees (that's why it's called a forest).

The process is repeated several times creating trees that isolate with paths to isolate the observations.

We can take the path lengths of these trees as the measure of anomalies, identifying that shorter paths are anomalies in the data.

In the graphic (**Fig. 2**) we have tried to use Isolation Forest in order to detect anomalies in data. We changed the hyper-parameters of the model without getting any significant improvement to the result.

The model seems not to detect the anomalies accurately, since the plot of the Anemometer #1 (**blue**) and the anomaly detection using isolation forest (**red**) doesn't seem to correlate.

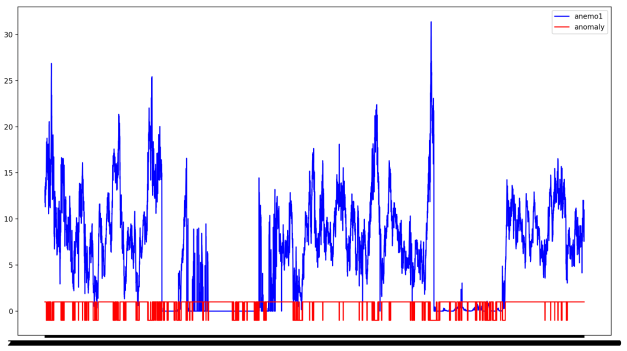

Fig. 2. Isolation Forest

B. *OneClass Support Vector Machines (OneClass SVM)*

1) Using the **average** of the data points

Another type of outlier detection method is One Class Support Vector Machines.

This model instead of labeling the data, tries to learn the boundaries of the normal data points trying to identify what could be the anomalies.

This is basically the same type of separation we do when we see a graphic and we quickly (and visually) identify where the data is in a graphic.

We tried then to add a new column called average which was the average of all the data points and tell the system that we wanted to predict the average for the data.

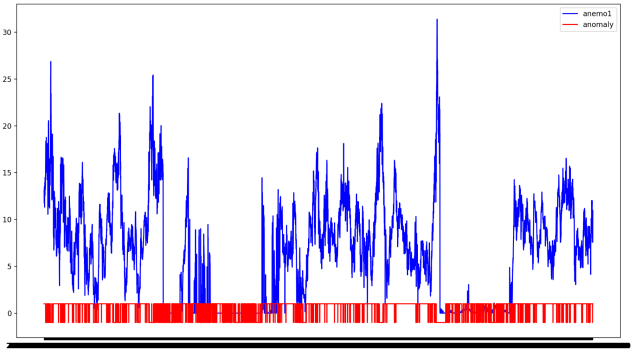

Fig. 3. OneClass SVM [calculated with average]

When plotting the results (**Fig. 3**) it becomes clear that there's no correlation between the Anemometer #1 graphic (**blue**) and the outlier detection by OneClass SVM (**red**).

2) Delta of sensors

We introduced a new calculation to include other sensors' data.

Since we already know that not all the sensors are faulty at the same time, they must behave similarly creating a cluster, therefore we introduced a new column called **difference** with the following calculation:

$$\text{Eq. (1)} \quad diff = 2 * s[i] - min(s[j])$$

a. i = sensor target / j = any but the sensor target

Which resulted in the (**Fig. 4**) which appears to be more accurate since there's a slight better correlation between the anomaly shown by Anemometer #1 (**blue**) and the anomaly detection (**red**). However this is still inaccurate to use in a real environment.

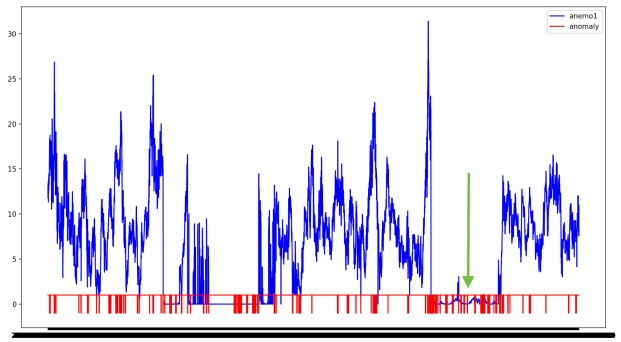

Fig. 4. OneClass SVM [calculated with difference]

After this, we decided to take a different approach because perhaps we were using a sophisticated approach for a problem that could be solved in simpler terms.

VIII. RMSE As Estimator of Anomaly

We still had the target to estimate the points of the time series data where we have anomalies when a sensor failed.

For this purpose we have used an RMSE (Root Mean Squared Error) approach to tag each point.

We define $aa[n]$ as the average of the *n-th* point of the time series dataset.

We define $anemo[n][i]$ as the *i-th* anemometer in the *n-th* point of the time series dataset.

We define the Root Mean Squared Error (**Eq. 2**) as follows:

$$\text{Eq. (2)}$$

$$RMSE = \sqrt{\sum^{anemometers} (anemo[n][i] - aa[n])^2}$$

We empirically determined that if $RMSE \geq 2$ we were in a point of the dataset that we had an anomaly.

We can see a difference between a dataset with **no anomalies** (Table 1) and a dataset **with anomalies** (marked in the sample in red color) where the RMSE behaves how we expected.

TABLE I. No Anomalies

n-th sample: [10, 11.5, 12, 14, 10.7, 9.8, 11.9, 13]		
aa	RMSE	anomaly
11.6125	1.3504	**false**

TABLE II. Anomalies

n-th sample: [10, 11.5, 12, 14, 10.7, 9.8, 11.9, **6**]		
aa	RMSE	anomaly
10.7375	2.1805	**true**

This *ad-hoc* method detected the anomalies correctly, and we can use it to tag our data as **correct** or **faulty** to apply any other machine learning algorithm.

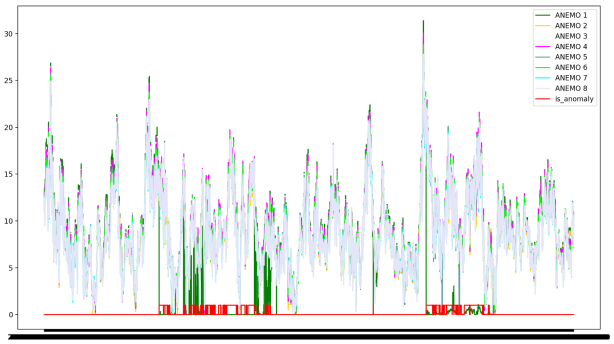

Fig. 5. RMSE as estimator

We can clearly see (**Fig. 5**) that this method provides a very accurate way to decide if a sensor is faulty or not, and also provides a way to tag our data as faulty or not faulty.

Unlike other examples the correlation between the Anemometer #1 (**green**) and the fault detector (**red**) are very strong and clear.

IX. K-means

K-means is an unsupervised machine learning algorithm used for clustering similar data points into groups based on their features.

The algorithm works by initializing "k" centroids, one for each cluster, and then assigning each data point to the nearest centroid.

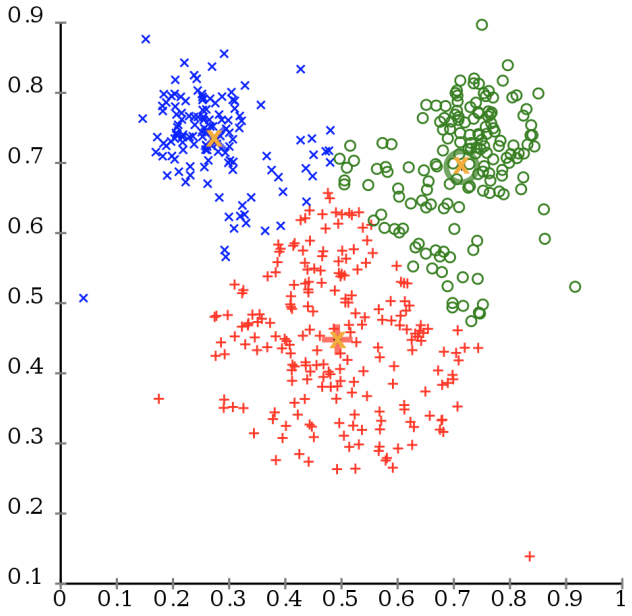

Fig. 6. K-means clusters with its centroids

The centroids are then re-calculated as the mean of all points in the cluster, and the process is repeated until the centroids no longer change significantly.

This feels pretty similar to what we were doing but it's also encapsulated in a ML algorithm, which can be less prone to error.

The K-means algorithm is computationally efficient and easy to implement but has some limitations because it is sensitive to the initial placement of centroids, which can result in different clustering outcomes.

Additionally, the algorithm assumes that clusters are spherical and equally sized, which may not always be the case in real-world data. Fine-tuning the number of clusters "k" is often necessary to achieve meaningful results.

X. Deciding the K value for K-means

In the previous section we stated that finding the right value for K (number of clusters) is crucial to get good results for this ML algorithm.

If we think about the question "What is the right k for this problem?" We can think that k=2 would be a good candidate, because what we want to see is an outlier, an element that it's moving out from the rest of the cluster. Let's take a look to an example (**Fig. 7**):

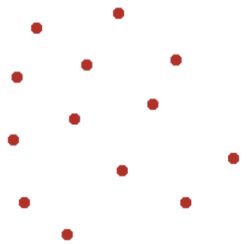

Fig. 7. Sample Cluster of points

If we have k=2 then the algorithm is always going to have two clusters with its centroid, even if there's no two different clusters, and literally anything could be the two clusters:

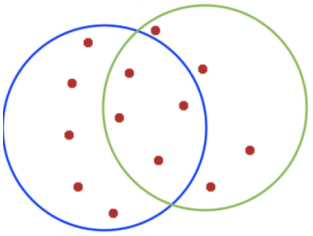

Fig. 8. K=2 in a dataset that doesn't have clusters

What if we set the K=1?

On one hand, if we pick **k=1** we are saying to the ML algorithm that we have only one cluster, and we want to find the centroid for our whole dataset.

On the other hand, picking k=1 under the premise that k-means would detect outliers that were outside the cluster worked surprisingly well (**Fig. 9**).

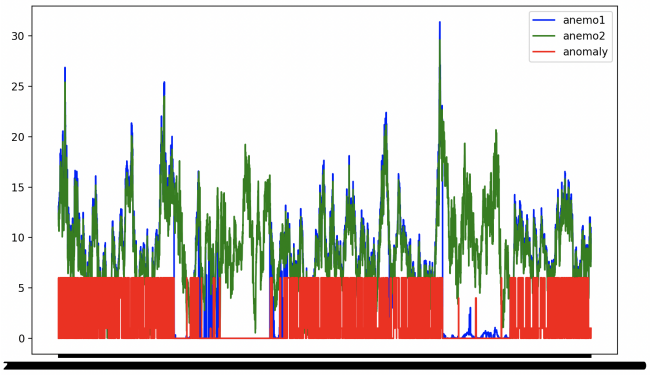

Fig. 9. K-means with k=1

If we compare the RMSE anomaly detection against the K-Means with K=1 we see the results are nearly identical (**Fig. 10**), besides the fact that the latter method not only detects that there's a failure but **also** identifies which device that is faulty making it more useful in the field.

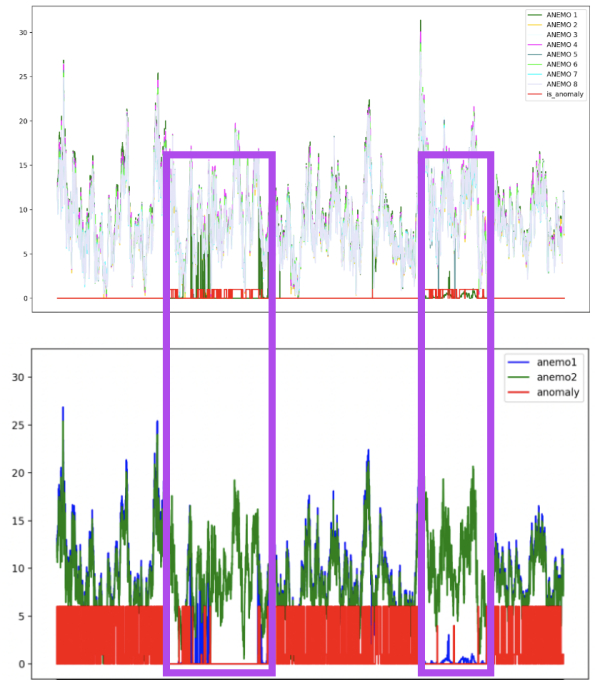

Fig. 10. Comparison between RMSE and K-means with k=1 methods

XI. Conclusion

The process of monitoring and diagnosing faults in weather stations it's a task that creates numerous challenges due to the volume and complexity of the data involved.

We had some initial setbacks with unsupervised models which failed to detect anomalies with accuracy.

This idea of exploiting the inherent redundancy of sensor systems in the weather mast by constructing an anomaly estimator which let us first create a good RMSE estimator, and later the use of K-means to identify the faulty sensor.

This system can be used already in production as SaaS. However there is a lot of room for improvement.

The most important improvement to tackle is to introduce a **prediction** algorithm. By using artificial intelligence we estimate that it's possible to **predict** anomalies before they occur.

References

[1] Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat

[2] Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat

[3] Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat

[4] Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat

[5] Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat

[6] Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat

[7] Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat