STORY OF SOLAR PANEL NO. 1.6.7 – PRACTICAL APPROACH TO ANOMALY DETECTION

Marek Wadinger

Faculty of Chemical and Food Technology STU in Bratislava

xwadinger@stuba.sk

Introduction

The main aim of this paper is to introduce an algorithm for automatic anomaly detection on device based on its measurements. The algorithm recognizes deviations from normal functioning as a result of degradation or novel patterns of behavior. Degradation of the modules can be both temporary due to shading, dirt, or snow and permanent considering hardware faults or aging. The main benefit of this algorithm is that it enables offline training on unlabeled data without previous knowledge of the nature of anomalies, unlike approaches with analyst in the loop. Furthermore, by removing the assumptions, like normal distribution of data and exact knowledge of the number of anomalies, we created a practical tool that is easy to use and suitable for deployment in a practical setting. The automatic diagnostic of the device has crucial role in timely maintenance planning and minimization of economic impact of malfunctioning. The algorithm is based on an ensemble of unsupervised classification models.

Implementation

The objective was to develop, deploy, and scale an algorithm that would label the most anomalous observations of the end-of-day production of photovoltaic modules. The time series data were queried from the database and missing values handled. In feature engineering, we extracted useful information about the entities from data to improve the performance of the algorithm. The data were split into the train set and the test set. In unsupervised learning, this step could be omitted, as there are no preexistent labels to be used for performance validation on test data. However, it is a good practice to evaluate the algorithm on data that were not seen during its training. For the anomaly detection, the statistical approach using machine learning was employed. Among all tunable parameters of the models, the most relevant was contamination rate, for its great influence on overall performance. Contamination estimator was used to update the outliers rate of detection models by means of the Tukey's method. Ensemble of models was trained based on the density-based K-Nearest Neighbors algorithm. and further used to evaluate the data in the test set, returning outlier scores (extend to which was the observation considered anomalous/novel) and binary labels (whether the outlier score was over threshold) and exported to a file for deployment. The results were interpreted by visualizing the time evolution of most anomalous entities and using a console printed report for summarization.

Experimental Part

For the algorithm development, we used Python, a widely supported high-level programming language [1]. Data were queried from InfluxDB database using API [2]. For data manipulation and handling, we employed the package for scientific computing numpy and the data analysis and manipulation tool pandas [3, 4]. Detection algorithms were covered in a toolkit for outlying object detection pyod and exported using tools included in the joblib package [5, 6].

Results and Discussion

The algorithm was successfully deployed on a group of photovoltaic modules. Without previous knowledge of anomalous patterns, the algorithm not only marked anomalies on the modules but also pointed out on anomalous behavior of Module 1.6.7 since 5th August 2021 - one month before the module raised suspicion after manual evaluation. The algorithm is of linear time complexity O(n) which means that the time grows linearly to the size of the input data set.

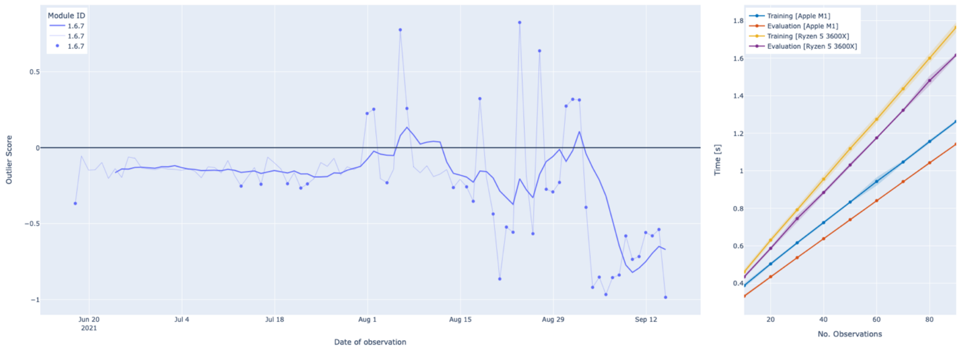

Figure 1. Left: Time series of outlier scores of modules with $n_{outliers} > 15$ (thick line: 7-day moving average; thin line: actual values; point: outlier labeled). Right: Avg. running time of training and evaluation of 50 runs on a single CPU core. No. observations are given for each of the 155 modules.

Conclusion

The developed algorithm proved its potential to automatically detect deviations from the expected behavior in early stages. The algorithm for automatic anomaly detection could be further used not only for maintenance planning but also as an input for the photovoltaic prediction algorithm. For internal operators of machinery and plants that perform tasks requiring extensive visual concentration, the algorithm could help to pinpoint information about novel or anomalous patterns that might be of great importance as visual fatigue and tiredness arises.

Acknowledgement

Special thanks to doc. Ing. Michal Kvasnica, PhD. for consultations during the project realization.

Literature

[1] van Rossum, G. Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam. 1995.

[2] Influxdb_client. Available at: https://github.com/influxdata/influxdb-client-python.git.

[3] Harris, C.R.; Millman, K.J.; van der Walt, S.J. et al. Array programming with NumPy. Nature. 2020, 585, 357–362

[4] McKinney, W. et al. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference. 2010, 445, 51-56.

[5] Zhao, Y.; Nasrullah, Z.; Li, Z. PyOD: A Python Toolbox for Scalable Outlier Detection. Journal of machine learning research (JMLR). 2019, 20(96), 1-7.

[6] Joblib Development Team. Joblib: running Python functions as pipeline jobs. 2020.