

Multivariate Anomaly Detection^{*}

William Chavarría^{a,b,*}, André Rodas^b, Pablo Armas^b and Francisco Manjon^b

^aTigo SRT, Avenida Petapa 24-50 z.12, Guatemala, 01012, Guatemala

^bTigo, Km 9.5 Ctra a El Salvador, Santa Catarina Pinula, 01051, Guatemala, Guatemala

ARTICLE INFO

Keywords:

collective anomalies

multivariate

IoT

ABSTRACT

Today, with the ubiquity of IoT, it is increasingly common for telecommunications companies to adopt and leverage these systems to improve the availability of their services. In this study, we investigated the possibility of using algorithms different from the current ones to enhance and better generalize the detection of anomalous cases in the temperature of the most important sites in the transmission network. We developed an unsupervised model to investigate its applicability, considering the use case and the current context of readings, behavior, and stability of measurements. Using an Isolation Forest model, we adjusted a model and achieved an f1-score of 84% in anomaly detection. The results obtained support the hypothesis on the effectiveness of conventional models, opening up a wide range of possibilities for expanding and generalizing to the rest of the sites. The prototype was packaged in an environment that allows for the evaluation and simulation of the algorithm's production deployment.

1. Introduction

The potential economic value that IoT could unlock is large and growing. It's estimated that by 2030, IoT could enable between \$5.5 and \$12.6 trillion in global value, including value captured by consumers and customers of IoT products and services (Chui and Collins, 2021). According to a report by McKinsey & Company, building an initial set of predictive models to extend the lifespan of an organization's assets is a significant investment and just the first step in an ongoing process of refinement and continuous improvement. Post-implementation, companies often seek to improve three aspects of their Predictive Maintenance (PdM) systems Cortez and Guillaume (2021).

Using IoT hardware, the aim is to identify anomalous behaviors across several dozen sites equipped with sensors measuring temperature, humidity, and current. These metrics are pivotal for detecting potential failures in industrial air handling units (HVAC) that cool the sites. Early anomaly detection in these metrics can prevent costly failures and revenue loss for the company.


Currently, an unsupervised algorithm named 'capa' is employed for its ability to detect collective anomalies, its computational efficiency, and its capability to work with multiple variables simultaneously. However, challenges are faced in standardizing the algorithm's hyperparameters across all sites, leading to false positives in some locations.


1.1. Motivation of this research

In 2017, due to a failure in one of the air conditioning units at the site called El Rodeo, equipment valued at over \$500K was damaged. Although the losses were substantial in terms of equipment and electronics, they paled in comparison to the revenue losses associated with service downtime and brand perception. A standalone IoT system with fixed thresholds is incapable of detecting and identifying early signs of operational degradation. An algorithm in production capable of detecting symptoms of an impending air conditioning unit failure at least two hours in advance is crucial for our operations.

^{*}This document is the result of the research conducted for the final project of the Data Products course.

^{*}Corresponding author

 wchavarria@tigo.com.gt (W. Chavarría)

 cha18982@uvg.edu.gt (W. Chavarría)

ORCID(s): 0009-0001-4084-9290 (W. Chavarría)

The objective of this document is to verify whether an anomaly detection algorithm, which operates by separating data points in a multidimensional space to identify outliers, is viable given the temperature conditions at our sites.

1.2. Problem description

Each site has two air conditioning units that operate in alternating cycles. One unit runs for a fixed period, and then the other takes over, causing a sudden change in temperature during the transition. This behavior is normal and should not be considered an anomaly. The current system detects it, and we have no way to filter it out. The time window for analysis can be extended, but this comes at the cost of increased computational load on the infrastructure running the algorithms. Changes in temperature due to external factors should also be considered normal.

Sites exhibit variable conditions that complicate the task of anomaly detection. For instance, although HVAC units are expected to receive regular maintenance, in practice, this maintenance is often minimal. Additionally, different brands and models of units are used across sites, some of which have outlived their expected lifespan. Other factors, such as the lack of airtightness in some sites and variability in maintenance practices, introduce additional noise into the data, challenging the efficacy of anomaly detection.

2. Methodology

Data were extracted from the IoT system's database, which records data at quasi-regular intervals. The site 'PARROQUIA' was selected with a one-week window, leaving anomalies that impacted at the end of the series. An SME manually labeled the actual anomalies in a column named truth. Subsequently, the data were transformed to meet the format requirements of the selected algorithm, which was Isolation Forest. A Jupyter Notebook was created, and a model was fitted. The efficacy of the model was visually verified, and metrics were determined, focusing on the macro-average F1-score

3. Results

We observed approximately 15 false positives in Figure 2, and the majority of anomalies were identified at the end of the series. The confusion matrix obtained is presented below

Multivariate Anomaly Detection

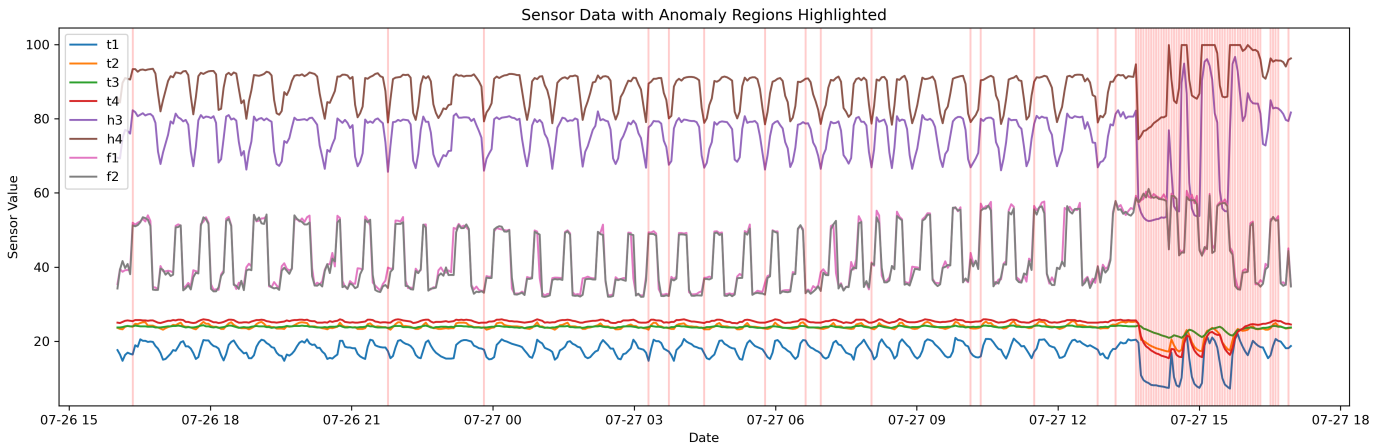


Figure 2: Isolation Forest

so that when they fall outside of some predefined range, they are classified as anomalies. Ball (2023)

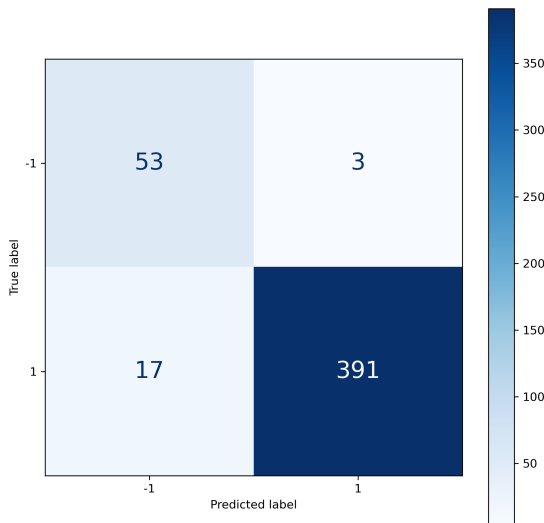


Figure 1: Isolation Forest

4. Discussion

The algorithm was only tested on a single site and within a very specific time frame with very clear anomalies. Although this is not sufficient to confidently assert its effectiveness, it does fulfill the objective of obtaining promising results that will allow for the initiation of more exhaustive tests with different sites and different types of anomalies, including collective ones. One of the main issues with Isolation Forest is the *contamination* parameter, which must be selected by the user and will depend on how much prior information we have about the number of anomalies in the series. A more sophisticated and appropriate approach would be to treat these cases as a time series forecasting problem and then apply the anomaly detection function, first decomposing the time series into its various components (seasonality, trend, residual) using something like statistical models in Python. The residual component will represent deviations from the expected seasonal pattern. What this would lead us to do then is to take those residuals and apply a threshold,

5. Conclusions

We were able to confirm our hypothesis that not only algorithms that detect collective anomalies can be useful for our use case. While collective anomalies often serve as a preamble or precursor to a critical failure, especially when both air conditioning units are operating simultaneously, it doesn't necessarily mean that algorithms need to detect them. Generally, there will be a gradual period of temperature increase, which is consistent with what most anomaly detection models do. It is highly likely that we will mostly rely on unsupervised algorithms, given that we do not have labels to help us determine whether anomalies occurred or not. More sophisticated algorithms like auto-encoders are currently ruled out due to the potential computational load they could represent. It is expected that in a short time, we will have the capability to deploy the best algorithm and its instances to each site, given the new understanding acquired related to production deployment and MLOps methodology.

References

- Ball, R., 2023. Queries on anomaly detection via email.
- Chui, M., Collins, M., 2021. The Internet of Things: Catching up to an accelerating opportunity. Technical Report. McKinsey & Company. URL: <https://mck.co/3DGLzvp>.
- Cortez, T., Guillaume, D., 2021. Prediction at scale: How industry can get more value out of maintenance. Operations Practice. McKinsey & Company. URL: <https://mck.co/3Bum39X>.