

Data Center IT Anomaly Prediction and Classification: an INFN CNAF experience (Remote Presentation)

Wednesday, March 27, 2024 11:00 AM (30 minutes)

INFN CNAF data center provides a huge amount of heterogeneous data through the adoption of dedicated monitoring systems. Having to provide a 24/7 availability, it has started to assess artificial intelligence solutions to detect anomalies aimed to predict possible failures.

In this study, the main goal is to define an artificial intelligence framework able to classify and predict anomalies in time series data obtained from different sensors and systems within the data center (i.e. electrical plant, cooling system, UPS system, and others). The framework takes into consideration the following data characteristics: the majority of the collected data has a time window that begins on January 6, 2022, and ends on July 7, 2023; the number of entries per file varies from 5000 up to 50000; most of the sensors values are sampled every 15 minutes, but some sensors (like the UPS system) are sampled every 10 minutes; during the merging phase of sensors data, using the timestamp as key, a tolerance of 15 minutes to get entries where every timestamp have the values of each sensor.

Having to deal with unlabeled data, the proposed framework performs as a first step a regression task to learn the behavior of the sensors and, given the previous 5 timestamps, provides the values of the sensors in the next timestamp. As a second step, it performs a classification task. Comparing the predicted and the actual behaviors of the sensors, in fact, evaluates the status of the system and possible anomalies.

The regression task can detect the relationship with other sensors with the usage of GATv2, Long-short term memory, and linear layers, and provide the trend of each sensor by using LSTM layers. To make the training phase faster and less influenced by the initial random initialization of the parameters, batch normalization has been performed after each GATv2 layer and LSTM layer. Once the regression network has provided the expected behavior of the sensors, the outcome is compared with the observed one by using two linear layers: after the first one, there is a batch normalization and a ReLU, providing two numbers between 0 and 1, that represents the probability that the timestamp is anomalous and the probability that it is not anomalous. The two layers have been trained with the mean squared error and the cross entropy as loss functions respectively. The network can properly learn from this unbalanced dataset.

To train the regression model we have only used the non-anomalous timestamps, instead to train the classifier we have considered both types of entries. To avoid missing data between timestamps, the samples for training, validation, and testing have not been created randomly, furthermore, the ratio between non-anomalous and anomalous timestamps has been preserved in all the sets. So to achieve all these points, the dataset has been divided into six different parts of equal length, and then the training set, the validation set, and the test set of each part have been created.

Primary authors: Mr TORZI, Luca (INFN CNAF); RONCHIERI, Elisabetta (INFN CNAF); GIOMMI, Luca (INFN CNAF); Dr COSTANTINI, Alessandro (INFN CNAF)

Presenter: RONCHIERI, Elisabetta (INFN CNAF)

Session Classification: Artificial Intelligence (AI)

Track Classification: Track 10: Artificial Intelligence (AI)