STI07-ETI-018

Data Quality Control and Validation Techniques in IoT

Jason Kabi[1]

[1]Centre for Data Science and Artificial Intelligence, Dedan Kimathi University of Technology. P.O. BOX - PRIVATE BAG – 10143, Dedan Kimathi - Nyeri, Kenya.

Email: jason.kabi@dkut.ac.ke

Abstract

Recently, the utilization of IoT systems in data gathering has greatly increased. This can be credited to factors such as low cost in the establishment and maintenance of the said systems, demand for machine learning modelling data, and also the fact that automation in data gathering can be realized. Broadly, an IoT system can be divided into 3 layers, the perception (data collection/ monitoring) layer which include actuators and sensors nodes such as temperature sensors used in data gathering, the network layer involving network servers and wireless networks which facilitate data transmission and storage and the application layer where a user can interact with the data. The success of all monitoring practices is highly dependent on the proper operation of the sensor nodes in the IoT perception layer. Since sensing elements are fragile and prone to damage which leads to malfunction, there are always anomalous data points in the data collected. The presence of outliers in raw data raises the need to ensure high quality data output from the sensor nodes. Sensor nodes generate large volumes of data hence

the data quality control methods utilized have to be automated extensible and quick enough for real time use. Outlier detection is one of the operations which fall under the quality control category. It is a widely studied area in machine learning and data acquisition. Nowadays, it is being utilized extensively in areas such as IoT. This work considers anomaly detection in time series sensor node data. The focus is on the performance-evaluation of various unsupervised classical machine learning algorithms such as Kernel Density Estimation in time series outlier detection. The aim is to test the robustness of known classical models which act as baselines in anomaly detection. IoT offers flexibility for various anomalies detection algorithms to be tested since the data collected is voluminous and the types of anomalies found are diverse. By deploying fine-tuned, long-established models, researchers can improve on the quality of the data they release from or use in various studies. This work also provides an insight into how time series data properties such as non-stationarity can affect anomaly detection and how operations such as windowing can be used to mitigate the effects and achieve desirable results. The experiments done show that, with some fine-tuning and data pre-processing, classical outlier detection methods' performance can be enhanced and utilized in IoT data quality control. This work also considers IoT data validation as a crucial step in building the needed confidence around data generated by IoT devices. When a temperature sensor deployed in the wild is generating data, a validator for the sensor is needed. The validator can be a separate dataset generated by standard instruments or manual records taken using a standard instrument. The validators instils confidence into data users interested in utilizing the data in different fields of study.

Keywords: Outlier Detection, Sensor Nodes, Time Series, Internet of Things,