ONLINE TIME SERIES SENSOR DATA CLEANING SYSTEM: A CASE STUDY IN WATER QUALITY

Yi-Fan Zhang, Peter Thorburn and Peter Fitch

INTRODUCTION

Water quality high-frequency monitoring offers a comprehensive and improved insight into the temporal and spatial variability of the target ecosystem. However, most monitoring system lacks the consideration of sensor data quality control. The sensor data missing, background noises and signal interference have long been a huge obstacle for the users in understanding and analysing the sensor data, therefore makes the utilisation of sensor data much inefficient.

Therefore, we present an online data cleaning system for water quality sensor data. After collecting the raw sensor data, the data cleaning system applied different data filters to corresponding water quality sensor streams. In this approach, the specific environmental effects and can be considered separately. Cleaned data streams are then sent to the web-based frontend interfaces for end users

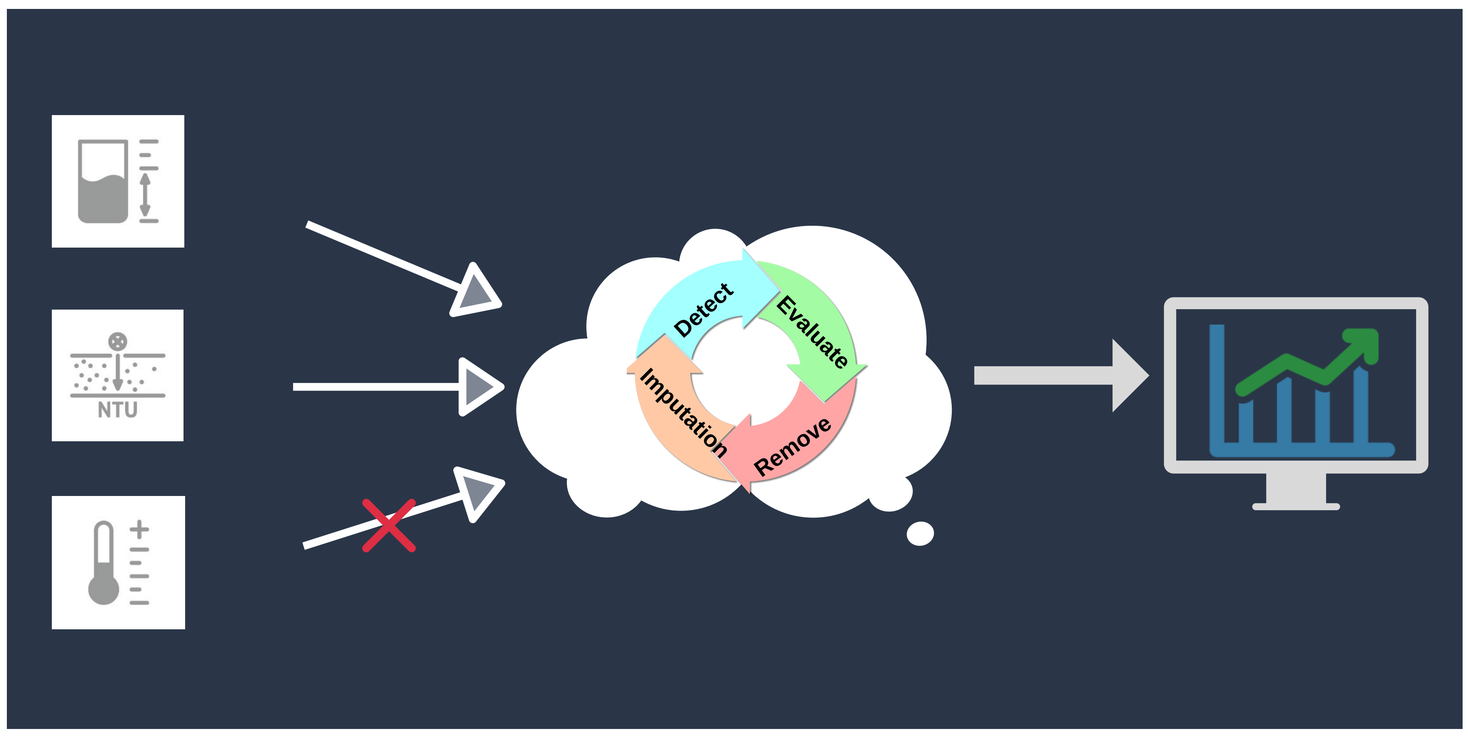

Figure 1 Workflow of the online sensor data cleaning system.

Our sensor data cleaning system uses four steps in solving the above challenging issue: detect, evaluate, remove, and imputation. In the workflow, the system checks incoming data streams in real-time; each new observations are evaluated based on the various filter algorithms we choose. After the detected outliers are removed, the blank gaps are filled by machine learning or statistical-based imputation algorithms. The data cleaning process is executed regularly to ensure that every observation can be processed.

There are two main tasks in this system: detect and remove water quality outliers, and recover the missing sensor data.

OUTLIER DETECTION

For the first task, the water quality filters are built based on the variable-specific threshold, value changing rate, time series moving average and statistical data distributions.

Multiple filters are selected based on the characteristics of the monitoring systems, For instance, the monitoring threshold and changing rate between consecutive timesteps can be set to remove outliers if only the error is obvious. While for filtering algorithms based on moving average or statistical data distribution, by utilising information from numerous previous observations, more outliers can be detected.

DATA IMPUTATION

Most data analytical methods require complete time series sensor data as inputs, and incomplete data can provide biased results. Hence, it is essential to recover missing sensor data before present the data in front of the end users.

In our sensor data cleaning system, instead of only removing outliers from the sensor data streams, the statistical and machine learning-based algorithms such as linear interpolation or sequence to sequence data imputation are applied in filling the sensor data gaps in the monitoring streams.

RESULTS & DISCUSSION

The prototype system is built on AWS Lambda service and eaglo.io environmental IoT platform. The data cleaning process is triggered hourly.

Nitrate sensor data collected from Behana Creek, Cairns is used for testing the sensor data cleaning system. The nitrate data is collected hourly by a in situ monitoring station.

It is illustrated in figure 2 that the range of nitrate concentration is from 0 mg/L to 6 mg/L, which is impossible in the real river ecosystem. Moreover, as shown in the green box, some extreme outliers are existing in the sensor data that is hard to be detected by using the maximum or minimum threshold. End users are hard to learn useful information from data streams with these outliers and errors.

Figure 3 demonstrates the filtered data generated by our sensor data cleaning system. First, the nitrate concentration is rearranged between 0 and 2 mg/L, which is trustworthy in the Cairns sugarcane catchment area. Second, by checking the changing rate and data distribution, the outliers located in the green box have been detected and filtered. Furthermore, all the gaps among the nitrate data stream have been filled.

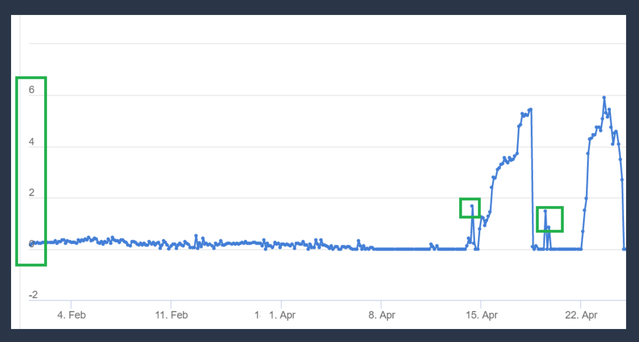

Figure 2 Original sensor data.

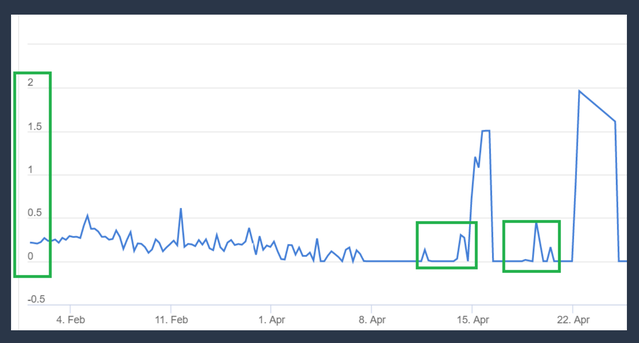

Figure 3 Filtered sensor data.

Yi-Fan Zhang and Peter Thorburn are with Agriculture & Food, CSIRO, Brisbane, QLD, Australia.
Peter Fitch is with Land & Water, CSIRO, Canberra, ACT, Australia
Email: (Yi-Fan.Zhang@csiro.au, Peter.Thorburn@csiro.au, Peter.Fitch@csiro.au)