New Algorithms Improve Water Quality Warnings

Contacts: Shawn Martin, Computer Science and Informatics, *smartin@sandia.gov*
David Hart, National Security Applications, *dbhart@sandia.gov*
William Hart, Discrete Mathematics and Complex Systems, *wehart@sandia.gov*
Sean McKenna, National Security Applications, *samcken@sandia.gov*

CANARY has been tested and successfully detects water quality changes on real-world datasets. … many real-world events produce water quality changes … detected by CANARY, resulting in what we would consider to be false positive identifications. In this project, we have developed alternative algorithms for the elimination of these false positives.

CANARY is a Matlab (www.mathworks.com) toolkit for detecting changes in water quality. It was developed for the U.S. Environmental Protection Agency (EPA) National Homeland Security Research Center Threat Ensemble Vulnerability Assessment program in conjunction with Sandia's Water Sentinel program.

CANARY has been tested and successfully detects water quality changes on real-world datasets. However, many real-world events produce water quality changes that are to be expected, such as addition of CL2 or the merging of different water sources throughout the day. (An example is shown in Figure 1.) These events are of course detected by CANARY, resulting in what we would consider to be false positive identifications. In this project, we have developed alternative algorithms for the elimination of these false positives.

In order to accurately assess our ability to identify false positives, we first generated simulated data using the EDDIES program with the help of Katherine Umberg at the EPA. The EDDIES program was able to generate multiple test data sets by combining data from actual locations in Cincinnati, Ohio. False positive events were added to the data (by EDDIES) in order to create test data for our algorithms.

According to past results [1], the most reliable change detection algorithms in CANARY include the linear filter (LPC) and a multivariate nearest neighbor (MVNN) method. In this work, we implemented the ability to combine these (and other) methods in serial or parallel in order to filter false positives. The serial and parallel modes for LPC and MVNN are illustrated in Figure 2.

We have performed extensive tests using various combinations of algorithms in both serial and parallel mode, using the EDDIES simulation data to quantify our success.

Our results indicate that the serial mode with LPC and MVNN works best for eliminating false positive water quality changes. However, this mode is often too restrictive and may eliminate actual events as well. The general conclusion is that expert knowledge is needed to accurately identify an event as a false positive. This knowledge may be in the form of an event profile that occurs periodically at certain times of the day, or the knowledge that an increase in CL2 will produce a

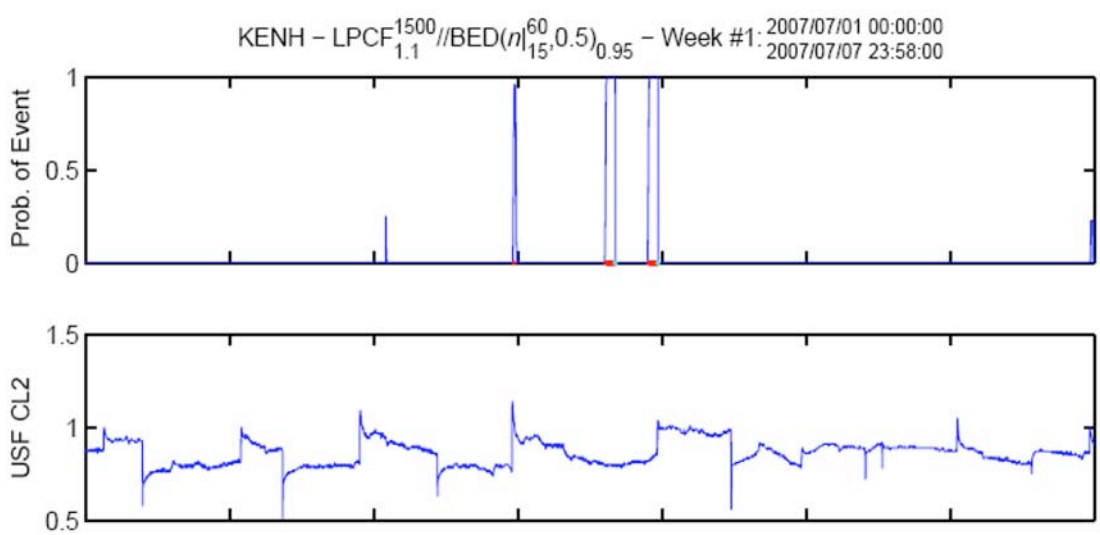

Figure 1. False positive event detected by CANARY. The first large spike in the upper plot is a false positive caused by the addition of CL2, as can be seen in the lower plot at the same time step.

known event. Future versions of CANARY will incorporate the ability to encode this type of expert knowledge.

This work has been supported by the U.S. Environmental Protection Agency.

Reference

1. McKenna, S., Wilson, M., and Klise, K. (2008). Detecting Changes in Water Quality Data, *Journal of the American Water Work Association* 100(1):74-85.

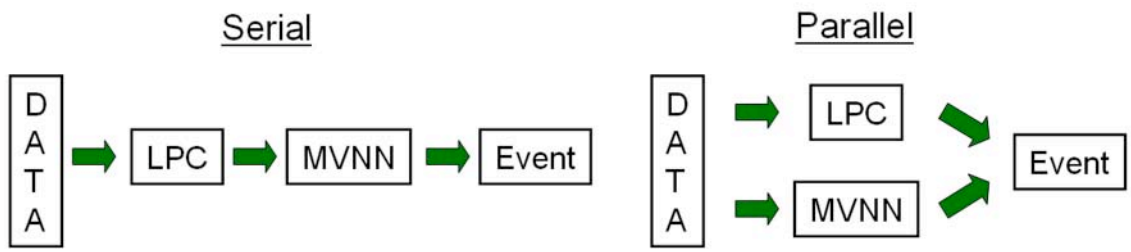

Figure 2. Left: Example of serial mode operation of Canary, using the LPC and MVNN algorithms. Right: Example of parallel mode operation of Canary using the LPC and MVNN algorithms.