

Learning Detectors of Events in Multivariate Time Series

Josep Roure, PhD, Artur Dubrawski, PhD, Jeff Schneider, PhD
The Auton Lab, Carnegie Mellon University, Pittsburgh, PA

Detection of events in multivariate time series is one of the key objectives of applied biomedical informatics. Vigilant monitoring, fast detection and rapid resolution of problems are key factors in effective mitigation of adverse events. Analysts require support of automated surveillance tools to process multiple streams of complex heterogeneous data in a timely and accurate manner. We present a pragmatic approach to constructing such tools. It takes advantage of aggregating evidence from multiple streams of data, it is able to use specific features of outbreaks to detect them faster, and it automates the design process using machine learning.

In the proposed approach, the individual streams of data are monitored with a temporal scan algorithm which tests null hypothesis that the current values of the observed time series do not differ from the baseline. The resulting p-values are then aggregated using e.g. Fisher's method, and the alert is generated whenever any of the streams is out of control, or if some of them are almost critical. This constitutes a basic non-specific multivariate detector. We then add a filter which suppresses alerts if the data characteristics are sufficiently similar to known false alerts. Such a filter can limit false positive rates resulting in higher sensitivity and improved timeliness of detection. We compare a manual method of constructing such filters with two machine learning approaches. They rely on features such as current and recent reads of the single streams' and aggregated p-values. If a substantial number of labeled outbreak examples are available, a classifier can be trained to tell apart false and correct detections. Otherwise, we build in feature space a model of the distribution of known false alerts (and, separately, known correct detections, if they are available). Candidate alerts which likely originate from the false alert distribution (and do not seem to belong to the correct detections distribution, if available) can then be suppressed.

We present results obtained using real-world food safety data in configurations ranging from training sets where experts label negative examples only (data instances which do not correspond to any known event of interest), to training sets where experts identify many known events. We experimentally show that powerful detectors can be learned from a few negative examples, and that the

power of detectors can be improved if more labeled data is available. That is a finding of strong practical significance because labeled data is often difficult and costly to get.