ISDS 2019 Conference Abstracts

An Algorithm for Early Outbreak Detection in Multiple Data Streams

Sesha K. Dassanayake[1], Joshua French[2]

[1]Mathematics and Computer Science, Rhodes College, Memphis, Colorado, United States, [2]University of Colorado Denver, Denver, Colorado, United States

Objective

To propose a computationally simple, fast, and reliable temporal method for early event detection in multiple data streams

Introduction

Current biosurveillance systems run multiple univariate statistical process control (SPC) charts to detect increases in multiple data streams [1]. The method of using multiple univariate SPC charts is easy to implement and easy to interpret. By examining alarms from each control chart, it is easy to identify which data stream is causing the alarm. However, testing multiple data streams simultaneously can lead to multiple testing problems that inflate the combined false alarm probability. Although methods such as the Bonferroni correction can be applied to address the multiple testing problem by lowering the false alarm probability in each control chart, these approaches can be extremely conservative. Biosurveillance systems often make use of variations of popular univariate SPC charts such as the Shewart Chart, the cumulative sum chart (CUSUM), and the exponentially weighted moving average chart (EWMA). In these control charts an alarm is signaled when the charting statistic exceeds a pre-defined control limit. With the standard SPC charts, the false alarm rate is specified using the in-control average run length (ARL_0). If multiple charts are used, the resulting multiple testing problem is often addressed using family-wise error rate (FWER) based methods – that are known to be conservative - for error control. A new temporal method is proposed for early event detection in multiple data streams. The proposed method uses p-values instead of the control limits that are commonly used with standard SPC charts. In addition, the proposed method uses false discovery rate (FDR) for error control over the standard ARL_0 used with conventional SPC charts. With the use of FDR for error control, the proposed method makes use of more powerful and up-to-date procedures for handling the multiple testing problem than FWER-based methods.

Methods

The proposed method can be applied to multiple univariate CUSUM or EWMA control charts. It can also be applied to a variation of the Hotelling T^2 chart which is a common multivariate process monitoring method. The Hotelling T^2 chart is analogous to the Shewart chart. Montgomery et. al [2] proposed a variation of the Hotelling T^2 chart where the T^2 statistic is decomposed into components that reflect the contribution of each data stream. First, a tolerable FDR level specified. Then, at each new time step disease counts from each of the m geographic regions $Y_{1t}, Y_{2t}, \ldots, Y_{mt}$ are collected. These disease counts are used to calculate the charting statistics $S_{1t}, S_{2t}, \ldots, S_{mt}$ for each region. Meanwhile by inspecting historical data from each region, a non-outbreak period is identified. Using data from the non-outbreak period, bootstrap samples are drawn with replacement from each region and charting statistics are calculated. Using the charting statistics, empirical non-outbreak distributions are generated for each region. With the empirical non-outbreak distributions and the current charting statistic for each region $S_{1t}, S_{2t}, \ldots, S_{mt}$, corresponding p-values $p_{1t}, p_{2t}, \ldots, p_{mt}$ are calculated. The multiple testing problem that occurs in comparing multiple p-values simultaneously is handled using the Storey -Tibshirani multiple comparison procedure [3] to signal alarms.

Results

As an illustration, all three methods – EWMA, CUSUM, and Hotelling T^2 (components) - were applied to a data set consisting of weekly disease count data from 16 German federal sates gathered over a 11 year period from 2004-2014. The first two years of data from 2004-2005 were used to calibrate the model. Figure 1 shows the results for the state of Rhineland Palatinate. The three plots in Figure 1 show (a) the weekly disease counts for Rhineland Palatinate (b) the EWMA statistic (shown in red), the CUSUM statistic (shown in dark green) and (c) the component of the Hotelling T^2 statistic corresponding to the illustrated state (shown in blue). The actual outbreak occurred on week 306 (shown by the orange line). Notice the two false alarms – alarms that occur before week 306 - with the Hotelling T^2 statistic (dark green) on weeks 34 and 292; similarly, the CUSUM statistic signals a false alarm on week 57. However, the EWMA statistic does not signal any false alarms before the outbreak (red). Figure 2 zooms on the alarm statistics for the time period from weeks 280 – 330. The Hotelling T^2 statistic misses the onset of actual outbreak on week 306.

The CUSUM statistic detects the outbreak on week 307 – one week later. However, the EWMA statistic detects the outbreak right at the onset on week 306.

Conclusions

Extensive simulation studies were conducted to compare the performance of the three control charts. Performance was compared in terms of (i) speed of detection and (ii) false alarm rates. Simulation results provide convincing evidence that the EWMA and the CUSUM are considerably speedier in detecting outbreaks compared to Hotelling T^2 statistic: compared to the CUSUM, the EWMA is relatively faster. Similarly, the false alarm rates are larger for Hotelling T^2 statistic compared to the EWMA and the CUSUM: false alarms are rare with both the EWMA and the CUSUM statistics with EWMA statistic having a slight edge. Overall, EWMA has the best performance out of the three methods with the new algorithm. Thus, the new algorithm applied to the EWMA statistic provides a simple, fast, and a reliable method for early event detection in multiple data streams.

References

1. Fricker RD. Introduction to Statistical Methods for Biosurveillance. New York, NY: Cambridge University Press; 2013. 399p.
2. Runger GC, Alt FB, Montgomery DC. 1996. Contributors to Multivariate Statistical Process Control Signal. *Commun Stat Theory Methods*. 25(10), 2203-13. https://doi.org/10.1080/03610929608831832
3. Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*. 100, 9440-45. PubMed https://doi.org/10.1073/pnas.1530509100

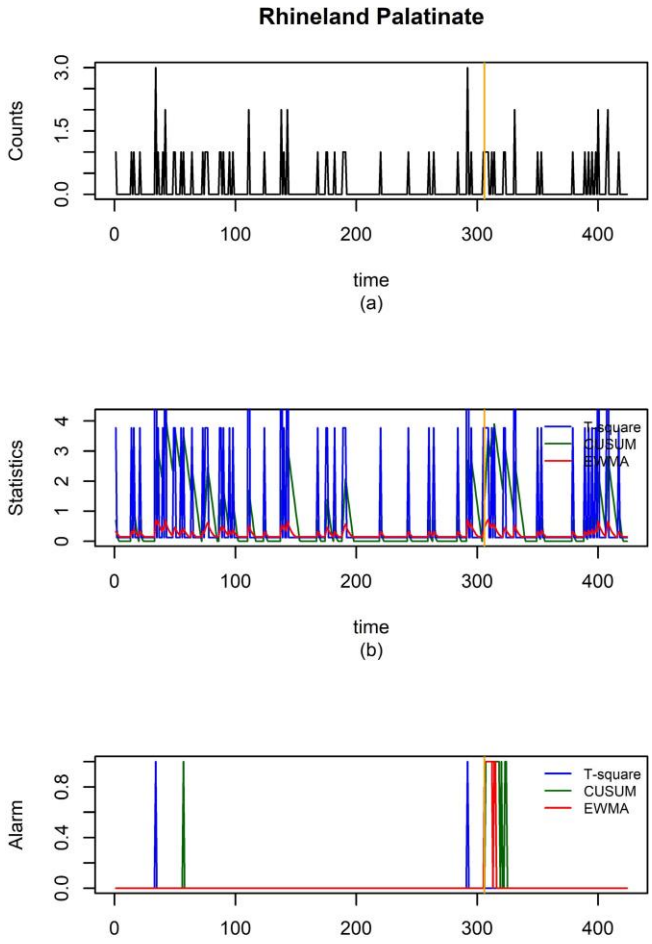

Figure 1

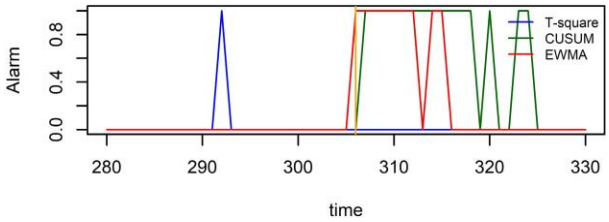

Figure 2