An Evaluation of Wavelet-Based Techniques for Prediction and Anomaly Detection in Univariate Syndromic Data

Thomas Lotze[1], Sean P. Murphy[3] and Galit Shmueli[1,2]

[1]Applied Math & Scientific Computation Program and the [2]Smith School of Business, University of Maryland College Park; [3]The Johns Hopkins University Applied Physics Laboratory

Objective

Syndromic data are created by processes that operate on different time scales (daily, weekly, or even yearly) and can include events of different durations from a 1-2 day outbreak of foodborne illness to a more gradual, protracted flu season. The duration of an outbreak caused by a new pathogenic strain or a bioterrorist attack is indeterminate. Wavelets are well suited for detecting signals of uncertain duration because they decompose data at multiple time and frequency scales. This study evaluates the use of several wavelet-based algorithms for both time series forecasting and anomaly detection using real-world syndromic data from multiple data sources and geographic locations.

Background

While several authors have advocated wavelets for biosurveillance, there are few published wavelet method evaluations using real syndromic data. Goldenberg et al performed an analysis using wavelet predictions as a way of detecting a simulated anthrax outbreak [1]. The commercial RODS application uses averaged wavelet levels to normalize for long-term trends and negative singularities [2]. In line with the implementation in [1] and in contrast to [2], we introduce two preconditioning steps to account for the strong day-of-week effect and holidays, and then use all levels of the wavelets to predict or alarm.

Methods

The wavelet algorithms used here expand upon [1] in two key aspects. First, we replace the initial cosine-transform smoothing with a preconditioning scheme that directly treats the day-of-week effect and holidays. Second, the wavelet decomposition uses a Haar basis with "backward shifted" coefficients to minimize edge problems and to enable prospective operation rather than retrospective analysis. In the first part of this study, numerous configurations for this family of wavelet algorithms were studied to yield the optimal performance. Variations on the length of the training window (including "infinite", using all previous data), the order of the AR model used for predicting each level's coefficients (or using a simple EWMA), and the forms of preprocessing (day-of-week, holiday, or neither) were examined to determine which yielded the best performance, when predicting next day and 7-day ahead forecasts. Performance was evaluated by examining the median absolute percent deviation of the forecast errors. In the second part of this study, the algorithm family was used to detect simulated single-day spike and lognormal outbreaks in a wide range of authentic syndromic data from a geographically diverse set of metropolitan areas. Detection performance was evaluated using Receiver Operating Characteristic (ROC) curves.

Results

A wavelet algorithm implemented with a 128-day sliding window, with a 7-day AR, combined with day-of-week and holiday preconditioning, proved most effective for both data prediction and outbreak detection. Overall, the detection probability is high for spike outbreaks, but the performance varies by event day of week, indicating residual serial correlation in the preconditioned data.

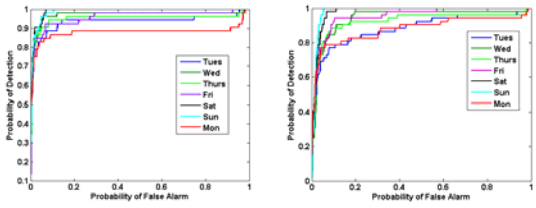

Figure 1: ROC curves for a respiratory series (left) and GI series (right), according to the day of week that an outbreak was injected.

Conclusions

From comparisons with other methods, the performance of the wavelet-based detector for univariate time series appears to be comparable to methods such as Holt-Winter's exponential smoothing. However, the distinguishing utility of the wavelet-based methods is likely to be their application to the multivariate problem. Our next steps are to investigate their performance on low-count data and on a larger variety of outbreak patterns. Another challenge is to include the day-of-week handling within the wavelet detector, eliminating the need for preconditioning.

References

[1] Goldenberg A, Shmueli G, Caruana RA, and Fienberg, SE (2002), "Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-Counter Medication Sales", *Proc. of the National Academy of Sciences*, vol. 99, issue 8, 5237-5240.

[2] Zhang J, Tsui FC, Wagner MM, and Hogan WR (2003), "Detection of Outbreaks from Time Series Data Using Wavelet Transform", *Proc. AMIA Symposium*, 748-752.

Acknowledgements

We thank Howard Burkom from the JH Applied Physics Lab for helpful discussions and feedback.