Ensemble Forecasting for Disease Outbreak Detection

Thomas H. Lotze and Galit Shmueli

University of Maryland, College Park
College Park, MD 20740
{lotze,gshmueli}@umd.edu

Abstract

We describe a method to improve detection of disease outbreaks in pre-diagnostic time series data. The method uses multiple forecasters and learns the linear combination to minimize the expected squared error of the next day's forecast. This combination adaptively changes over time. This adaptive ensemble combination is used to generate a disease alert score for each day, using a separate multi-day combination method learned from examples of different disease outbreak patterns. These scores are used to generate an alert for the epidemiologist practitioner. Several variants are also proposed and compared. Results from the International Society for Disease Surveillance (ISDS) technical contest are given, evaluating this method on three syndromic series with representative outbreaks.

Problem Description

In modern biosurveillance, time series of pre-diagnostic health data are monitored for disease outbreaks. Pre-diagnostic time series typically consist of daily counts of regional emergency department chief complaints such as cough, daily sales of cough remedies at pharmacy or grocery stores, daily counts of school absences, or in general, data that are expected to contain an early signature of a disease outbreak. Outbreaks of interest include terrorist-driven attacks, e.g. a bioterrorist anthrax release, or naturally occurring epidemics, such as an avian influenza outbreak. In either setting, the goal is to alert public officials and create an opportunity for them to respond in a timely manner.

To do this effectively, alerts must occur quickly after the outbreak begins, should detect most outbreaks, and have a low false alarm rate. There are a host of difficulties in achieving such performance (Fienberg and Shmueli, 1995), foremost among them the seasonal, nonstationary, and autocorrelated nature of the health data being monitored.

In order to identify outbreaks in pre-diagnostic health data, most modern algorithms use some type of forecasting and then monitor the residuals (i.e., forecast errors) using a control chart. These residuals provide a strong indication of outbreak: any day on which the actual count is significantly higher than expected is likely to be evidence of an outbreak. Thus, two pieces are necessary: forecasting which adapts to local trends in the data and incorporates seasonality; and good evaluation of the forecast residuals for alerting. In this paper, we examine methods for improving both.

Methodology

Multiple forecasters are generated for each time series. For these results, we used five methods: a 7-day difference, a Holt-Winters Exponential Smoother, a linear regression (using as predictors day-of-week dummy variables, cos and sin seasonality terms, and a linear index) and two windowed linear regressions (using only the past 56 or 112 days to estimate the coefficients). More details on these methods can be found in (Lotze, et. al, 2007).

For each day, the linear combination of forecasters which had the minimum squared error on past days is determined; this can be found by running a simple linear regression using the past time series values, with the past forecasts as predictors. The resulting linear combination is used to combine the five forecasts, creating an ensemble forecast value for the next day. As the nature of the series changes over time, each of the forecasters has a different accuracy. By changing the linear coefficients to reflect this, the ensemble forecaster adapts to take advantage of the local accuracy of different individual forecasters.

Residuals are then generated by subtracting the forecast from the observed value for each day. Since the variance of these residuals is different at different time points, the standard deviation is estimated using past residuals. This is then used to standardize the residuals, creating scores, $s_t = r_t / \hat{\sigma}_t$, which have approximately the same variance. This provides a common scale for deciding if a day's score is significant enough to alert. Because variance changes by day of week, seven normalization scales were used.

Once these scores are determined, new alert scores are generated by combining scores from the past few days. By doing this, the method becomes more sensitive to multi-day outbreaks and less sensitive to false alarms from random one-day deviations. The new score is a weighted average of previous scores,

$$s_{t,new} = \sum_{i=1}^{k} w_i s_{t-k+1,old} \;/ \sum_{i=1}^{k} w_i$$

Three weighting systems are considered, with varying values of k: standard average ($w_i = 1$), linear increasing ($w_i = i$), and exponentially increasing ($w_i = e^i$). A grid search on weighting method and number of days k is used to maximize performance on the training data.

Results

Three types of series were considered in the ISDS contest:

1. Patient emergency room visits (ED) with gastrointestinal symptoms
2. Aggregated over-the-counter (OTC) anti-diarrheal and anti-nauseant sales
3. Nurse advice hotline calls (TH) with respiratory symptoms

Each of these series had five years of non-outbreak data. Forecasting methods were trained on two years of data, and their mean squared error (MSE) tested on the last three. The ensemble method had the lowest MSE on each series. Adaptive sliding window variants were also considered, using only a sliding window of recent days to estimate parameters; however, this did not improve performance.

	ED	OTC	TH
Regression	268.93	12874.63	36.97
Holt-Winters	287.50	12127.18	40.69
Ensemble	266.38	10745.64	35.24

Table 1: MSE of different forecasting methods on each series

Outbreaks were created and added to the non-outbreak data, corresponding to the following three outbreak types:

1. ED: waterborne *E.coli* 0157:H7
2. OTC: waterborne outbreak of Cryptosporidium. Due to the prolonged, less severe nature of Cryptosporidium, many infected residents self-medicate, evidenced by an increase of OTC anti-diarrheal and anti-nauseant product sales.
3. TH: large-scale, seasonal influenza epidemic

Each outbreak type had 30 stochastic outbreaks generated, each following the same general shape, as seen in Figure 1.

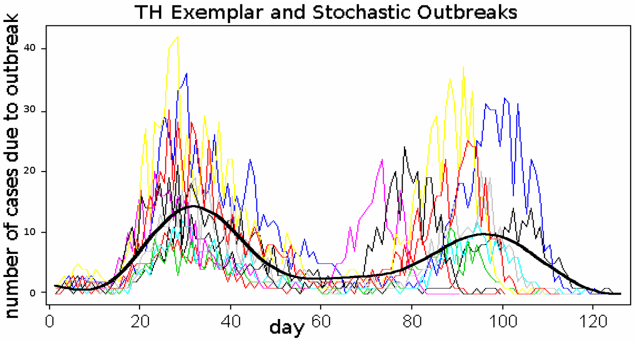

Figure 1: TH exemplar outbreak (bold) and stochastic outbreaks

Detection performance was compared across all 30 outbreak scenarios. To test detection performance, the alert limit was set to seven false alert levels, ranging from 0 to 6 false alerts per year. Under each rate of false alerts, the number of true alerts was calculated. For ED, the average delay (number of days to detection) was calculated instead, since the ED outbreaks are large enough to guarantee detection of all outbreaks, even with 0 false alerts. The final results average all seven levels.

	ED		OTC		TH	
	normal	reweight	normal	reweight	normal	reweight
Regression	8.46	6.81	0.57	0.61	0.78	0.90
Holt-Winters	7.78	7.14	0.37	0.34	0.53	0.53
Ensemble	6.92	6.71	0.51	0.61	0.78	0.81

Table 2: Avg. delay (ED) and detection rate (OTC and TH) results

Conclusions

Lesion analysis indicates that the multi-day score reweighting provided the most improvement, that ensemble forecasting can provide some improvement, and that adaptive estimation actually reduced performance on this data. The combined forecast was a better forecaster than any of the individual forecasters; however, its improvements do not always provide improved detection. Multi-day combination of alert values significantly improves performance over one-day alert values. If the multi-day outbreak distribution is known, this can be optimized and tuned for the specific outbreak shape; otherwise, when a specific shape is unknown, a general linear combination can be used to improve performance.

More work should be done to determine the relative value of different methods of combining the ensemble forecasters. This includes different windowing schemes, normalization techniques, and multi-day combinations. However, the methods proposed here, especially multi-day score reweighting, show potential for improving disease outbreak detection from pre-diagnostic data.

Acknowledgements

This work was partially supported by NIH grant RFA-PH-05-126 and was performed under an appointment to the DHS Scholarship and Fellowship Program, administered by ORISE under DOE contract number DE-AC05-06OR23100. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DHS, DOE, or ORAU/ORISE.

References

Fienberg, S. E. & Shmueli, G. *Statistical Issues and Challenges Associated with Rapid Detection of Bio-terrorist Attacks* Statistics in Medicine, 2005, 24, 513-529

Lotze, Thomas H., Murphy, Sean P., Shmueli, Galit. *Preparing Biosurveillance Data for Classic Monitoring* Advances in Disease Surveillance, 2008, forthcoming.