HWR at the Contest: A Holt-Winters-Based Method Applied to Simulated Outbreaks

Sylvia Halász, Ph..D., Colin Goodall, Ph.D., Arnold Lent, Ph.D.
AT&T Labs - Research, Florham Park and Middletown, New Jersey

OBJECTIVE

Ideal anomaly detection algorithms should detect both sudden and gradual changes, while keeping the background false positive alert rate at a tolerable level. Our objective was to develop an anomaly detection algorithm that adapts to the time series being analyzed and reduces false positive signals.

BACKGROUND

Earlier we have presented studies with HWR, where the alerts were generated using a logical OR of several different criteria [1]. The anomaly detection contest required a continuous score for each day of the time series. This gave the impetus to develop a new version of our algorithm

METHODS

We combine updated tools from industrial quality control techniques. We also account for day of week effects. The result is an exponentially weighted moving average (EWMA)-type adaptive algorithm with several parameters. For the ISDS contest, we generated anomaly alerts in HWR using extreme values of a score function, which is the cumulative sum of the normalized residuals, $\Sigma\{\xi(t) / \sigma\text{^}(\xi(t))\}$, and some quantities based on the $\xi(t) / \sigma\text{^}(\xi(t)$ which are derived from statistical quality control (SQC)-type probabilities.

The quantity $\sigma\text{^}(\xi(t))$ is the estimated standard deviation of $\xi(t)$; it is updated at each time interval using exponential weighting with parameter η. The summation is over a random interval, where the sum is reset when certain criteria are met. For the probability calculations, we assume that the $\xi(t)$ are independent random quantities. These SQC-type probabilities are based on the AT&T (Western Electric) Run Rules: hence the R in the name of the algorithm.

We collect the following statistics as we work with the time series:

- *sumres*, cumulative sum of the residuals; we reset *sumres* to 0 whenever the current standardized residual is less than *param_reset_lower* or *sumres* is larger than *param_reset_upper*
- *pos*, the number of consecutive positive residuals
- *pos1*, the number of times in the 5 day period *t-4, t-3, t-2, t-1, t* the residual is over 1, corresponding to values above 1 standard deviation
- *pos2*, the number of times in the 3 day period *t-2, t-1, t* the residual is over 2, corresponding to values over 2 standard deviations.

For the contest the score on a given day is the sum *sumres + c0 + c1 + c2 + c3 + c4 + c5*, where all *c(i)* are zero, except

c0 = p_pos5 * 1.862	if pos=5
c1 = p_pos6 * 2.155	if pos=6
c2 =2.155 + (pos – 6) * 0.265	if pos >= 7
c3 = 2.87	if pos1 >=4 in 5 days
c4 = 2.96	if pos2 >= 2 in 3 days
c5 = p_kappa * κ	if residual >= κ

The constants in the score follow from Gaussian probability calculations. Assuming that the (standardized) residuals are independent N(0,1) random variables, the probability of the events considered will be less than or equal to $1 - \Phi(c)$, where Φ is the standard normal distribution function and *c* is any of the constants in the score. p_pos5, p_pos6 and p_kappa are tuning constants for the different time series.

RESULTS

With threshold = *param_reset_upper* = 8*; param_reset_lower* = -1 and κ = 3 we obtain the results shown in Table 1. We report averages over all the test series for this fixed alert threshold. The column "maximum number of false positives in 5 years" gives the total number of false positives in the 5-year series with the highest number of false positives.

Table 1

Data type	Outbreaks missed	Avg. no. of false positives per year	Max no. of false positives in 5 years
ED	1	0.43	7
TH	4	0.62	7
OTC	8	2.08	16

CONCLUSIONS

The continuous version seems to provide good detection properties, but further studies are needed.

REFERENCES

[1] S. Halász, C. Goodall, A. Lent, J. Allegra, D. Cochrane: Detecting the Start of the Flu Season, BioSurveillance 2007, Springer Lecture Notes in Computer Science Vol. 4506.