ISDS 2018 Conference Abstracts

Nonparametric Models for Identifying Gaps in Message Feeds

Andrew Walsh*

Health Monitoring, Pittsburgh, PA, USA

Objective

Characterize the behavior of nonparametric regression models for message arrival probability as outage detection tools.

Introduction

Timely and accurate syndromic surveillance depends on continuous data feeds from healthcare facilities. Typical outlier detection methodologies in syndromic surveillance compare predictions of counts for an interval to observed event counts, either to detect increases in volume associated with public health incidents or decreases in volume associated with compromised data transmission.

Accurate predictions of total facility volume need to account for significant variance associated with the time of day and week; at the extreme are facilities which are only open during limited hours and on select days. Models need to account for the cross-product of all hours and days, creating a significant data burden. Timely detection of outages may require sub-hour aggregation, increasing this burden by increasing the number of intervals for which parameters need to be estimated.

Nonparametric models for the probability of message arrival offer an alternative approach to generating predictions. The data requirements are reduced by assuming some time-dependent structure in the data rather than allowing each interval to be independent of all others, allowing for predictions at sub-hour intervals.

Methods

Healthcare facility data was collected as HL7 messages via the EpiCenter syndromic surveillance system from June 1, 2017 through August 31, 2017. 713 facilities sent at least 1,000 messages during this period and were included in the analysis.

Standard Poisson regression models were fit to counts of messages per quarter hour. Predictors were indicators for day of week, hour of day, and quarter of hour, along with interaction terms between them.

Nonparametric logistic regression models were fit to data on the presence or absence of any message for each minute of the first two months of the study period, using the minute within the week as a predictor. The last month of data was scanned for outages at 15-minute intervals and calculating the probability of no messages since the last received message per facility as:

$$P(\text{Gap from } m_{last} \text{ to } m_{now}) = \prod_t 1 - P_{message}(t)$$

Four consecutive intervals with probability below 1^{-10} were considered outages.

Results

A total of 12,710,275 ADT A04 messages were received from 713 facilities from June 1, 2017 through August 31, 2017.

Estimation of Poisson regression models averaged 1 minute, while nonparametric models averaged 1.5 minutes to estimate. Poisson models required 672 parameters to specify, whereas nonparametric models required 29. Calculating predictions from fitted models averaged 0.2 seconds for Poisson models and 2 seconds for nonparametric models. Although predictions from the two models are not on identical scales and thus not directly comparable, they did correlate well with each other with an average correlation of 0.8.

The nonparametric regression method detected 175 resolved outages and 9 open outages in August, 2017. The resolved outages lasted an average of 1.5 days (1.75 hours to 15 days). The likelihood of these outages averaged 6e-13 (3e-160 to 4e-11).

Figure 1 illustrates how the nonparametric models can be used in a dashboard for all 713 connections. Likelihood of an outage is available for each facility based on how long it has been since the last message was received; this can be updated every minute as needed. Figure 2 illustrates the predictions from a nonparametric model for a single facility and a detected outage.

Conclusions

Nonparametric regression models of message arrival demonstrated suitable performance for use in detecting connection outages. Compared to standard Poisson regression models, computation time for nonparametric models was longer but within acceptable ranges for operational needs and storage was significantly reduced. Further, storage and computation time for standard models will increase if greater time granularity is desired, whereas the nonparametric models require no additional storage or computation. Model predictions were sufficiently similar between both models for the two to give comparable performance in detecting outages. Given the greater time flexibility of the nonparametric models and the smaller data requirements for initial model estimation (due to fewer estimated parameters), the nonparametric approach represents a promising new option for monitoring syndromic surveillance data quality.

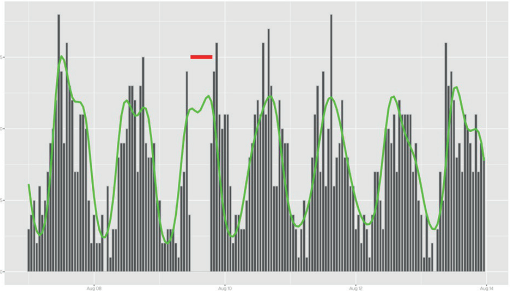

Fig. 1: Dashboard of facility connection status for all 713 facilities. Color scale indicates likelihood of outage (green - least likely; red - most likely) based on the probability of receiving no messages since last one was received.

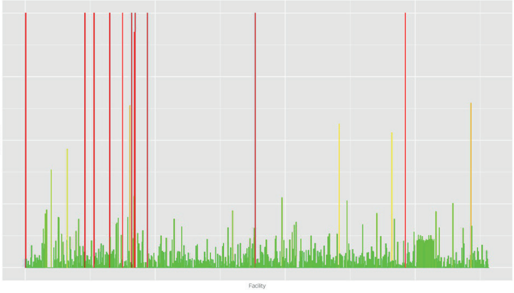

Fig 2: Hourly time series of expected (green) and observed (grey) messages, with a red bar indicating a detected outage.

Keywords

nonparametric regression; data quality; monitoring; hospital connection; real-time

Acknowledgments

We wish to thank our public health customers for funding support and data for this work.

***Andrew Walsh**

E-mail: andy.walsh@hmsinc.com