Supervised Learning for Automated Infectious-Disease-Outbreak Detection

Benedikt Zacher, Alexander Ullrich, Stephane Ghozzi

Infectious-Disease Epidemiology, Robert Koch Institute, Berlin, Germany

Objective

By systematically scoring algorithms and integrating outbreak data through statistical learning, evaluate and improve the performance of automated infectious-disease-outbreak detection. The improvements should be directly relevant to the epidemiological practice. A broader objective is to explore the usefulness of machine-learning approaches in epidemiology.

Introduction

Within the traditional surveillance of notifiable infectious diseases in Germany, not only are individual cases reported to the Robert Koch Institute, but also outbreaks themselves are recorded: A label is assigned by epidemiologists to each case, indicating whether it is part of an outbreak and of which. This expert knowledge represents, in the language of machine leaning, a "ground truth" for the algorithmic task of detecting outbreaks from a stream of surveillance data. The integration of this kind of information in the design and evaluation of algorithms is called supervised learning.

Methods

Reported cases were aggregated weekly and divided into two count time series, one for endemic (not part of an outbreak) and one for epidemic cases. Two new algorithms were developed for the analysis of such time series: *farringtonOutbreak* is an adaptation of the standard method *farringtonFlexible* as implemented in the *surveillance* R package: It trains on endemic case counts but detects anomalies on total case counts. The second algorithm is *hmmOutbreak*, which is based on a hidden Markov model (HMM): A binary hidden state indicates whether an outbreak was reported in a given week, the transition matrix for this state is learned from the outbreak data and this state is integrated as factor in a generalised linear model of the total case count. An explicit probability of being in a state of outbreak is then computed for each week (one-week ahead) and a signal is generated if it is higher than a user-defined threshold.

To evaluate performance, we framed outbreak detection as a simple binary classification problem: Is there an outbreak in a given week, yes or no? Was a signal generated for this week, yes or no? One can thus count, for each time series, the true positives (outbreak data and signals agree), false positives, true negatives and false negatives. From those, classical performance scores can be computed, such as sensitivity, specificity, precision, F-score or area under the ROC curve (AUC).

For the evaluation with real-word data we used time series of reported cases of salmonellosis and campylobacteriosis for each of the 412 German counties over 9 years. We also ran simple simulations with different parameter sets, generating count time series and outbreaks with the *sim.pointSource* function of the *surveillance* R package.

Results

We have developed a supervised-learning framework for outbreak detection based on reported infections and outbreaks, proposing two algorithms and an evaluation method. *hmmOutbreak* performs overall much better than the standard *farringtonFlexible*, with e.g. a 60% improvement in sensitivity (0.5 compared to 0.3) at a fixed specificity of 0.9. The results were confirmed by simulations. Furthermore, the computation of explicit outbreak probabilities allows a better and clearer interpretation of detection results than the usual testing of the null hypothesis "is endemic".

Conclusions

Methods of machine learning can be usefully applied in the context of infectious-disease surveillance. Already a simple HMM shows large improvements and better interpretability: More refined methods, in particular semi-supervised approaches, look thus very promising. The systematic integration of available expert knowledge, in this case the recording of outbreaks, allows an evaluation of algorithmic performance that is of direct relevance for the epidemiological practice, in contrast to the usual intrinsic statistical metrics. Beyond that, this knowledge can be readily used to improve that performance and, in the future, gain insights in outbreak dynamics. Moreover, other types of labels will be similarly integrated in automated surveillance analyses, e.g. user feedback on whether a signal was relevant (reinforcement learning) or messages on specialised internet platforms that were found to be useful warnings of international epidemic events.

Acknowledgement

The authors would like to thank Bettina Rosner for helpful explanations and comments on outbreak data.