Incorporating Learning into Disease Surveillance Systems

Daniel B. Neill, Ph.D.

Heinz School of Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213

OBJECTIVE

We argue that the incorporation of machine learning algorithms is a natural next step in the evolution and improvement of disease surveillance systems. We consider how learning can be incorporated into one recently proposed multivariate detection method, and demonstrate that learning can enable systems to substantially improve detection performance over time.

BACKGROUND

Current state-of-the-art outbreak detection methods [1-3] combine spatial, temporal, and other covariate information from multiple data streams to detect emerging clusters of disease. However, these approaches use fixed methods and models for analysis, and cannot improve their performance over time.

Here we consider two methods for overcoming this limitation, learning a prior over outbreak regions and learning outbreak models from user feedback, using the recently proposed multivariate Bayesian scan statistic (MBSS) framework [1]. Given a set of outbreak types $\{O_k\}$, set of space-time regions S, and the multivariate dataset D, MBSS computes the posterior probability $Pr(H_1(S, O_k) \mid D)$ of each outbreak type in each region, using Bayes' Theorem to combine the prior probabilities $Pr(H_1(S, O_k))$ and the data likelihoods $Pr(D \mid H_1(S, O_k))$. Each outbreak type can have a different prior distribution over regions, as well as a different model for its effects on the multiple streams. The set of outbreak types, as well as the region priors and outbreak models for each type, can be learned incrementally from labeled data or user feedback.

METHODS

We consider a simple framework for incremental learning from labeled data, given a fixed set of outbreak types and initial models for each type. For each day of data, MBSS uses the current models for outbreak detection, then updates the models based on the user's label (either "no outbreak", or outbreak type O_k and region S). An "outbreak" label enables MBSS to update its region priors for outbreak type O_k, as well as its model for the effects of O_k on each stream. Other outbreak types are not affected, except for updating the overall prior probability of each type; similarly, a "no outbreak" label only updates the overall priors. Because the number of search regions is huge, we parameterize the priors in terms of region size, shape, and duration, and learn each parameter separately. Similarly, outbreak models are parameterized in terms of severity and relative impacts on each stream. Each parameter's distribution was learned using a Bayesian maximum likelihood approach, in which the learned distribution is a weighted average of the observed distribution of labels (weighted by the number of examples) and the initial model. In our sequential framework, these distributions can be updated incrementally, by computing a weighted average of the previous distribution and current example.

RESULTS

We compared the detection performance of MBSS, with and without learning, on simulated multivariate outbreaks injected into three OTC streams (cough, fever, thermometers). Priors on region size and relative effects on each stream were learned, while other model components were fixed. Learning a single model for outbreaks injected into all three streams improved detection time by 0.5 days and increased proportion of outbreaks detected from 80% to 88%. Learning models for two outbreak types injected into the cough and fever streams (one primarily affecting cough, and one affecting fever) improved detection time by 0.6 days and proportion detected from 70% to 83%. After 10-20 examples the two outbreak types could be distinguished by the third outbreak day.

CONCLUSIONS

Using MBSS as an example, we demonstrated that the performance of outbreak detection systems can be substantially improved by learning. However, having a user "in the loop" to actively provide feedback for the system allows for much more interaction and learning than the simple framework given here. Our future work will allow users to define new outbreak types by example, including "relevant" types for true outbreaks and "irrelevant" types for other causes of a detected cluster. Models and priors for each new type will be actively learned by the system, choosing potential examples of that type for the user to label. The system will then selectively report only those clusters that are most likely to be relevant to the user and those that are most informative for learning.

Supported by NSF (IIS-0325581) and CDC (1 R01 PH000028-01).

REFERENCES

[1] Neill DB, Moore AW, Cooper GF, A multivariate Bayesian scan statistic. Advances in Disease Surveillance 2007, 2: 60.

[2] Reis BY, Kohane IS, Mandl KD, An epidemiological network model for disease outbreak detection. PLoS Medicine 2007, 4: 210.

[3] Neill DB, Lingwall J, A nonparametric scan statistic for multivariate disease surveillance. Submitted for publication.

Further Information:
Daniel B. Neill, neill@cs.cmu.edu
www.cs.cmu.edu/~neill