Learning Outbreak Regions for Bayesian Spatial Biosurveillance

Maxim Makatchev, M.S., Daniel B. Neill, Ph.D.

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

OBJECTIVE

This work incorporates model learning into a Bayesian framework for outbreak detection. Our method learns the spatial characteristics of each outbreak type from a small number of labeled training examples, assuming a generative outbreak model with latent center. We show that using the learned models to calculate prior probabilities for a Bayesian scan statistic significantly improves detection performance.

BACKGROUND

The multivariate Bayesian scan statistic (MBSS) [1] enables rapid detection and accurate characterization of emerging events by combining evidence from multiple data streams. Given a set of space-time regions S, set of event types E_k, and multivariate dataset D, MBSS uses Bayes' Theorem to compute the posterior probability $\Pr(H_1(S, E_k) \mid D)$ that each event E_k has affected each region S. The region prior probabilities $\Pr(H_1(S, E_k))$ must either be specified by a domain expert or learned from past data. In [2], we considered simple maximum likelihood methods of learning the region priors and the effects of each outbreak type. However, the huge number of regions and the limited availability of outbreak data prevent learning a multinomial distribution over all possible regions, while simpler models (e.g. uniform priors) have reduced detection power when the outbreak has predictable size and shape or when some locations are affected more frequently than others. Thus we wish to parameterize our models in a way that captures common properties of outbreaks (size, shape, connectivity, and biases toward urban or rural areas) while reducing the number of model parameters to learn.

METHODS

We propose a simple generative model that assumes a latent center location s_c and radius parameter r for each outbreak. Each location is assumed to be affected with probability $(1 + e^{(d-r)/h})^{-1}$, where d is the location's distance from the center. The distributions over centers and radii (and the steepness parameter h) are learned from a set of labeled training examples. Since each example specifies the outbreak locations but not the underlying model parameters, we estimate the parameter distributions using a generalized expectation-maximization (GEM) algorithm. The prior probability $\Pr(H_1(S, E_k))$ that a given region S is affected by a given outbreak type E_k can be easily calculated from the learned models, and these priors can improve the detection ability of MBSS. More details of our model learning approach are provided in [3].

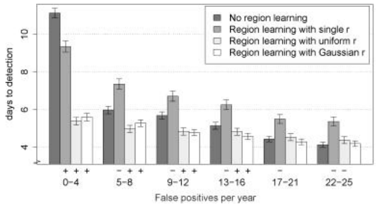

Figure 1 – Average days to detect vs. false positive rate for MBSS with and without region learning.

RESULTS

We compared the detection performance of MBSS, with and without region learning, on simulated disease outbreaks injected into real-world Emergency Department visit data from Allegheny County, PA. In the simulations shown in Figure 1, models were learned from 50 labeled outbreaks and tested on an additional 1000 outbreaks. Region models assuming a uniform or Gaussian distribution for the radius parameter r performed significantly better than the original MBSS method for false positive rates of 0-16 fp/year, and comparably for 17-25 fp/year. A simpler region model (assuming constant radius) underperformed the original MBSS method, demonstrating the importance of choosing a model representation which can capture the variability between outbreaks.

CONCLUSIONS

To our knowledge, this is the first work which incorporates a generative outbreak model into the spatial cluster detection framework. Our results demonstrate that spatial outbreak models can be learned from a small amount of training data, significantly improving detection performance. Further simulations [3] under a variety of outbreak conditions demonstrate that these models are robust to estimation errors, even when outbreaks violate our modeling assumptions.

This work was partially supported by NSF grant IIS-0325581 and CDC grant 8-R01-HK000020-02.

REFERENCES

[1] Neill DB, Cooper GF, A multivariate Bayesian scan statistic for early event detection and characterization. Machine Learning, 2008, in press.

[2] Neill DB, Incorporating learning into disease surveillance systems. Adv Disease Surveillance, 2007, 4: 107.

[3] Makatchev M, Neill DB, Learning outbreak regions in Bayesian spatial scan statistics. Proc. ICML/UAI/COLT Workshop on Machine Learning for Health Care Applications, 2008.

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