A Bayesian Scan Statistic for Spatial Cluster Detection

Daniel B. Neill[1], M.S., Andrew W. Moore[1], Ph.D., Gregory F. Cooper[2], M.D., Ph.D.

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

[2]Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15213

OBJECTIVE

This paper develops a new Bayesian method for cluster detection, the "Bayesian spatial scan statistic," and compares this method to the standard (frequentist) scan statistic approach on the task of prospective disease surveillance.

BACKGROUND

The spatial scan statistic [1] is one of the most important statistical tools for cluster detection, and is commonly used in the public health community for detection of disease clusters. However, this approach has two main disadvantages: first, it is difficult to incorporate prior knowledge, e.g. our beliefs about the size and shape of an outbreak and its impact on disease rate. Second, it is very time-consuming, and infeasible for large datasets, due to the need to calculate statistical significance by computationally expensive randomization testing. Though the "fast spatial scan" algorithm [2] can dramatically reduce computation time, we must still perform this faster search both for the original data set and for a large number of randomly generated replica data sets.

METHODS

Here we consider the natural Bayesian extension of Kulldorff's spatial scan statistic, moving from a Poisson to a conjugate Gamma-Poisson model. Given a set of spatial regions S, our goal is to compute the posterior probability $\Pr(H_1(S) \mid \text{Data})$ of an outbreak in each spatial region, as well as the posterior probability $\Pr(H_0 \mid \text{Data})$ that no outbreak has occurred. Because we have chosen a conjugate prior, we can obtain a closed-form solution for the marginal likelihood $\Pr(\text{Data} \mid H_1(S))$ for each region S, efficiently computable as a function of the aggregate count (i.e. number of disease cases) and aggregate baseline (i.e. expected count) of the region. The parameter priors (α and β for the Gamma distributions) are learned from the time series of past counts. Combining the marginal likelihoods with our region priors $\Pr(H_1(S))$ using Bayes' Theorem, we obtain the posterior probabilities of each hypothesis given the data, as desired. More details are available in the full paper [3].

To test detection power, we compared the Bayesian and frequentist scan statistic approaches on seven types of simulated respiratory outbreaks, injected into real (anonymized) Emergency Department records and over-the-counter sales data for Allegheny County, PA. These included simulated anthrax outbreaks generated by the BARD simulator [4], as well as simpler outbreaks with linear onsets, each with various parameter settings. We also compared runtime (assuming that data points are mapped to a uniform grid, and searching over all rectangular regions on the grid) for the Bayesian spatial scan, naïve frequentist spatial scan, and fast frequentist spatial scan.

RESULTS

On the outbreaks tested, the Bayesian spatial scan was shown to have higher detection power than the frequentist approach, detecting an average of 0.15 days faster at a false positive rate of 1/month. Additionally, because no randomization testing was necessary, the Bayesian spatial scan ran 900-1200x faster than the naïve frequentist spatial scan for all grid sizes. We also found that the Bayesian spatial scan was faster than the fast frequentist spatial scan for grid sizes up to 128x128, and slower for grid sizes of 256x256 and above. Both the (naïve) Bayesian and fast frequentist methods can search a 128x128 grid in under 80 minutes on our test system, as compared to over a month for the naïve frequentist spatial scan.

Thus we now have two ways of making the spatial scan computationally feasible: using the frequentist approach with the fast spatial scan algorithm of [2], and using the Bayesian approach given here. Even larger grid sizes might be searched by extending the fast spatial scan to the Bayesian method; we are currently investigating this potentially useful synthesis.

CONCLUSIONS

As compared to the frequentist method, the Bayesian spatial scan has several advantages, including higher detection power, faster computation, easier visualization and calibration, and easier combination of evidence from multiple detectors. All of these issues are considered in more detail in the full paper [3].

REFERENCES

[1] Kulldorff M, A spatial scan statistic. Communications in Statistics: Theory and Methods, 1997, 26(6): 1481-1496.

[2] Neill DB, Moore AW, Rapid detection of significant spatial clusters. Proc. 10th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2004, 256-265.

[3] Neill DB, Moore AW, Cooper GF, A Bayesian spatial scan statistic. Accepted to Neural Information Processing Systems 18.

[4] Hogan W, Cooper G, Wagner M, Wallstrom G, A Bayesian anthrax aerosol release detector. Technical Report, RODS Laboratory, University of Pittsburgh, 2004.

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