Emerging Health Threats Journal 2011, **4**:s43. **doi:** 10.3134/ehtj.10.043

www.eht-journal.org

ABSTRACT

Generalized fast subset sums for Bayesian detection and visualization

DB Neill, and Y Liu

Event and Pattern Detection Laboratory, Carnegie Mellon University, Pittsburgh, PA, USA
E-mail: neill@cs.cmu.edu

Objective

We propose a new, computationally efficient Bayesian method for detection and visualization of irregularly shaped clusters. This Generalized Fast Subset Sums (GFSS) method extends our recently proposed MBSS and FSS approaches, and substantially improves timeliness and accuracy of event detection.

Introduction

The multivariate Bayesian scan statistic (MBSS)[1] enables timely detection and characterization of emerging events by integrating multiple data streams. MBSS can model and differentiate between multiple event types: it uses Bayes' Theorem to compute the posterior probability that each event type E_k has affected each space-time region S. Results are visualized using a 'posterior probability map' showing the total probability that each location has been affected. Although the original MBSS method assumes a uniform prior over circular regions, and thus loses power to detect elongated and irregular clusters, our Fast Subset Sums (FSS) method[2] assumes a hierarchical prior, which assigns non-zero prior probabilities to every subset of locations, substantially improving detection power and accuracy for irregular regions.

Methods

We propose GFSS, a generalized Bayesian framework, which includes both FSS and the original MBSS method as special cases. As in FSS, we define a hierarchical prior over all 2^N subsets of the N locations. We first choose the center location s_c and size $n \in \{1 \ldots N\}$ uniformly at random. Given the 'neighborhood' Z consisting of s_c and its $n-1$ nearest neighbors, each location $s_i \in Z$ is independently included with probability P, where the parameter P defines the 'sparsity' of the region. FSS assumes a uniform prior over all 2^n subsets of Z, and thus corresponds to GFSS with $P = 0.5$, whereas MBSS only considers circular regions, and thus corresponds to $P = 1$.

Naïve computation of the posterior probability map using GFSS would require us to compute and sum over an exponential number of region probabilities, which is computationally infeasible for $N > 25$. However, we show that, for any value $0 < P \leqslant 1$, the posterior probability map can be computed without computing each individual region probability, thus reducing the run time from exponential to polynomial in N. In practice, this means that we can monitor hospital data from 97 Allegheny County zip codes in less than 10 s per day of data using GFSS (Figure 1).

Results

We evaluated the detection power and spatial accuracy of GFSS for 10 values of the sparsity parameter P ranging from 0.1 to 1.0. We tested these methods on 10 differently-shaped semi-synthetic outbreaks (200 injects of each type) injected

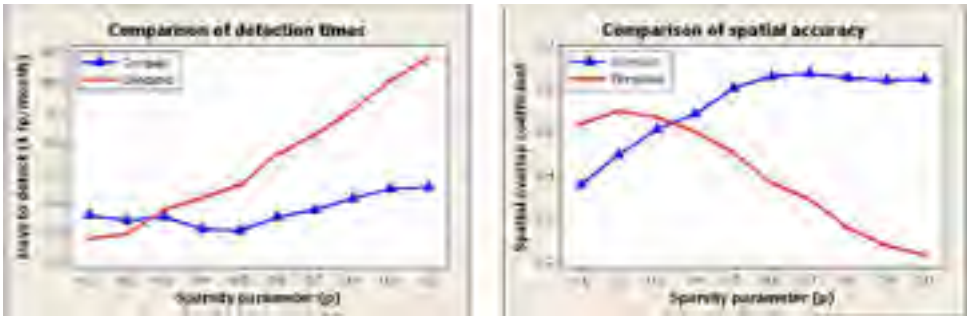

Figure 1 Detection time and spatial accuracy for GFSS, as a function of the sparsity parameter *P*.

into two streams of real-world Emergency Department data from Allegheny County, PA. Figure 1 shows the average detection time (days to detect at 1 false positive per month) and spatial accuracy (overlap coefficient between true and detected clusters) for each method. Our results show that the optimal value of P depends strongly on the shape of the outbreak: for compact clusters, the original FSS method (GFSS with $P = 0.5$) minimizes detection time, while spatial accuracy was maximized at $P = 0.7$. For highly elongated outbreaks, however, GFSS with $P = 0.2$ achieved substantial improvements as compared with FSS, including 0.8 days faster detection and 10% higher spatial accuracy. These results suggest that incorporating previous information about the density or sparsity of an outbreak can improve detection power. Additionally, GFSS enables us to more accurately distinguish between multiple outbreak types with different sparsity parameters. The optimal value of P for each outbreak type can be learned automatically, using labeled data from outbreaks of that type.

Conclusions

Our results demonstrate that GFSS can dramatically improve event detection and visualization as compared with MBSS and FSS, while still enabling fast, exact computation of the posterior probability map.

Acknowledgements

This work was partially supported by NSF Grants IIS-0916345, IIS-0911032 and IIS-0953330. This paper was presented as an oral presentation at the 2010 International Society for Disease Surveillance Conference, held in Park City, UT, USA on 1–2 December 2010.

References

1 Neill DB, Cooper GF. A multivariate Bayesian scan statistic for early event detection and characterization. *Mach Learn* 2010;**7**:261–82.
2 Neill DB. Fast multivariate Bayesian scan statistics for event detection and visualization. *Stat Med* 2010 (in press).