Fast Subset Sums for Multivariate Bayesian Scan Statistics

Daniel B. Neill, Ph.D.

H.J. Heinz III College, Carnegie Mellon University, Pittsburgh, PA 15213

OBJECTIVE

We extend our recently proposed multivariate Bayesian scan statistic (MBSS) framework to enable the detection and visualization of irregularly-shaped clusters, improving spatial accuracy and timeliness of detection while maintaining efficient computation.

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. MBSS can model and differentiate between multiple event types, and improves timeliness and accuracy of detection as compared to previous approaches [1-2]. Results can be visualized using a "posterior probability map" (Fig. 1) showing the total probability P_i that each location s_i has been affected, where $P_i = \sum_{E_k} \sum_{S: s_i \in S} \Pr(H_1(S, E_k) \mid D)$.

METHODS

While the original MBSS method assumes a uniform prior over circular regions, we extend this framework to enable detection and visualization of irregularly-shaped clusters. We define a hierarchical region prior over all subsets of the N locations, where we first choose a center location s_c and neighborhood size $n \in \{1 \ldots N\}$ uniformly at random, then consider a uniform prior over all 2^n subsets of the center location and its n – 1 nearest neighbors. Naïve computation of the posterior probability map using this prior would require us to compute and sum over an exponential number of region probabilities, which is computationally infeasible for N > 25. However, we have developed an efficient algorithmic method, "fast subset sums" (FSS), which enables us to efficiently and exactly compute the posterior probability map without computing each individual region probability. In [1], the likelihood ratio of spatial region S for a given event type E_k and event severity θ can be found by multiplying the likelihood ratios $LR(s_i \mid E_k, \theta)$ for all locations $s_i \in S$. FSS uses a similar method to compute the average likelihood ratio of the 2^n subsets for a given center s_c and neighborhood n, without considering each subset individually, by multiplying the quantities $((1 + LR(s_i \mid E_k, \theta)) / 2)$. More details of the FSS method are provided in the full paper [3].

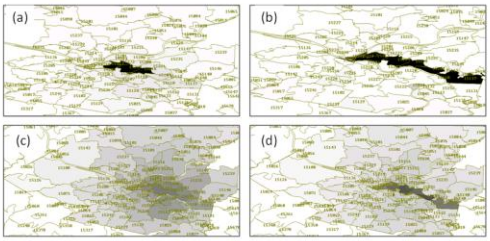

Figure 1: (a) and (b) represent average posterior probability maps of MBSS and FSS for severe, elongated outbreaks injected into 7 Allegheny County zip codes. FSS detects all 7 zip codes; MBSS only detects 4. (c) and (d) represent average posterior probability maps for MBSS and FSS for less severe outbreaks injected into the same set of zip codes.

RESULTS

We compared the run time and detection power of the FSS method to our original MBSS approach (assuming a uniform prior over circular regions) on ten differently-shaped semi-synthetic outbreaks (200 injects of each type) injected into real-world, multivariate Emergency Department data from 97 Allegheny County zip codes. Run time for FSS was ~9s per day of data, as compared to ~1s for searching over circles only. However, FSS improved timeliness of detection by an average of one day at 1 false positive/month, including an average improvement of 2.1 days for highly elongated clusters, halved the number of missed outbreaks, and significantly improved spatial accuracy (as measured by the average overlap coefficient between injected and detected clusters) from 67% to 82%. Fig. 1 compares the average posterior probability maps produced by MBSS and FSS for a highly elongated outbreak region, demonstrating that FSS more precisely identifies the affected region for both severe and less severe outbreaks.

CONCLUSIONS

Our results demonstrate that irregular clusters can be accurately detected and visualized in the multivariate Bayesian scan statistic framework using a hierarchical region prior, and that FSS enables efficient computation of the resulting posterior probability map.

This work was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0325581.

REFERENCES

[1] Neill DB, Cooper GF. A multivariate Bayesian scan statistic for early event detection and characterization. Machine Learning, 2009, in press.

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

[3] Neill DB. Fast multivariate Bayesian scan statistics for event detection and visualization. In preparation, 2009.

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