ISDS

Emerging Health Threats Journal 2011, **4**:s42. **doi:** 10.3134/ehtj.10.042

www.eht-journal.org

ABSTRACT

Fast subset scan for multivariate spatial biosurveillance

DB Neill, E McFowland III, and H Zheng

Event and pattern detection laboratory, Carnegie Mellon University, Pittsburgh, PA, USA
E-mail: neill@cs.cmu.edu

Objective

We extend the recently proposed 'fast subset scan' framework from univariate to multivariate data, enabling computationally efficient detection of irregular space-time clusters even when the numbers of spatial locations and data streams are large. These fast algorithms enable us to perform a detailed empirical comparison of two variants of the multivariate spatial scan statistic, demonstrating the tradeoffs between detection power and characterization accuracy.

Introduction

The spatial scan statistic[1] detects significant spatial clusters of disease by maximizing a likelihood ratio statistic over a large set of spatial regions. Several recent approaches have extended spatial scan to multiple data streams. Burkom[2] aggregates actual and expected counts across streams and applies the univariate scan statistic, thus assuming a constant risk for the affected streams. Kulldorff *et al.*[3] separately apply the univariate statistic to each stream and then aggregate scores across streams, thus assuming independent risks for each affected stream. Neill[4] proposes a 'fast subset scan' approach, which maximizes the scan statistic over proximity-constrained subsets of locations, improving the timeliness of detection for irregularly shaped clusters. In the univariate event detection setting, many commonly used scan statistics satisfy the 'linear-time subset scanning' (LTSS) property, enabling exact and efficient detection of the highest-scoring space-time clusters.[4]

Methods

In the multivariate setting, we wish to search over proximity-constrained subsets of locations and all subsets of the monitored data streams, but an exhaustive search over subsets is computationally infeasible, scaling exponentially with the number of streams and the maximum neighborhood size. We develop computationally efficient algorithms for both the Burkom and Kulldorff multivariate scan approaches. For Burkom's method, we iterate between two steps, optimizing over subsets of streams for the current subset of locations, and optimizing over subsets of locations for the current subset of streams. For Kulldorff's method, we iterate between optimizing over subsets of locations for fixed values of the relative risks for each stream, and recalculating the maximum likelihood risk values for the current subset of locations. Each optimization over subsets can be performed efficiently for any statistic satisfying the LTSS property: we sort the locations (streams) by a priority function, and then consider subsets consisting of the top-k highest priority locations (streams), for $k = 1 \ldots N$. We can prove that one of these will be the highest scoring subset. Both fast algorithms converge to a local maximum of the score function, and our experiments demonstrate that each closely approximates the global maximum with high probability.

Results

We compared fast localized scan (searching over proximity-constrained subsets of locations) and circular scan approaches, for both the Burkom and Kulldorff methods, monitoring multiple streams of real-world Emergency Department data from Allegheny County, PA. Our fast algorithms enable both multivariate scan statistics to be optimized over proximity-constrained subsets of the 97 zip codes and all subsets of the 16 monitored data streams in less than 2 s per day of data, whereas exhaustive search would require hundreds of millions of years. Comparing the Burkom and Kulldorff methods, we find tradeoffs between detection and characterization performance: Kulldorff's method exhibits slight but significant improvements in detection time, whereas Burkom's method more accurately characterizes the affected subset of streams. For both methods, our fast localized scan approach improved timeliness of detection by 1 to 2 days as compared with circular scan, and also increased spatial accuracy (weighted overlap coefficient between true and detected regions) from 70 to 83%. More details of our methods and results are provided in the full paper.[5]

Conclusions

By extending fast subset scan to the multivariate setting, we enable more timely detection of emerging events using

multiple data streams, as well as accurate characterization of the affected subset of streams.

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 Kulldorff M. A spatial scan statistic. *Commun Statist: Theory and Methods* 1997;**26:**1481–96.
2 Burkom HS. Biosurveillance applying scan statistics with multiple disparate data sources. *J Urban Health* 2003;**80:**i57–i65.
3 Kulldorff M, Mostashari F, Duczmal L, Yih K, Kleinman K, Platt R. Multivariate spatial scan statistics for disease surveillance. *Stat Med* 2007;**26:**1824–33.
4 Neill DB. Fast and flexible outbreak detection by linear-time subset scanning. *Adv Dis Surveill* 2008;**5:**48.
5 Neill DB, McFowland E, Zheng H. Fast subset scan for multivariate event detection and visualization. *Stat Med* 2011 (in press).