Fast Multidimensional Subset Scan for Outbreak Detection and Characterization

Daniel B. Neill* and Tarun Kumar

Event and Pattern Detection Laboratory, Carnegie Mellon University, Pittsburgh, PA, USA

Objective

We present Multidimensional Subset Scan (MD-Scan), a new method for early outbreak detection and characterization using multivariate case data from individuals in a population. MD-Scan extends previous work on multivariate event detection by identifying the characteristics of the affected subpopulation, and enables more timely and accurate detection while maintaining computational tractability.

Introduction

The multivariate linear-time subset scan (MLTSS) [1] extends previous spatial and subset scanning methods [2-3] to achieve timely and accurate event detection in massive multivariate datasets, efficiently optimizing a likelihood ratio statistic over proximity-constrained subsets of locations and all subsets of the monitored data streams. However, some disease outbreaks may only affect a subpopulation of the monitored population (age group, gender, individuals engaging in a specific high-risk behavior, etc.), and MLTSS is unable to use this additional information to enhance detection ability.

Methods

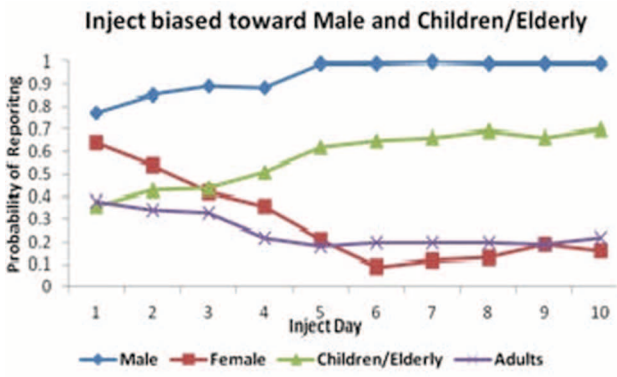
Rather than using the aggregate counts for each monitored location and data stream, we assume a set of multivariate data records representing each affected individual, with attributes such as date, home zip code, prodrome, gender, and age decile. MD-Scan jointly optimizes the likelihood ratio statistic over subsets of the values for each monitored attribute, identifying a space-time region (subset of locations and time steps) and subpopulation (including gender(s) and age groups) where the number of recent cases for a subset of the monitored prodromes is significantly higher than expected. To do so, the linear-time subset scanning property [3] is used to efficiently and exactly optimize over subsets of a given attribute, conditioned on the current subsets of all other attributes. MD-Scan then iterates over all attributes until convergence to a local optimum, and performs multiple random restarts to approach the global optimum. Additional constraints can be incorporated into each conditional optimization step, including spatial proximity, temporal contiguity, and connectedness. More details are provided in [4].

Results

We evaluated MD-Scan using simulated disease outbreaks injected into real-world Emergency Department data from Allegheny County, PA. Each outbreak was assumed to differentially affect a specific subpopulation (e.g. "adult females" or "children and the elderly"). MD-Scan achieved significantly earlier detection than MLTSS when the distribution of injected cases for the monitored attributes was sufficiently different from the background data, particularly when multiple attributes were affected or the inject was biased toward a less common attribute value. For simulated gender-specific and age-biased injects which affected only children and the elderly, MD-Scan detected over one day faster than MLTSS, and achieved 10% higher spatial accuracy. MD-Scan was also able to accurately identify the affected age and gender groups (Figure 1), while MLTSS does not characterize the affected subpopulation. Runtime of MD-Scan, while 9x slower than MLTSS, was still extremely fast, requiring an average of 4.15 seconds per day of data.

Conclusions

Our results demonstrate that MD-Scan is able to accurately identify the subpopulation affected by an outbreak, as represented by a subset of values for each monitored attribute. Additionally, MD-Scan substantially improves timeliness and accuracy of detection for outbreaks which differentially affect a subset of the monitored population. Detection performance was further enhanced by incorporating additional constraints such as spatial proximity and graph connectivity into the iterative MD-Scan procedure.



Keywords

event detection; disease surveillance; scan statistics

Acknowledgments

This work was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0953330, and a UPMC Healthcare Technology Innovation grant.

References

[1] Neill DB, McFowland E, Zheng H, Fast subset scan for multivariate spatial biosurveillance. Emerging Health Threats Journal, 2011, 4: s42.

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

[3] Neill DB, Fast subset scan for spatial pattern detection. J. Royal Statistical Society B, 2012, 74: 337-360.

[4] Kumar T, Neill DB, Fast tensor scan for event detection and characterization. Submitted.

*Daniel B. Neill
E-mail: neill@cs.cmu.edu