ISDS

Emerging Health Threats Journal 2011, **4**:s52. **doi:** 10.3134/ehtj.10.052

www.eht-journal.org

ABSTRACT

Detection of multiple overlapping anomalous clusters in categorical data

M Sabhnani, A Dubrawski, and J Schneider

Auton Lab, Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
E-mail: sabhnani@cs.cmu.edu

Objective

We present Disjunctive Anomaly Detection (DAD), a novel algorithm to detect multiple overlapping anomalous clusters in large sets of categorical time series data. We compare performance of DAD and What's Strange About Recent Events (WSARE) on a disease surveillance data from Sri Lanka Ministry of Health.

Introduction

Syndromic surveillance typically involves collecting time-stamped transactional data, such as patient triage or examination records or pharmacy sales. Such records usually span multiple categorical features, such as location, age group, gender, symptoms, chief complaints, drug category and so on. The key analytic objective to identify potential disease clusters in such data observed recently (for example during last one week) as compared with baseline (for example derived from data observed over previous few months). In real world scenarios, a disease outbreak can impact any subset of categorical dimensions and any subset of values along each categorical dimension. As evaluating all possible outbreak hypotheses can be computationally challenging, popular state-of-the-art algorithms either limit the scope of search to exclusively conjunctive definitions[1] or focus only on detecting spatially co-located clusters[2] for disease outbreak detection. Further, it is also common to see multiple disease outbreaks happening simultaneously and affecting overlapping subsets of dimensions and values. Most such algorithms[1,2] focus on finding just one most significant anomalous cluster corresponding to a possible disease outbreak, and ignore the possibility of a concurrent emergence of additional clusters.

Methods

DAD model assumes that there are multiple anomalous clusters in data where each cluster is defined as a conjunction over data dimensions and disjunctions over values along each dimension. The cluster definitions are allowed to overlap across multiple dimensions and values. It is convenient to visualize the data aggregated in a multidimensional cube with as many cells as there are unique conjunctions of all data dimensions. Each cluster spans a sub-tensor in this view of data. It is defined by two factors: location (the sub-tensor), which defines the scope of disease outbreak, and intensity, which defines the disease rate. DAD assumes that effect of overlapping clusters on any cell of the data cube are additive.

During detection, DAD algorithm iteratively adds new clusters to the model and optimizes their distribution along the data cube simultaneously. It alternately fits cluster intensities using non-negative least squares approach, and cluster locations using best subset selection approach. The algorithm uses AIC regularization to control the number of clusters reported by the model.

Results and Conclusions

We evaluated DAD against WSARE on Sri Lanka Weekly Epidemiological Reports[3]. The data stores patient visits spanning 26 regions and 9 diseases reported over 2.5 years. We injected multiple overlapping disease outbreaks in the data and then executed both algorithms to see how well they could be detected. Figure 1 (left) shows the detection accuracy (ROC) of DAD (shown in solid) and WSARE (dotted). Each experiment involved three simultaneous overlapping clusters, and the graph shows average performance over 100 such experiments. Figure 1 (right) shows time-to-detection (AMOC) characteristic. When both

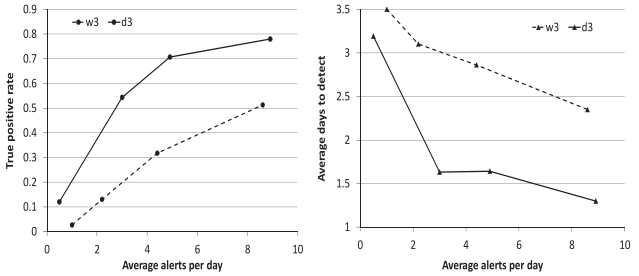

Figure 1 Accuracy (left) and timeliness of detection (right) performance of DAD and WSARE algorithms.

algorithms are allowed to generate at most three alerts per day, DAD can detect 55% of injected clusters, whereas WSARE can only detect 20%. Also, DAD can detect them in 1.5 days after onset, whereas WSARE takes almost 3 days. We found similar results for evaluations across various injection parameters: number of clusters, size of clusters, and extend of overlap between predicted and injected clusters.

Acknowledgements

This work was supported, in part, by National Science Foundation (Grant 0911032). 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 Wong W, Moore A, Cooper G, Wagner M. What's Strange About Recent Events (WSARE): an algorithm for the early detection of disease outbreaks. *J Mach Learn Res* 2005;**6**:196–98.
2 Neill D, Cooper G. A multivariate Bayesian scan statistic for early event detection and characterization. *Mach Learn J* 2009.
3 Weekly Epidemiological Reports. Sri Lanka Ministry of Health and Nutrition. http://wwwepid.gov.lk/wer.htm.