COACTION PUBLISHING

Fast graph structure learning from unlabeled data for outbreak detection

Sriram Somanchi* and Daniel B. Neill

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

Objective

Our goal is to learn the underlying network structure along which a disease outbreak might spread and use the learned network to improve the timeliness and accuracy of detection.

Introduction

Disease surveillance data often have an underlying network structure (e.g., for outbreaks that spread by person-to-person contact). If the underlying graph structure is known, detection methods such as GraphScan (1) can be used to identify an anomalous subgraph, indicative of an emerging event. Typically, however, the network structure is unknown and must be learned from unlabeled data, given only the time series of observed counts (e.g., daily hospital visits for each zip code).

Methods

Our solution builds on the GraphScan (1) and Linear Time Subset Scan (LTSS) (2) approaches, comparing the most anomalous subsets detected with and without the graph constraints. We consider a large set of potential graph structures and efficiently compute the highest-scoring connected subgraph for each graph structure and each training example using GraphScan. We normalize each score by dividing by the maximum unconstrained subset score for that training example (computed efficiently using LTSS). We then compute the mean normalized score averaged over all training examples. If a given graph is close to the true underlying structure, then its maximum constrained score will be close to the maximum unconstrained score for many training examples, while if the graph is missing essential connections, then the maximum constrained score given that structure will be much lower than the maximum unconstrained score. Any graph with a large number of edges will also score close to the maximum unconstrained score. Thus, we compare the mean normalized score of a given graph structure to the distribution of mean normalized scores for random graphs with the same number of edges and choose the graph structure with the most significant score given this distribution.

Results

We generated simulated disease outbreaks that spread based on the zip code adjacency graph with additional edges added to simulate travel patterns and injected these outbreaks into real-world hospital data. We evaluated detection time and spatial accuracy using the learned graphs for these simulated injects (Fig. 1). This figure also shows the detection performance given the true (adjacency plus travel) graph, the adjacency graph without travel patterns and the average performance given randomly generated graphs. We observe that the learned graph achieves comparable spatial accuracy to the true graph, while the adjacency graph has lower accuracy. Additionally, the learned graph is able to detect outbreaks over a day earlier than the true graph and 1.5 days earlier than the adjacency graph. Thus, our method can successfully learn the additional edges due to travel patterns, substantially improving detection performance.

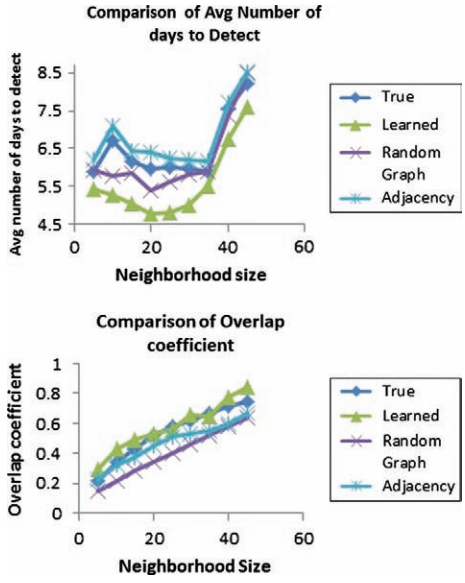

Fig. 1. Comparison of detection performance of the true, learned and adjacency graphs.

Conclusions

We proposed a novel framework to learn graph structure from unlabeled data. This approach can accurately learn a graph structure, which can then be used by graph-based event detection methods such as GraphScan, enabling more timely and accurate detection of outbreaks, which spread based on that latent structure. Our results show that the learned graph structure is similar to the true underlying graph structure. The resulting graph often has better detection power than the true graph, enabling more timely detection of outbreaks, while achieving similar spatial accuracy to the true graph.

Keywords

Event detection; biosurveillance; graph learning

Acknowledgements

This work was partially supported by National Science Foundation grants IIS-0916345, IIS-0911032 and IIS-0953330.

References

1. Speakman S, Neill DB. Fast graph scan for scalable detection of arbitrary connected clusters. In: Proceedings of the 2009 International Society for Disease Surveillance Annual Conference.
2. Neill DB. Fast subset scan for spatial pattern detection. J R Stat Soc (Ser B). 2011, to appear.

*Sriram Somanchi
E-mail: somanchi@cmu.edu

Citation: Emerging Health Threats Journal 2011, **4**: 11017 - DOI: 10.3402/ehtj.v4i0.11017