OJPHI

ISDS 2014 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Identifying Emerging Novel Outbreaks In Textual Emergency Department Data

Mallory Nobles[1], Lana Deyneka[2], Amy Ising[3] and Daniel B. Neill*[1]

[1]Carnegie Mellon University, Event and Pattern Detection Laboratory, Pittsburgh, PA, USA; [2]North Carolina Department of Health and Human Services, Communicable Disease Branch, Raleigh, NC, USA; [3]University of North Carolina, Chapel Hill, NC, USA

Objective

We apply a novel semantic scan statistic approach to solve a problem posed by the NC DETECT team, North Carolina Division of Public Health (NC DPH) and UNC Department of Emergency Medicine Carolina Center for Health Informatics, and facilitated by the ISDS Technical Conventions Committee. This use case identifies a need for methodology that detects emerging, potentially novel outbreaks in free-text emergency department (ED) chief complaint data.

Introduction

Typical approaches to monitoring ED data classify cases into pre-defined syndromes and then monitor syndrome counts for anomalies. However, syndromes cannot be created to identify every possible cluster of cases of relevance to public health. To address this limitation, NC DETECT's approach clusters cases by arrival times and monitors the textual chief complaint data associated with each identified cluster for relevant similarities [1]. This approach is time consuming and limited in its ability to detect emerging outbreaks that are dispersed across time. A new method is needed to automatically identify clusters of interest that would not be detected by existing syndromes. Clusters may be based on symptoms, events, place names, arrival time, or hospital location.

The NC DPH dataset describes 198,511 de-identified ED visits over one year at 3 North Carolina hospitals. The data include chief complaint, altered date and time of arrival, hospital A/B/C, and age group. About 40 simulated outbreaks were injected into the data set by the NC DETECT team. For example, an inject cluster might consist of 4 patients who report getting sick after eating at a particular restaurant.

Methods

Our semantic scan approach [2] is well suited to this problem. Here, we first infer a set of topics (probability distributions over words) from the free-text data using Latent Dirichlet Allocation [3]. Then, we assign each case to its most likely topic and use a variant of spatial scan [4] to identify anomalous counts of these topics. Our approach learns one set of topics using past data and a second set of topics for the most recent data (3-hour moving window). The first set of topics describes commonly occurring illness types (e.g., fever or rash). The second set of topics are chosen to be maximally different from the first set, and thus can capture clusters related to one-time events (e.g., common-source exposure) or novel disease types. Scans can be performed over combinations of other data attributes (e.g., age groups) to identify the outbreak type and affected subpopulations.

Results

Our methods successfully identified clusters of cases referring to specific locations, unusual sets of symptoms, or affected subpopulations. For example, we found a cluster of 10 cases that all mentioned a local middle school within a 4 hour span. Other detected clusters had related chief complaints, like a chemical spill, motor vehicle accident or contagious disease such as head lice or scabies. Some discovered clusters identified a location and symptoms, such as a sudden onset of rashes at the beach. Other clusters found specific subpopulations, e.g., 7 young adults complaining of smoke inhalation.

Conclusions

The topics learned by our semantic scan approach act as illness categories and eliminate the need for classifying cases into pre-defined syndromes. Since topics dynamically adapt to current data, semantic scan can identify emerging clusters that public health officials could not have predicted in advance. Finally, semantic scan is automated and sufficiently fast to identify outbreaks as they occur. We expect that the method could also be useful to other public health surveillance problems.

Keywords

text mining; event detection; semantic scan

Acknowledgments

This work was supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0953330. Data was provided by the NC DHHS/DPH NC DETECT system. The NC DETECT Data Oversight Committee does not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented.

References

1. Li M, Loschen W, Deyneka L, et al. Time of arrival analysis in NC DETECT to find clusters of interest from unclassified patient visit records. Online J Pub Health Inform 2013; 5(1): e13.
2. Murray K, Dyer C, Liu Y, Neill DB. A semantic scan statistic for novel outbreak detection. Tech. rept., Carnegie Mellon University, 2014.
3. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. J Mach Learn Res. 2003; 3:993-1022.
4. Kulldorff M. A spatial scan statistic. Commun Stat Theor Meth. 1997; 26:1481-96.

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