Detecting previously unseen outbreaks with novel symptom patterns

Yandong Liu and Daniel B. Neill*

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

Objective

We propose a new text-based spatial event detection method, the semantic scan statistic, which uses free-text data from emergency department chief complaints to detect, localize and characterize newly emerging outbreaks of disease.

Introduction

Commonly used syndromic surveillance methods based on the spatial scan statistic (1) first classify disease cases into broad, preexisting symptom categories (prodromes) such as respiratory or fever, then detect spatial clusters where the recent case count of some prodrome is unexpectedly high. Novel emerging infections may have very specific and anomalous symptoms, which should be easy to detect even if the number of cases is small. However, typical spatial scan approaches may fail to detect a novel outbreak if the resulting cases are not classified to any known prodrome. Alternatively, detection may be delayed because cases are lumped into an overly broad prodrome, diluting the outbreak signal.

Methods

We propose a new approach to detect emerging patterns of keywords in the chief complaint data. Our semantic scan statistic has three steps: automatically inferring a set of topics (probability distributions over words) from the data using Latent Dirichlet Allocation (2), classifying each chief complaint to the most likely topic, and then performing a spatial scan using the case counts for each topic. We compare three variants of the semantic scan: static (topics are learned from historical data and do not change from day to day), dynamic (topics are recalculated each day using the most recent two weeks of data) and incremental (not only using the static topics but also learning additional 'emerging' topics that differ substantially from the static topics).

Results

We compared the three semantic scan methods to the standard, prodrome-based spatial scan using synthetic disease outbreaks injected into real-world emergency department data from Allegheny County, PA. We first considered 55 different outbreak types, corresponding to all distinct ICD-9 codes with at least 10 cases, which were mapped to one of the existing prodromes. For each outbreak type, we generated spatially localized injects with chief complaints sampled from the cases with that ICD-9 code (Fig. 1). The static, dynamic and incremental methods required an average of 7.7, 7.1 and 6.9 days, respectively, to detect and were able to precisely characterize the outbreak based on the detected topic (e.g., top keywords for ICD-9 code 569.3 were 'rectal', 'bleed', and 'bleeding'). The prodrome method achieved more timely detection (5.0 days to detect) but with much less precise characterization (e.g., 'hemorrhagic' for ICD-9 code 569.3). Next, we considered both randomly selected, unmapped ICD-9 codes and synthetically generated unprecedented events, such as an outbreak that makes the patient's nose turn green. The prodrome method required 10.9 days to detect these outbreaks, while the semantic scan was able to achieve much faster detection. For example, for the green nose outbreak, the static, dynamic and incremental methods detected in 6.4, 5.3 and 5.6 days, respectively. The dynamic and incremental methods correctly identified the emerging topic (keywords 'green', 'nose', 'nasal', etc.), while the static method did not, since the outbreak did not correspond to any of the topics learned from historical data.

Conclusions

The semantic scan statistic can successfully capture emerging spatial patterns in free-text chief complaint data, enabling more timely detection of novel emerging outbreaks with previously unseen patterns of symptoms. Other advantages include more accurate characterization of outbreaks (identifying a set of keywords that precisely describe the disease symptoms) and the ability to detect outbreaks without preexisting syndrome definitions. Additionally, our methods have the potential to achieve more timely detection by incorporating free-text data sources, such as Twitter and other social media tools, into the surveillance process.

Keywords

Text mining; event detection; semantic scan

Acknowledgments

This work was partially supported by NSF grants IIS-0916345, IIS-0911032 and IIS-0953330.

References

1. Kulldorff M. A spatial scan statistic. Commun Stat Theor Meth. 1997;26:1481–96.
2. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.

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

Citation: Emerging Health Threats Journal 2011, **4**: 11074 - DOI: 10.3402/ehtj.v4i0.11074