Non-Parametric Scan Statistics for Disease Outbreak Detection on Twitter

Feng Chen* and Daniel B. Neill

CMU, Pittsburgh, PA, USA

Objective

We present a new method for disease outbreak detection, the "Non-Parametric Heterogeneous Graph Scan (NPHGS)". NPHGS enables fast and accurate detection of emerging space-time clusters using twitter and other social media streams where standard parametric model assumptions are incorrect.

Introduction

Disease outbreak detection based on traditional surveillance datasets, such as disease cases reported from hospitals, is potentially limited in that the collection of clinic information is costly and time consuming. However, social media provides the vast amount of data available in real time on the internet at almost no cost. Our solution, NPHGS, provides a nonparametric statistical approach for outbreak detection that well addresses the key technical challenges in social media streams: 1) large volume; 2) highly informal, ungrammatical and dynamic; and 3) heterogeneous correlations.

Methods

Given twitter and other social media streams, we first model the streams as a heterogeneous graph, in which: 1) each node can be of different types, such as user, tweet, geographic location, keyword, and hashtag; 2) the relationships between nodes can be different types, such as retweet, reply, and follower; and 3) each node type can have different attributes, such as the number of tweets and users for a given geographic location; the number of followers for a given user; and the number of sentiment score for a given tweet. Second, we further model the network as a "sensor" network, in which each node senses its "neighborhood environment" and reports empirical p-values measuring the current anomalousness levels of various neighborhood-related attributes. Third, we efficiently maximize a nonparametric scan statistic over connected sub-graphs to identify the most anomalous network clusters. Each cluster is returned as the indicator of an ongoing or upcoming outbreak event.

Results

We randomly selected ten percent of all the raw twitter data from 2012-June to 2013-April in the country Chile, where 19 rare Hantavirus disease outbreaks in more than eight different states have been reported in local news reports, such as La Tecera and Las Ultimas Notlcias. For each gold standard event reported by news, we determine whether our method: a) Have an alert in the same state 1-7 days before the event, named as successful forecasting. Record the number of days of lead time for that event based on the earliest such alert; b) did not have an alert in that state 1-7 days before the event, but did have an alert in that state 0-7 days after the event, named as successful detection. Record the number of days of lag time for that event based on the earliest such alert; c) otherwise, the alert is regarded as false alarm. Note that, sophisticated preprocesses have been conducted to extract high quality related tweets and to accurately estimate their geographic locations.

Promising results are show in Figures 1 and 2. First, we observe from Figure 1 that the detection true positive rate (TPR) reaches up to around 65% with only the reasonable low false positive rate around 22%. The forecasting rate is lower (up to around 20% TPR with 22% FPR), which potentially means that either 1) twitter is better fitted for event detection rather for forecasting or 2) our current approach, NPHGS, have the limited ability to capture the trigger signals for forecasting. However, it has been well-known that event forecasting is a very hard problem. Second, Figure 2 shows that when FPR is the lowest, NPHGS has the lag time around six days, but when FPR is the reasonable low value around 22%, NPHGS has the lag time only around 1 day, which is usually better than detection methods (around 3 to 4 days lag time) using traditional public surveillance data.

Conclusions

NPHGS is a new nonparametric statistical approach for heterogeneous social media streams. The case study on twitter demonstrates promising results using twitter data.

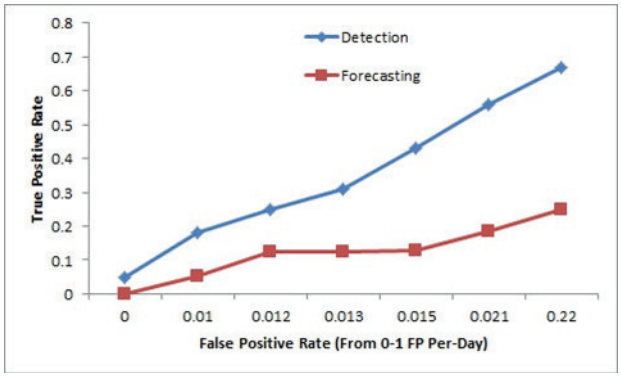

Figure 1: FPR vs. TPR (detection and forecasting)

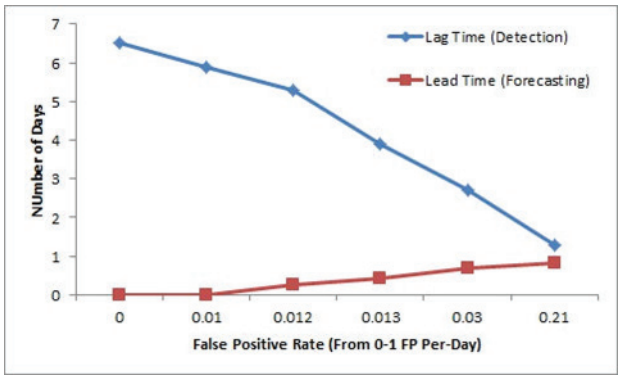

Figure 2: FPR vs. Lead Time and Lag Time

Keywords

non-parametric scan statistics; social media; disease outbreak

*Feng Chen
E-mail: fchen1@cmu.edu