Forecasting Rare Disease Outbreaks with Spatio-temporal Topic Models

Saurav Ghosh[1], **Theodoros Rekatsinas**[2], **Sumiko R. Mekaru**[3], **Elaine O. Nsoesie**[3]
John S. Brownstein[3], **Lise Getoor**[3], **Naren Ramakrishnan**[1]
sauravcsvt@cs.vt.edu, thodrek@cs.umd.edu
{sumiko.mekaru,elaine.nsoesie, john.brownstein}@childrens.harvard.edu,
getoor@soe.ucsc.edu, naren@cs.vt.edu
[1]Virginia Tech, [2]University of Maryland
[3]University of California, Santa Cruz, [4]Boston Children's Hospital

Abstract

Rapidly increasing volumes of news, tweets, and blogs are proving to be extremely valuable resources in helping anticipate, detect, and forecast significant societal events. In this paper, we focus on the problem of forecasting rare disease outbreaks and demonstrate how spatio-temporal topic models over health-related newspaper articles can successfully be used to forecast outbreaks. More precisely, we present a novel framework that integrates topic models with one-class SVMs, so that modeling the underlying topic evolution and forecasting its prominence can be used as a surrogate for making near-term predictions of disease outbreaks. We demonstrate the effectiveness of our proposed technique using incidence data for Hantavirus in multiple countries of Latin America.

1 Introduction

There has been a growing interest in developing statistical models for detecting infectious diseases as they arise, in a sufficiently timely fashion to enable effective control measures to be taken. Most of the early approaches targeted specific diseases and relied on highly specialized data, including medical records or environmental time series [11, 10]. Recently, however, there has been a growing interest in monitoring disease outbreaks using publicly available data on the Web, including news articles [2, 7], blogs [3], search engine logs [6] and micro-blogging services, such as Twitter [4]. Due to their volume, ease of availability, and 'citizen participation', such 'open source indicators' have been shown to be quite effective at monitoring disease emergence and progression.

While effective at detecting outbreaks of common diseases, such as influenza, the above techniques have significant limitations at predicting outbreaks of *rare*, yet deadly, diseases, such as Hantavirus. Here we propose a novel framework for spatially targeted prediction of rare disease outbreaks. The proposed framework leverages the temporal topic models formalism and auto-regression techniques proposed by Matsubara et al. [8] as well as one-class SVMs [9]. More precisely we show how the temporal topic models over archival and ongoing news articlces can enable detection of emerging disease-related topics for a collection of predefined locations. Furthermore, we show how one-class SVMs can be used on the output topic distribution to detect anomalous topic distributions that constitute early indicators of the onset of an outbreak. We evaluate and demonstrate the effectiveness of the proposed framework for forecasting Hantavirus outbreaks in Latin America.

2 Framework Overview

Temporal Topic Models for Newspaper Articles. We assume as input a collection of time-stamped health related news articles over a time period T, which is assumed to be discretized into fixed time intervals $(t_1, t_2, \ldots, t_N)$ of length l (typically one week). Each news article is associated with (i)

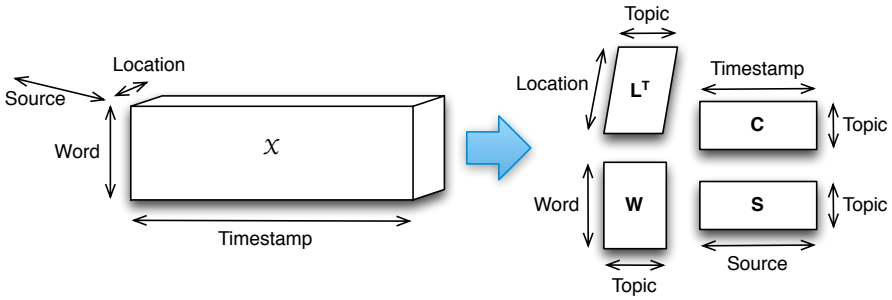

Figure 1: Illustration of the tensor decomposition used to define the temporal topic model.

a disease incidence location, (ii) a time stamp, and (iii) a news provider, which we will refer to as the *source*. Each article can be viewed as a collection of words, and hence, our input can be viewed as a sequence of words from different sources. It is convenient to treat the given sequence as a four dimensional tensor, i.e., $\mathcal{X} \in \mathbb{N}^{|W|\times|S|\times|L|\times N}$, where W denotes the set of the words under consideration, S is the set of the news article sources, and L is the set containing the locations of interest. As stated earlier, our time granularity is one week, defined as the 7-day period from Sunday to Saturday referred to as an *epidemiological week*, or *epi week* for short. For example, a tuple ("hanta",("Los Lagos", "Chile"), "www.biobiochile.cl", '28': 2) means that the count of news articles mentioning the word "hanta" is 2 in the state of Los Lagos in Chile over the Epi week 28.

Each entry in $\mathcal{X}$ is assumed to be generated from a *latent topic*. The hidden topics can be modeled in terms of the four dimensions mentioned above, namely, "word", "source", "location" and "time". We consider decomposing the input with respect to source since we assume that different sources focus on different diseases. Assuming K latent topics, one can decompose tensor $\mathcal{X}$ into four matrices $\mathbf{W}(|W| \times K)$, $\mathbf{L}(K \times |L|)$, $\mathbf{S}(K \times |S|)$, and, $\mathbf{C}(K \times N)$ with non-negative entries (see Figure 1). Rows in $\mathbf{L}, \mathbf{C}$ and $\mathbf{S}$, associated with a topic $r \in K$, correspond to categorical distributions describing the probability of observing each word, location, epi-week or source given r while rows in $\mathbf{W}$ correspond to probabilities of observing each topic r given a word w . By analyzing these distributions, we can identify which of the K topics are related to rare disease topics of interest (i.e., *rare disease topics*).

We estimate the probability of a rare disease topic becoming prominent for a given location l at a future time interval t, denoted by PS_r as:

$$PS_r = \Pr(r|l,t) = \sum_{s\in S} \Pr(r|s,l,t)\Pr(s|l,t) = \sum_{s\in S}\sum_{w\in W} [\Pr(r|w,s,l,t)\Pr(w|s,l,t)]\Pr(s|l,t) \quad (1)$$

where s and w represent a source and a word respectively. In Equation (1), $\Pr(s|l,t)$ corresponds to the probability that given a location l and a time interval t an article will be reported by source s; this corresponds to the *coverage* of source s for the location l at time t. Finally, $\Pr(r|s,l,t)$ corresponds to the probability that topic r will be covered by s for location l at time t.

Given a word w, the topic r is independent of s, l, and t, thus, $\Pr(r|w,s,l,t) = \Pr(r|w)$. By the chain rule we also have that $\Pr(w|s,l,t) = \frac{\Pr(w,s,l,t)}{\Pr(s,l,t)}$. The term $\Pr(w,s,l,t)$ can be estimated using the count of news articles corresponding to the tuple $< w,s,l,t >$ denoted by $x_{w,s,l,t}$. We have that $\Pr(w,s,l,t) = \frac{x_{w,s,l,t}}{\sum\limits_{w} x_{w,s,l,t}}$. Notice that $x_{w,s,l,t}$ corresponds to a future time interval and hence needs to be estimated. We estimate $x_{w,s,l,t}$ as

$$x_{w,s,l,t} \propto \bar{x}_w \sum_{r=1}^{K} W_{w,r} \cdot L_{r,l} \cdot C_{r,t} \cdot S_{r,s} \quad (2)$$

where $W_{w,r}$, $L_{r,l}$, and $S_{r,s}$ can be retrieved by the corresponding matrices. However $C_{r,t}$ corresponds to a future time interval and needs to be estimated. We use matrix $\mathbf{C}$ to forecast the values $C_{r,t}$with $r \in \{1,2,\ldots,K\}$. We use an autoregressive model over the values of topic r for the n previous time intervals, denoted by $C_{r,t-1}, C_{r,t-2}, \ldots, C_{r,t-N}$.We have:

$$C_{r,t} = a_1 \cdot C_{r,t-1} + a_2 \cdot C_{r,t-2} + \cdots + a_n \cdot C_{r,t-n} \quad (3)$$

where $a_1, a_2, ..., a_n$ are the regression coefficients. Combining the equations above, we have $\Pr(r|l,t) = \sum_s \sum_w \frac{\Pr(r|w)\Pr(w,s,l,t)}{\Pr(l,t)}$. The term $P(l,t)$ is constant for a given location l and future timestamp t, and hence, $\Pr(r|l,t) \propto \sum_s \sum_w \Pr(r|w) \cdot \Pr(w,s,l,t) \propto \sum_s \sum_w W_{w,r} \frac{x_{w,s,l,t}}{\sum_w x_{w,s,l,t}}$. To derive a proper probability distribution over rare disease topics we normalize the term $PS_r = \Pr(r|l,t)$ over all topics for a given location and timestamp.

Detecting Anomalies with One-Class SVMs. To predict the incidence of a rare disease outbreak for a location l at time t, we reason about the predicted prominence probabilities of rare disease topics, detecting if the probabilities indicate an anomalous point, i.e., the incidence of a disease outbreak. To detect anomalous points we use one-class SVMs [9] (OCSVM). A OCSVM maps input data X into a high dimensional feature space H via a kernel $\Phi : X \rightarrow H$ and finds the maximal margin hyperplane which best separates the training data from the origin. The classification rule corresponds to $f(x) = \mathsf{sign}(\mathbf{w}\Phi(x) - b)$, where $\mathbf{w}$ is a weight vector and b is a bias term. We use this classification rule to detect if a new point x is an anomalous point (i.e., $f(x) < 0$) or not. The sets of training examples for each of these OCSVMs is comprised by the predicted rare disease topic distributions for each location over the time intervals in the time window T.

Recall that our goal is to detect the incidence of a particular disease outbreak for a specific location in L. Operationally, we train a separate OCSVM for each location and disease pair and forecast outbreaks on a weekly basis. This approach thus predicts *if* a disease outbreak will happen and *where* it will happen (since we are training and forecasting for each location). For the *when* since we are predicting for an epi week we adopt a standard relative date within the epi week to be the date at which the rare disease incidence will occur, and tune it using cross-validation.

3 Experimental Evaluation

Datasets. Our corpus of public health-related news articles is drawn from HealthMap [5] (http://healthmap.org), a prominent online source of news articles and tweets for disease outbreak monitoring and real-time surveillance of emerging public health threats. In this paper, we focus on HealthMap articles from Latin America. Traditional IR pre-processing such as stopword removal and term frequency modeling is performed over a fixed vocabulary of words. The dictionary contains words that are either commonly associated with diseases (e.g., "contagious") or words associated with a specific disease (e.g., "rodents","hanta" for Hantavirus). The latter are also used to identify the topics that are most probable to correspond to rare disease topics. When predicting for an epiweek we use historical (weekly) data from June 2012 up to the previous week to construct the tensor decomposition and train the OCSVM. We evaluate the performance of our approaches from January 2013 to May 2013. The size of the input tensor varies over time, as new articles are added every epiweek. The number of words in the tensor ranges from 20908 to 45163, the number of locations from 74 to 144 and the number of HealthMap data sources from 381 to 798.

GSR. We also make use of a gold standard report (GSR) which gives ground truth determinations of whether a disease incidence (Hantavirus) happened in a given country. The GSR is determined by analysts (not co-authors of this paper) poring over multiple Latin American or international news sources and studying bulletins issued by health reporting organizations such as ProMED [1]. In practice, outbreak alerts are useful only when there has not been an outbreak in the near past. The analysts adopt a 6-month rule wherein the GSR data does not include rare disease incidences in locations for which there has been an earlier outbreak reported within the past six months.

Metrics. We adopt four key measures of performance. Given our predictions, we compute the recall and precision at a country level, grouping together predictions for locations in the same country. We also compute an average warning quality for each country. Each prediction for a location in the country under consideration is assigned a quality score $Q = \frac{4}{3}(1 + a_{loc} + a_{date})$, where a_{loc} and a_{date} denote the location and date accuracy of the prediction. The quality score takes values between 0 and 4. Finally, we consider the lead time of our predictions, which is calculated as the time between the date of alerting and the actual date of reporting of the outbreak (not the incidence date of the outbreak).

BRM. We also compare the performance of our framework against a *base rate model* (BRM). This model assumes a fixed rate for the occurrence of rare disease outbreaks for each country and for each month. To determine this rate, the model extracts the average frequency of outbreak occurrences reported over a past time window of four months. BRM reports disease outbreaks for that country at

Table 1: Performance results for spatio-temporal topic models with OCSVMs and BRM.

Month	Country	Base Rate Model			Spatio-temporal topic models + OCSVMs			
		Qual.	Rec.	Prec.	Qual.	Lead Time	Rec.	Prec.
Jan.	Chile	2.83	0.17	0.50	**2.92**	7.5	**0.67**	**1.0**
	Overall	2.83	0.17	0.50	**2.92**	7.5	**0.67**	**0.57**
Feb.	Chile	2.58	**0.68**	0.45	**3.36**	10.0	0.5	**0.5**
	Overall	2.58	**0.68**	**0.45**	**3.36**	10.0	0.5	0.4
Mar.	Chile	**2.54**	**0.8**	**0.8**	**2.54**	13	0.5	0.5
	Argentina	-	0	-	**2.54**	1.0	**1.0**	**1.0**
	Uruguay	-	0	-	**2.92**	5.0	**1.0**	**1.0**
	Overall	2.54	0.32	**0.8**	**2.66**	6.33	**0.75**	0.75
Apr.	Chile	2.59	**0.7**	**0.7**	**2.73**	6.5	0.67	0.67
	Argentina	-	0	-	**2.92**	3.00	**1.0**	**0.25**
	Overall	2.59	0.53	**0.7**	**2.79**	5.33	**0.75**	0.42
May	Chile	2.63	0.86	0.57	**3.015**	7.5	**1.0**	**0.67**
	Argentina	2.48	0.24	**0.72**	**3.015**	7.5	**0.67**	0.67
	Overall	2.55	0.48	0.61	**3.015**	6.0	**0.8**	**0.66**

a frequency equal to the extracted rate. Alerting dates are assigned to the beginning of each month while events dates are assigned uniformly at random to a day within the corresponding month. (Thus lead time is not a meaningful criterion to evaluate the BRM.) The performance of BRM is enhanced by taking the average performance over 25 independent runs.

Mapping events to alerts. Since there could be multiple events (and/or alerts) in a given month, a strategy is necessary to map events to alerts. We conduct a maximum bipartite matching between events and alerts where i) an edge exists if the alert was issued prior to the reporting date of the event, ii) the weight on the edge denotes the putative quality score.

Results. We focus on three countries, i.e., Chile, Argentina, and Uruguay, for which cases of hantavirus outbreaks were reported from January to May 2013. We perform the tensor decomposition described in Section 2 using 12 disease topics. We run our OCSVM model - using a linear kernel for the SVMs corresponding to Chile and Argentina and a radial basis function (rbf) kernel for the SVMs corresponding to Uruguay - on the predicted topic probabilities for all available locations in these three countries. The results are shown in Table 1. As shown our approach can effectively detect outbreaks with an average lead-time of 6.4 days. Moreover, our approach outperforms BRM in all cases in terms of quality, meaning that our approach is better at predicting diseases for particular locations. Notice that by definition the BRM model performs poorly when the hantavirus season rises (or begins) in a particular country. If we look at the recall values carefully, the recall value for BRM is very low (0.17) in January since in Chile the hantavirus season picks up in January with respect to December because in December there are no incidents while in January there are 6 incidents. In February, the recall value for BRM in Chile increases to 0.65 because in February the hantavirus season falls down in Chile with respect to January. Since the spatio-temporal topic model is trained on the news articles data (more ground truth), it can adapt well to changes in seasonality as evident from its recall values. Finally, it was capable of detecting outbreaks in Argentina and Uruguay during March and April without any incident being reported in the GSR for the past few months in these 2 countries; BRM fails as it is dependent on the GSR data.

Acknowledgments

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

[1] International society for infections diseases. `http://www.promedmail.org/`.

[2] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl. Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. *PLoS Medicine*, 5(7), 2008.

[3] C. Corley, D. Cook, A. Mikler, and K. Singh. Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, 7(2), 2010.

[4] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. SOMA '10, pages 115–122, 2010.

[5] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association*, 15:150–157, 2008.

[6] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457, 2009.

[7] J. P. Linge, R. Steinberger, T. P. Weber, R. Yangarber, and E. van der Goot. Internet surveillance systems for early alerting of health threats. *Eurosurveillance*, 14(13), 2009.

[8] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa. Fast mining and forecasting of complex time-stamped events. KDD 2012.

[9] B. Schoelkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *NEURAL COMPUTATION*, 13:2001, 1999.

[10] W.-K. Wong, A. Moore, G. Cooper, and M. Wagner. Bayesian network anomaly pattern detection for disease outbreaks. ICML 2003.

[11] W.-K. Wong, A. Moore, G. Cooper, and M. Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. AAAI 2002.