Editor: **Daniel B. Neill,** H.J. Heinz III College, Carnegie Mellon University, neill@cs.cmu.edu

New Directions in Artificial Intelligence for Public Health Surveillance

Daniel B. Neill, *Event and Pattern Detection Laboratory, H.J. Heinz III College, Carnegie Mellon University*

Public health surveillance is the process of detecting, characterizing, tracking, and responding to disease outbreaks, other health threats (such as a bioterrorist attack, radiation leak, or contamination of the food or water supply), and other patterns relevant to the health of populations (such as obesity, drug abuse, mental health, or malnutrition). This surveillance takes place at the local, state, national, and global scales, and often requires coordination of multiple entities (for example, hospitals, pharmacies, and local, state, and federal public health organizations) to achieve a timely, focused, and effective response to emerging health events. In this work, we focus on the role that AI and machine learning methods can play in assisting public health through the early, automatic detection of emerging outbreaks and other health-relevant patterns.

The last decade has seen major advances in analytical methods for outbreak detection, including (but not limited to) analysis of spatial and temporal data,[1] integration of multiple data streams,[2] Bayesian methods to model and differentiate between multiple event types,[3,4] more realistic outbreak simulations,[5] and improved metrics to evaluate detection performance.[6] Deployed systems have incorporated many such methods, monitoring a variety of public health data sources such as Emergency Department visits and over-the-counter medication sales[7] and enabling more timely and accurate identification of disease outbreaks in practice.

While most existing surveillance systems rely heavily on basic statistical methods such as time series analysis together with the expert knowledge of public health practitioners, we believe that the disease surveillance field is entering a major paradigm shift due to a dramatic increase in the number, quantity, and complexity of available data sources. Current disease surveillance systems are relying more and more on massive quantities of data from nontraditional sources, ranging from Internet search queries and user-generated Web content, to detailed electronic medical records, to continuous data streams from sensor networks, cellular telephones, and other location-aware devices. This shift toward analysis of data at the societal scale will require a corresponding shift in the methodologies employed in practical disease surveillance systems, incorporating techniques from AI, machine learning, and data mining to make sense of the massive quantity of data, to detect relevant patterns, and to assist public health decision making. Practitioners will increasingly rely on tools and systems that use advanced statistical methods to accurately distinguish relevant from irrelevant patterns, scalable algorithms to process the massive quantities of complex, high-dimensional data, and machine learning approaches to continually improve system performance from user feedback. Thus we believe that the next decade of disease surveillance research will require us to address three main challenges:

1. Making disease surveillance systems more interactive, enabling users to make sense of the mass of available data.
2. Exploiting the richness and complexity of novel data sources at the societal scale.
3. Creating new methods which can scale up to massive quantities of data and can integrate information from large numbers of data sources.

In the remainder of this article, we briefly discuss two current lines of research with potential to address some of these challenges. While these

1541-1672/12/$31.00 © 2012 IEEE
Published by the IEEE Computer Society

examples are taken from the work of our own Event and Pattern Detection Laboratory, we note that many initiatives along these lines are underway in the broader disease surveillance, statistics, AI, and machine learning communities.

Detecting Outbreaks Through Rich Text Analysis

While the future holds great potential for utilizing a multitude of data sources for outbreak detection, the fact that much of this data (including electronic health records and nontraditional public health data sources such as Twitter feeds) exists only as unstructured free text poses a major challenge. Many existing disease surveillance systems approach this problem by predefining a set of broad syndrome groupings (or "prodromes"), such as "respiratory illness," "gastrointestinal illness," "influenza-like illness," and others, and monitoring Emergency Department (ED) visits and other data sources for unexpected increases in the number of cases for each prodrome. The analysis classifies each ED case into one of the existing set of prodromes (or "unknown") by keyword matching, Bayesian network-based classification,[8] or other text classification approaches, and then applies standard temporal or spatio-temporal surveillance methods to the resulting case counts.

The prodrome-based approach has two main disadvantages: first, mapping specific chief complaints (such as "coughing up blood") to a broader symptom category ("respiratory" or "hemorrhagic") is likely to dilute the outbreak signal, delaying or preventing detection of an emerging outbreak. For example, an extra 10 cases of respiratory illness in an area might not be sufficient to detect an outbreak; but if all 10 patients were coughing up blood, this anomalous pattern might enable much earlier detection. Additionally, approaches based on mapping cases to existing prodromes have little or no ability to detect novel outbreaks with previously unseen symptom patterns. For example, a newly emerging infection that makes the patient's nose turn green should only require a small number of cases for detection (since each individual observation is highly anomalous); but if a system maps such cases to the "unknown" category, it might require many such cases to detect the outbreak, dramatically reducing the ability of public health practitioners to respond to it in a timely and effective manner.

In recent work, we have addressed this problem by proposing a new text-based spatial event-detection approach, the "semantic scan statistic," which uses free-text data from Emergency Department chief complaints to detect, localize, and characterize emerging outbreaks of disease[9] (see Figure 1). This approach detects emerging spatial patterns of keywords in the chief complaint data by learning "topic models" (probability distributions over words) from the data using Latent Dirichlet Allocation[10] and classifying cases into one or more of the newly learned topics. We then identify space-time clusters of cases corresponding to any of the learned topics.

We developed and compared three variants of the semantic scan. In the "static" approach, we learned a set of 25 topics from historical data; these topics did not change from day to day. This approach was able to produce topics that approximated common syndrome groupings, but was unable to detect emerging keyword patterns.

In the "dynamic" approach, we recalculated the set of 25 topics every day using the most recent two weeks of data, and in the "incremental" approach, we not only used the static topics but also learned five

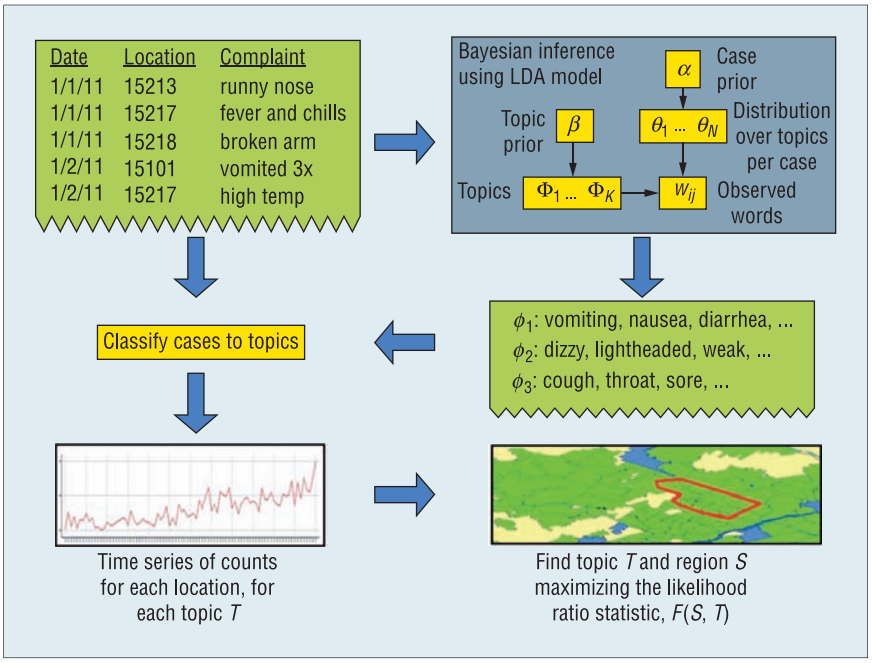

Figure 1. The semantic scan statistic learns a set of topics from the data using Latent Dirichlet Allocation, classifies each case into the most likely topic(s), and then maximizes a likelihood ratio statistic *F(S, T)* over all topics *T* and all space-time regions *S*.

additional topics from the recent data. These "emerging topics" were constrained to differ substantially from the static topics by recalculating the LDA models using all 30 topics while keeping the static topic distributions fixed. The dynamic and incremental semantic scan methods dramatically outperformed the prodrome-based method for synthetically generated "novel" outbreaks and for symptoms that did not correspond to any of the pre-existing prodrome types, reducing the average time to detect (for a fixed false-positive rate of one/month) from 11 days to five.

An additional benefit of such methods is that they can characterize the novel outbreak, not only by identifying the affected spatial area and time period but also by providing the set of emerging keywords. For example, for a simulated "green nose" outbreak, the semantic scan correctly identified an emerging spatial cluster of a learned topic distribution with highest probabilities corresponding to words such as "green," "greenish," "colored," "nose," and "nasal." It should be noted, however, that the detection time of the text-based method was one to two days slower than the prodrome-based method for simulated outbreaks corresponding to known prodrome types, suggesting that the use of pre-existing syndrome definitions is still beneficial for the majority of typically occurring diseases. This result is not surprising, since the prodrome definitions were manually generated and incorporate existing medical knowledge about common classes of disease symptoms. Nevertheless, our work demonstrates that text-based detection methods can supplement the prodrome-based method in order to effectively detect novel disease outbreaks. They can also be useful in the many cases where existing prodrome definitions are unavailable or inadequate, including surveillance of online data, multilingual surveillance systems, monitoring low-resource areas, and "drop-in" surveillance for major events.

Scaling Up Disease Surveillance

A second major challenge of future disease surveillance systems will be dealing with the scale of the data. Many potential sources of health information contain massive amounts of data, and we wish to detect emerging disease patterns in such datasets in near real time. Additionally, future systems will integrate information from a large number of data sources to achieve more timely detection and improved situational awareness. Thus, future detection systems will require new, computationally efficient algorithms to scale up to the huge amounts of data.

One useful class of methods for solving large-scale detection problems is what we call "subset scanning:" we treat the pattern-detection problem as a search over subsets of the data, finding those subsets which maximize some measure of interestingness or anomalousness (a "score function"), often subject to additional constraints. For example, when finding emerging clusters of disease cases in space-time data, we can identify subsets of spatial locations with higher-than-expected counts for some subset of the monitored data streams that also satisfy constraints on spatial proximity.[11,12] For more general datasets, we might wish to identify a subset of records that is self-similar and for which some subset of attributes is anomalous.[13]

In each case, we rely on a surprising and remarkably useful property of many likelihood ratio statistics that we call *linear time subset scanning* or LTSS.[11] It is possible to efficiently maximize any score function satisfying the LTSS property over subsets of the data by defining another function (the "priority" of a data record), sorting the data records by priority, and then only evaluating subsets consisting of the top k highest-priority records. For a dataset with N data records, this reduces the number of function evaluations required for optimization from 2^N to N, while still correctly identifying the most anomalous (highest-scoring) subset. In practice, this allows us to solve detection problems in milliseconds that would have required millions of years for exhaustive computation.

While LTSS provides an efficient and exact solution to the optimization problem of finding the most interesting unconstrained subset, incorporating relevant constraints such as spatial proximity or self-similarity of the detected pattern requires additional steps. One simple approach is to define the "local neighborhood" of each spatial location or each data record, use LTSS to efficiently optimize over each neighborhood, and then identify the most interesting subset over all such neighborhoods. A more challenging case is when we start with an underlying graph or network structure and wish to constrain our search to connected subgraphs. The resulting algorithm[14] still requires exponential time in the worst case, but it can scale to graphs of several hundred nodes, an order of magnitude larger than previous approaches.[15] Finally, we can use iteration to optimize over the most anomalous subsets of records and attributes (or locations and data streams). We alternate between optimizing the score function over subsets of records for the current subset of attributes, and optimizing over subsets of attributes for the

current subset of records, where the LTSS property allows us to perform each optimization step efficiently.[11,12] The resulting algorithms can scale to hundreds of thousands of data records and integrate information from hundreds of data streams.

We consider these approaches only a first step toward addressing truly societal-scale data such as Internet search queries, user-generated Web content, and location and proximity data from cellular telephones. For these and other massive data sources, even algorithms that scale linearly with the size of the data will be insufficient. Enabling disease surveillance systems to scale to the data-driven public health practices of the future will require other techniques such as approximate subset scan algorithms that can sample aggregate data at multiple resolutions. ■

Acknowledgments

This article is based on the author's invited talk, "Research Challenges for Biosurveillance," presented at the 2010 Annual Conference of the International Society for Disease Surveillance (ISDS). The author wishes to thank the ISDS community, as well as the members of his Event and Pattern Detection Laboratory, for their comments and feedback. This work was partially supported by the National Science Foundation under grants IIS-0916345, IIS-0911032, and IIS-0953330. All views are solely those of the author, and have not been endorsed by NSF.

References

1. M. Kulldorff, "Prospective Time-Periodic Geographical Disease Surveillance Using a Scan Statistic," *J. Royal Statistical Society A*, vol. 164, 2001, pp. 61–72.
2. M. Kulldorff et al., "Multivariate Scan Statistics for Disease Surveillance," *Statistics in Medicine*, vol. 26, no. 8, 2007, pp. 1824–1833.
3. G.F. Cooper et al., "Bayesian Biosurveillance of Disease Outbreaks," *Proc. 20th Conf. Uncertainty in Artificial Intelligence, ACM*, 2004, pp. 94–103.
4. D.B. Neill and G.F. Cooper, "A Multivariate Bayesian Scan Statistic for Early Event Detection and Characterization," *Machine Learning*, vol. 79, 2010, pp. 261–282.
5. W.R. Hogan et al., "The Bayesian Aerosol Release Detector: An Algorithm for Detecting and Characterizing Outbreaks Caused by an Atmospheric Release of Bacillus Anthracis," *Statistics in Medicine*, vol. 26, 2007, pp. 5225–5252.
6. T. Fawcett and F. Provost, "Activity Monitoring: Noticing Interesting Changes in Behavior," *Proc. 5th Int'l Conf. Knowledge Discovery and Data Mining* (KDD 99), *ACM*, 1999, pp. 53–62.
7. M. Wagner et al., "A National Retail Data Monitor for Public Health Surveillance," *Morbidity and Mortality Weekly Report*, vol. 53, 2004, pp. 40–42.
8. W.W. Chapman et al., "Classifying Free–Text Chief Complaints into Syndromic Categories with Natural Language Processing," *Artificial Intelligence in Medicine*, vol. 33, no. 1, 2005, pp. 31–40.
9. Y. Liu and D.B. Neill, "Detecting Previously Unseen Outbreaks with Novel Symptom Patterns," *Emerging Health Threats J.*, vol. 4, 2011, in press.
10. D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, 2003, pp. 993–1022.
11. D.B. Neill, "Fast Subset Scan for Spatial Pattern Detection," *J. Royal Statistical Society B*, 2011, to be published.
12. D.B. Neill, E. McFowland III, and H. Zheng, "Fast Subset Scan for Multivariate Spatial Biosurveillance," *Emerging Health Threats J.*, vol. 4, 2011, p. s42.
13. E. McFowland III, S. Speakman, and D.B. Neill, "Fast Generalized Subset Scan for Anomalous Pattern Detection," *Proc. INFORMS Annual Conf.*, 2010.
14. S. Speakman and D.B. Neill, "Fast Graph Scan for Scalable Detection of Arbitrary Connected Clusters," *Proc. Int'l Soc. Disease Surveillance Annual Conf.*, 2009.
15. T. Tango and K. Takahashi, "A Flexibly Shaped Spatial Scan Statistic for Detecting Clusters," *Int'l J. Health Geographics*, vol. 4, 2005, p. 11.