ISDS 2019 Conference Abstracts

Streamlined Development of Analytic Fusion Capability for Health Surveillance

Susama Agarwala, Howard Burkom, Daniel Wernig

REDD, Johns Hopkins Applied Physics Lab, Washington, District of Columbia, United States

Objective

Our project goal is to enhance the capability of automating health surveillance[MOU1] by US Department of Defense (DoD) epidemiologists. We employ software tools that build and train Bayesian networks (BNs) to facilitate the development of analytic fusion of multiple, disparate data sources comprising both syndromic and diagnostic data streams for rapid estimation of overall levels of concern for potential disease outbreaks. Working with previously developed heuristic BNs, we evaluate the ability of machine learning algorithms to detect outbreaks with greater accuracy. We use historical training data on the ability to detect outbreaks of influenza-like illness (ILI).

Introduction

The motivation for this project is to provide greater situational awareness to DoD epidemiologists monitoring the health of military personnel and their dependents. An increasing number of data sources of varying clinical specificity and timeliness are available to the staff. The challenge is to integrate all the information for a coherent, up-to-date view of population health.

Developers at the Johns Hopkins Applied Physics Laboratory, in collaboration with medical epidemiologists at the Armed Forces Health Surveillance Branch, previously designed a multivariate decision support tool to add to the DoD implementation of the Electronic Surveillance System for Early Notification of Community-Based Epidemics (ESSENCE). Data sources included clinical encounter records including free-text chief complaints, filled prescription records, and laboratory test orders and results. Filtered data streams were derived from these sources for daily monitoring, and alerting algorithms were customized and applied to the resulting time series. We built BNs to derive overall levels of concern from the collection of data streams and algorithm outputs to derive, in the form of daily fusion alerts, the overall level of various outbreak concerns. Visualizations made apparent which data features accounted for these concerns, including drill-down to the level of patient record details. Advantages of the BN approach are this transparency and the capacity for assessments using incomplete data and incorporating novel and report-based data streams. The need for such fusion was nearly unanimous in a global survey of public health epidemiologists [1].

Our proof-of-concept system based on commercial BN software was well received by a cross-section of DoD health monitors. The new software tools we apply in this project use freely available R packages which provide more comprehensive tools for BN training and development. These results will allow us to improve the analytic fusion abilities of DoD ESSENCE, as well as in civilian surveillance systems

Our testing procedures and results are presented below.

Methods

We employ a 3.75-year dataset (2006-2010) with information from 502 US medical treatment facilities including 289 hospitals. Our data include time series of daily counts and alerting algorithm outputs from each facility for syndrome groups based on a) chief complaints and diagnosis codes from clinic visits, b) groups of laboratory test orders and influenza test results, and c) selected groups of filled prescriptions. For each facility group, the challenge is to combine these data streams into a daily assessment of levels of concern for an ILI outbreak. The software developed in this project facilitates the formation, training, and testing of BNs for outbreak alerting based on the datasets above.

Underlying each BN is a directed acyclic graph whose leaf nodes represent discrete states of concern for each data stream, ranging from general streams such as ILI-related chief complaints to specific ones such as positive flu test results. The states are derived from both daily stream counts and alerting algorithm outputs. Internal nodes represent mid-level combinations of indicators, such

as ILI concern based only on clinic data, and parent nodes which represent the calculated level of concern based on all data sources. The connectivity between the nodes, and the orientation of the edges of the graph are determined by the heuristic relationships adopted in previous projects [2]. Each node is associated with a conditional probability table (CPT). Training a BN requires state assignments for every node in the underlying graph, which must be obtained from a trusted source. These assignments can generate CPTs at each node of the BN such that, when given only a set of evidence nodes, the levels of concern can be propagated up the network to provide the desired levels of concern at the decision nodes.

Truth data for ILI outbreaks comes from two sources: 33 documented outbreaks with dates supplied by DoD surveillance reports or media articles and, because many modest-sized outbreaks are unreported, 81 unconfirmed data-derived events with algorithmic alerts across multiple data sources. We use data from the more numerous unconfirmed events to train the BNs.

To avoid commercial software constraints, we use the free R package *gRain* [3] to create, train, and test the BNs. We test multiple BNs for multivariate ILI outbreak detection, all based on the same nodal structure with 18 parent, intermediate, and leaf nodes. Candidate BNs are created and trained using either a) CPTs determined with a multivariate stochastic search in the previous project, augmented with ground truth data or b) a heuristic lookup table of state combinations.

Results

Table 1 shows the high odds ratios calculated for candidate BNs. These statistics are calculated for decision node outputs for event vs non-event dates in the truth data with the constraint that every event is detected for at least one date. The machine learning advantage from the training data is evident from comparing the two rows. To show the advantage of fusing data sources, Table 2 gives analogous odds ratios based on single-stream alerting algorithms. Aside from the lower detection statistics, single streams offer no corroboration of statistical alerts.

Conclusions

Analytic fusion is essential for the efficient, timely use of a growing collection of complex, streaming information by a limited workforce of human health monitors. This project builds upon previous fusion capability for automated health surveillance by expediting new development and facilitating software implementation through open source tools. The detection results indicate a significant advantage in both sensitivity and alert rates of automated systems achievable with machine learning. Moreover, if basic challenges of multivariate data acquisition and determination of truth datasets for supervised learning can be met, further improvements are likely using BN structure discovery methods as well as other machine learning approaches.

References

1. Hopkins RS, Tong CC, Burkom HS, et al. 2017. A Practitioner-Driven Research Agenda for Syndromic Surveillance. *Public Health Rep*. 132(1_suppl), 116S-126S. PubMed https://doi.org/10.1177/0033354917709784
2. Burkom HS, Elbert Y, Ramac-Thomas L, Cuellar C, Hung V. 2013. Refinement of a Population-Based Bayesian Network for Fusion of Health Surveillance Data. *Online J Public Health Inform*. 5(1), e6. https://doi.org/10.5210/ojphi.v5i1.4413
3. Højsgaard S. 2012. Graphical Independence Networks with the gRain Package for R. *J Stat Softw*. 46(10). doi:10.18637/jss.v046.i10.

Table 1. Fusion Odds Ratios

Intermediate/Parent Nodes	ILI Outbreak	Influenza Outbreak	Severe ILI Outbreak
Odds Ratio (CPTs made by hand)	75.90951	17.00181	61.52413
Odds Ratio (CPTs modified with Training)	107.2841	61.42208	63.97279

Table 2. Single Stream Odds Ratios

	CAPER ACD-9 Syndrome	CAPER Chief Complaint	Influenza Antivirals	Influenza Lab Test
Median Odds Ratio	3.9	7.5	4.7	2.1