Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2008)
J. Li, D. Aleman, R. Sikora, eds.

AN ASYMPTOTICALLY OPTIMAL STRATEGY IN SEQUENTIAL CHANGE DETECTION AND IDENTIFICATION APPLIED TO PROBLEMS IN BIOSURVEILLANCE

Savas Dayanik, Warren Powell, and Kazutoshi Yamazaki
Operations Research and Financial Engineering
Princeton University
Princeton, NJ
[sdayanik, powell, kyamazak]@princeton.edu

Abstract

We describe a statistical model of detecting and identifying an outbreak of bioagents. Based on a sequence of observations, one needs to detect the outbreak of disease and identify its cause. To maintain public safety, it is important to detect it at the earliest, and at the same time, identify its cause as accurately as possible in order to take the most appropriate countermeasures. This boils down to solving optimally the tradeoff between the detection delay cost and the false alarm and misidentification costs to the society.

We focus on developing an early detection system of anthrax, which is considered to be the most likely weapon in biological warfare. Despite its importance, early detection of anthrax is difficult because its initial symptoms resemble a cold. The detection system therefore needs to differentiate anthrax from flu and other bioagents that cause similar symptoms.

We propose a new sequential decision strategy which triggers an alarm when the posterior probability of a certain type of change exceeds some fixed threshold for the first time, show its asymptotic optimality under the Bayesian formulation, and verify the result numerically.

Keywords: biosurveillance of disease outbreaks, anthrax, change-point detection, sequential hypothesis testing, asymptotic theory.

Introduction

Biosurveillance detects and characterizes outbreaks of disease, whether natural or bioterrorist-induced. It systematically collects and analyzes data in order to take the most appropriate response and treatment.

Biosurveillance has two conflicting goals: minimizing detection delay and false-alarm/misidentification probabilities. In order to maintain public safety and minimize casualties, it is important to detect outbreaks *at the earliest*. At the same time, one needs to identify *with high accuracy* its cause such as its disease name, causative biological

agent, source and route of transmission in order to take the most appropriate countermeasures.

In this paper, we develop a simple and near-optimal algorithm of detecting the outbreak of bioagents, focusing on anthrax, which is considered to be the most likely weapon in biological warfare. The best defense against biological attacks is an early warning system. Vaccination, for example, is not a practical solution because of the large number of bioagents that could be used in bioterrorism. An early detection system is thought to be the only practical solution against bioterrorism.

Early detection of inhalational anthrax is important and at the same time difficult. If left untreated, it will usually result in death. Moreover, if the disease is fatal, a person's body infected with anthrax bacilli becomes a potential source of infection to others. The difficulty of early detection mainly comes from the fact that its initial symptoms resemble a common cold. For this reason, we need a mechanism for efficiently differentiating anthrax from flu and other bioagents that cause similar symptoms.

Data commonly used in biosurveillance includes visits to emergency medical services, 911 calls and ambulance dispatch records, laboratory and mortality records, veterinary reports, and school or work absence records [1].

Grocery and over-the-counter (OTC) medication sales, such as cough medicine, are potentially used in detecting anthrax [2]. According to [2], they have three main advantages; 1) these datasets are typically very large and rich, including information on each purchased item and in many cases include customer information, and are also available on a more frequent basis such as daily and even hourly; 2) the outbreak footprint exist in these data earlier than in medical or public health data, because of self treatment that people usually pursue before seeking medical assistance; 3) specific symptoms experienced by purchasers may be inferred at a relatively early stage of the onset of the disease.

Although cough medicine sales data themselves are not sufficient in differentiating anthrax from flu, they can be complemented by the sales data of tissues because anthrax does not lead to a runny nose. For this reason, we use the sales data of both cough medicine and tissues for anthrax detection.

We want to solve the tradeoff between minimizing the delay before making the call that a disease has entered the population (the detection delay) and maximizing the accuracy (both in the call that an outbreak has occurred, and the cause of the outbreak). We model the surveillance of disease outbreaks using a combination of quickest change-point detection and sequential hypothesis testing. We focus on the Bayesian formulation where one wants to minimize a Bayes risk which consists of a linear combination of 1) expected detection delay, 2) false alarm probability and 3) misidentification probability.

Although we focus on the detection of anthrax in this paper, it is important to note that the strategy we will propose is implementable in more complicated cases with higher number of disease types. We will give a mathematical model and verify its near-optimality by showing its asymptotic optimality together with numerical results.

Mathematical Model:

Let θ be the outbreak time and μ be its cause, either "(A) anthrax" or "(F) flu", both of which are unobservable and therefore are random variables. At time $t = 1,2, \ldots$, one observes

$$X_t = (X_t^{(C)}, X_t^{(T)})$$

where

$X_t^{(C)} :=$ sales of cough medicine on day t, and
$X_t^{(T)} :=$ sales of tissues on day t.

We initially have rough estimates (*prior* distributions) about θ and μ, but they change over time as captured by the belief process $\Pi_t = (\Pi_t^{(0)}, \Pi_t^{(A)}, \Pi_t^{(F)})$ where

$\Pi_t^{(0)} := P(no\ outbreak\ before\ t \mid X_1, \ldots, X_t) = P(\theta > t \mid X_1, \ldots, X_t)$,
$\Pi_t^{(A)} := P(outbeak\ of\ anthrax\ before\ t \mid X_1, \ldots, X_t) = P(\theta \le t, \mu = A \mid X_1, \ldots, X_t)$,
$\Pi_t^{(F)} := P(outbreak\ of\ flu\ before\ t \mid X_1, \ldots, X_t) = P(\theta \le t, \mu = F \mid X_1, \ldots, X_t)$.

Conditionally given θ and μ, the observation process X is i.i.d. and F_0-distributed before θ and F_μ-distributed thereafter. We can obtain F_0, F_A, F_F using historical data.

We want to find a strategy, a pair of a detection time τ and a declared cause d, so that the following costs are minimized;

1) $ED(\tau, d) := expected\ delay = E[\tau - \theta]$,
2) $FA(\tau, d) := false\ alarm\ probability = P(\theta > \tau)$,
3) $MI(\tau, d) := misidentification\ probability = P(d \neq \mu, \theta \le \tau)$.

The Bayesian formulation minimizes a Bayes risk, a linear combination of the abovementioned costs,

$$R(\tau, d) := \alpha ED(\tau, d) + \beta FA(\tau, d) + \gamma MI(\tau, d),$$

for some fixed parameters α, β and γ.

A strategy (τ, d) has to be *well-defined*. Because at each time t, one needs to decide either to stop or continue based on the information obtained up until time t, whether $\{\tau > t\}$ or $\{\tau \le t\}$ must be determined by $X_1, \ldots, X_t$. Moreover, because when stopped at τ, one needs to declare its cause based on the information up until τ, d has to be a deterministic function of τ and $X_1, \ldots, X_\tau$.

It is intuitive that one should trigger an alarm when either $\Pi_t^{(A)}$ or $\Pi_t^{(F)}$ gets sufficiently high. The problem will be an optimal stopping problem where one should trigger an alarm when Π enters some region for the first time.

Although an optimal strategy that minimizes the Bayes risk $R(\tau, d)$ exists [3], our focus in this paper is to obtain a simpler strategy that can be also implemented in more complicated cases. We propose a strategy, which triggers an alarm when the posterior probability of a certain type of change exceeds some fixed threshold for the first time.

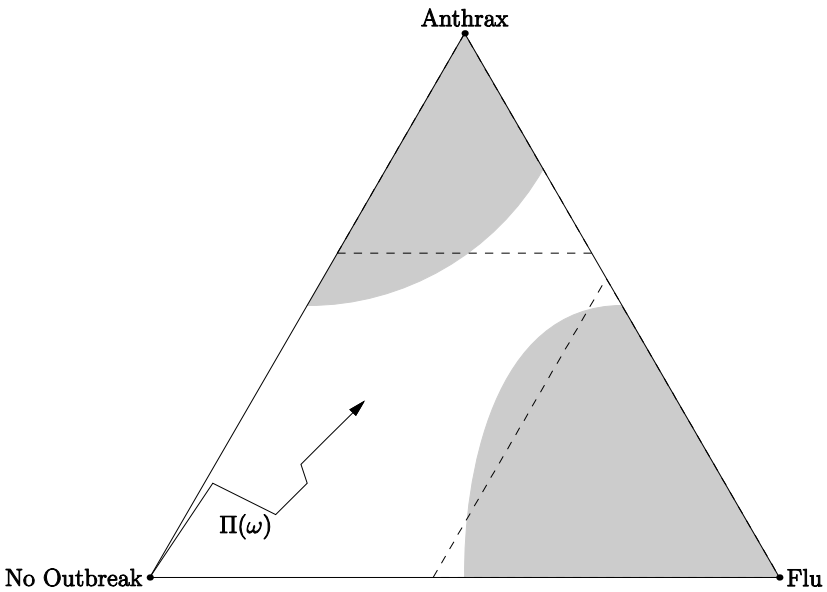

Figure 1: Illustrations of (i) the optimal stopping region indicated by the gray regions and (ii) the stopping region of the proposed strategy indicated by the dotted triangles.

Figure 1 illustrates graphically the proposed strategy in comparison with the optimal strategy. The process starts in the lower left-hand corner, where we are in the state of "no outbreak." As observations are made, the process progresses through the white region, and stays in the white region as long as we are not willing to make the call that an outbreak has occurred. If the process falls in the shaded area at the top, then this means that we have an outbreak of anthrax. If it falls in the shaded area at the lower right, then it is an outbreak of flu. These shaded areas, properly designed, represent an optimal policy in terms of minimizing the delay in making a call, and making the right call.

The problem with the gray regions is that they are difficult to compute. Instead, we propose to use the triangles shown in the two corners. Our algorithmic challenge, then, is designing the triangles to minimize the Bayes risk. The size of these triangles represents a stopping and selection policy.

Notice that the figure will be three-dimensional if there are three possible causes (e.g. SARS in addition to anthrax and flu), and the problem will be significantly more difficult than the two-dimensional case. In more realistic applications in biosurveillance, specific characteristics of outbreaks, such as causative biological agent and source and route of transmission, must be identified in order to take appropriate countermeasures. The problem therefore tends to be high-dimensional and obtaining the optimal strategy is practically impossible. The complexity of the proposed strategy, however, is not affected by the dimensionality.

Practical Solution

We define strategy (τ_A, d_A) for some fixed constants $A = \{A_i > 0; i = A, F\}$ by

$$\tau_A := \inf\left\{n \geq 1: \Pi_n^{(i)} > \frac{1}{1+A_i} \text{ for some } i = A, F\right\} \text{ and } d_A \in argmax_i \Pi_{\tau_A}^{(i)}.$$

Namely, it triggers an alarm when the posterior probability of a certain type of change exceeds $1/(1 + A_i)$ for the first time. The values of A determine the size of the dotted triangles in Figure 1, and have to be determined in order to implement the strategy.

The smaller the unit cost of detection delay α is, the longer one should wait, i.e., $\tau_A \uparrow \infty$ as $\alpha \downarrow 0$. Therefore, A should decrease to 0 as $\alpha \downarrow 0$. In order to avoid technical details, we omit the formula of A in terms of α. However, it can be shown that if the values of A are chosen in a certain way as a function of α, we can obtain the asymptotic optimality

$$\frac{\inf_{(\tau,d)} R(\tau,d)}{R(\tau_A,d_A)} \to 1 \quad \text{as } \alpha \downarrow 0.$$

We refer the reader to [4] for the formula for determining the value of A; our model generalizes their model and the formulae are similar. The asymptotic optimality result can be shown by (nonlinear) renewal theory-type arguments; see [5,6]. The values of A can be obtained analytically, and no complicated computation is needed to implement the strategy. For other asymptotically optimal strategies, see [7,8] for the sequential hypothesis testing problem and [9,10] for the quickest detection problem.

Numerical Example

The asymptotic optimality result indicates that it is near-optimal when the value of unit detection delay α is small. This section verifies numerically that it is also near-optimal even when it is not small.

The prior distribution of θ is assumed to be geometric with success probability $p = 0.01$,

$$P(\theta = t) = (1-p)^{t-1}p$$

so that the average interarrival time of outbreaks is 100 days, and that of μ is assumed to be

$$P(\mu = A) = 0.1 \text{ and } P(\mu = F) = 0.9.$$

We assume that given θ and μ, the sales of cough medicine and tissues X are conditionally mutually independent and are normally distributed with common variance 1 million dollars and mean $\lambda_0^{(C)}$ and $\lambda_0^{(T)}$ before θ and with $\lambda_\mu^{(C)}$ and $\lambda_\mu^{(T)}$ thereafter in units of thousands of dollars with

$$\begin{aligned}
\lambda_A^{(C)} &= \lambda_0^{(C)} + 1.0, \\
\lambda_F^{(C)} &= \lambda_0^{(C)} + 1.0, \\
\lambda_A^{(T)} &= \lambda_0^{(T)} + 0.0, \\
\lambda_F^{(T)} &= \lambda_0^{(T)} + 0.5.
\end{aligned}$$

Table 2 compares the strategy (τ_A, d_A) and the optimal strategy (τ^*, d^*) with various values of unit cost of detection delay α and common unit costs of false-alarm/misidentification $\beta = \gamma = 1.0$. We can see the asymptotic optimality of (τ_A, d_A),

i.e., the ratio converges to 1 as $\alpha \downarrow 0$ as listed in the last column. Moreover, we can see from the table that it is near-optimal even when α is not small.

α	$R(\tau^*, d^*)$	$R(\tau_A, d_A)$	$R(\tau_A, d_A)/R(\tau^*, d^*)$
0.020	0.2896362	0.30860624	1.065496
0.015	0.2422770	0.25750238	1.062843
0.010	0.1869979	0.19718571	1.054481
0.005	0.1203246	0.12367423	1.027838

Figure2: Comparison between (τ^*, d^*) and (τ_A, d_A): their Bayes risks and the ratios.

Reference and Citation

1. Fienberg, S. and Shmueli, G., Statistical issues and challenges associated with rapid detection of bio-terrorist attacks, Statist. Med. 2005: 24:513-52
2. Goldenberg, A., Shmueli, G., Caruana, R., Fienberg, S. (2002): Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales.
3. Dayanik, S., Goulding, C. and Poor, V., Bayesian Sequential Change Diagnosis, Mathematics of Operations Research. 2008: 33-2:475-496
4. Baum, C.W. and Veeravalli, V.V. (1994): A sequential procedure for multihypothesis testing. IEEE Trans.Inform. Theory, 6 , 1994–2007.
5. Woodroofe, M. (1982): Nonlinear Renewal Theory in Sequential Analysis. SIAM Philadelphia, Pa.
6. Siegmund, D. (1985): Sequential Analysis. Springer Series in Statistics. Springer-Verlag, New York.
7. Dragalin, V. P., Tartakovsky, A. G., and Veeravalli, V.V. (1999): Multihypothesis sequential probability ratio tests I. Asymptotic optimality. IEEE Trans. Inform. Theory, 45(7), 2448–2461.
8. Dragalin, V. P., Tartakovsky, A. G., and Veeravalli, V.V. (1999): Multihypothesis sequential probability ratio tests II. Accurate asymptotic expansions for the expected sample size. IEEE Trans. Inform. Theory,46(4), 1336–1383.
9. Tartakovsky, A.G. and Veeravalli, V.V. (2004): Change-point detection in multichannel and distributed systems. Applied Sequential Methodologies, 173, 339–370.
10. Baron, M. and Tartakovsky, A. G. (2006): Asymptotic optimality of change-point detection schemes in general continuous-time models. Sequential Anal., 25(3), 257-296.