Detectors should be characterized by likelihood ratios, not posterior probabilities

Peter Szolovits, MIT

June 18, 1999
Slightly revised February 1, 2000.

Thanks to Kathy Laskey and Hamish Fraser for suggestions.

This is a short note, I hope, to lay out a case for how best to estimate the useful certainty of an intrusion detector. It is based on experience gained from probabilistic modeling in medical and other diagnosis domains. The note was triggered by a presentation at the TIC in June, 1999 by Dan Schnackenberg, in which he showed a CIDF report containing a line "Certainty: 60", which set off a debate about the meaning of this numerical estimate.

Dan's characterization of the meaning of "60" is that it is the detection algorithm designer's subjective estimate that when this detector goes off, in 60% of the cases, an attack of type H (*hypothesis*) has actually occurred. This note argues that the designer cannot (or at least should not) be in a position to make such an estimate, and to suggest instead that we ask the designer to estimate a different number, what is called the *likelihood ratio* or *conditional odds* for the detector. To explain and justify this, it will be helpful to introduce some notation and basic ideas.

Bayesian reasoning assumes that one can estimate the *a priori* (before observations are made) probability of some events of interest. For example, that we can estimate, before we even try to detect anything, that H attacks will happen in the scope of our interest, say once every hundred hours. Another way of saying this might be that, for any one-hour period, there is approximately a 1% chance of such an attack (forgetting Poisson). This rate will, naturally, vary greatly with circumstances. For example, whether our locale is on a public internet or a behind a carefully-controlled firewall will affect it, as will our understanding of whether there are likely to be people interested in attacking us at this particular time with the means to do so. For example, Kathy Laskey gave an example where on on-campus student organization announced its intent to bring down engineering computers at George Mason; such an event would make the prior probability of an attack much higher, even before we begin to actually try to run any of our monitors.

Now suppose that we begin to run our monitoring system, and within an hour the detector for attack H goes off. According to Bayes' Rule, the proper way to estimate the likelihood that we have actually detected an attack is, informally, to combine our a priori expectation of this type of attack with the reliability of our detector. The formula is

$$P(H \mid D) = \frac{P(H)P(D \mid H)}{P(H)P(D \mid H) + P(\overline{H})P(D \mid \overline{H})}$$

where $P(H)$ is the *a priori* probability of the attack H, $P(\overline{H}) = 1 - P(H)$ is the probability that H has not occurred, and $P(D \mid H)$ and $P(D \mid \overline{H})$ are the probabilities that detector D would go off in the face of an H attack and the probability that it would go off in the absence of such an attack. Fortunately, in practice all that is needed for computation is the *likelihood ratio* of these last two numbers (see below):

$$L(D \mid H) = \frac{P(D \mid H)}{P(D \mid \overline{H})}$$

which is the estimate of *how much more likely is this detector to go off when it should than when it should not!*

This ratio is a good measure of the reliability of the detector, and should be independent of the actual rate at which H attacks really happen. It is in fact the ratio of the detector's true positive rate to its false positive rate, and will be greater than 1 for useful detectors. For highly reliable detectors, the ratio will be large.

Probabilities can be computed using *odds* instead, which leads to a particularly simple form for incorporating evidence and *a priori* belief. Thus, if we define

$$O(H) = \frac{P(H)}{P(\overline{H})} = \frac{P(H)}{1 - P(H)}$$

then Bayes' Rule may be written as

$$O(H \mid D) = O(H)L(D \mid H)$$

In other words, the posterior odds that an H attack occurs are just the product of the prior odds times the likelihood ratio for the detector. Thus, a likelihood ratio greater than one will increase the posterior odds; this is true just in case the detector is more likely to go off when H occurs than when it does not.

This formulation cleanly separates two issues that matter in figuring out how strongly you should believe that an H attack has actually happened after you observe a positive response from the detector. The likelihood ratio characterizes the quality of the detector, and the prior odds characterize the prior risk of this attack. Together, these estimate the posterior risk.

I claim that the writer of a detector should in fact be happier to estimate the likelihood ratio than to try directly to estimate Dan's characterization, the posterior probability. This is because the detector writer has no idea in what circumstances the detector will actually be deployed and run, and thus has no ability to factor in all the circumstantial issues that will determine the prior probability of attack. (See note A, below.)

The other main reason for this approach to computing certainty is that Dan's approach provides no guidance on how to combine certainties of evidence from multiple sources, because each of them has already incorporated a priori judgments into his or her estimate. (See note B, below.) By contrast, the likelihood ratio approach allows us to combine the effects of multiple detectors simply. If we can assume that each of detectors D_1 and D_2 operates independently, so their results depend only on whether *H* is actually happening but not on each other, then the posterior probability of *H* after both detectors go off is

$$O(H \mid D) = O(H)L(D_1 \mid H)L(D_2 \mid H)$$

This correctly takes the prior probability of *H* into account once, and properly combines the evidential impact of both detectors.

I want to emphasize that the lessons outlined above have been painfully learned by twenty years of bad experience with other alternatives. In particular, the design that asks a test creator to estimate the posterior probability of what the test is trying to detect has consistently led to errors in other fields, and thus should be avoided in this domain.

Note A—The Effects of Prior Probabilities

Consider an intrusion detector that has a likelihood ratio for detecting intrusions of 100:1. In other words, it is an exceptionally good detector, and is 100 times as likely to go off when an intrusion occurs than to go off in the absence of one. Now consider two scenarios, suggested by Kathy Laskey based on a bit of GMU history:

1. A student who normally connects via a dial-up modem, is given a dynamic IP address, and has slow connectivity, observes that the intrusion detector on his computer alarms. If we assume that a person with such a connection has 10,000:1 odds against an intrusion (about 0.01% chance), the posterior odds after seeing the detector alarm are about 100:1 against an actual intrusion, or about a 1% chance of intrusion.
2. After an announcement by school hackers angered over IT policies that they plan to bring down the computer system, the same detector as in scenario one alarms on a machine belonging to a member of the IT staff. If, in the context of the immediate threat, we assume that there are 1:2 odds that any particular IT staff computer will be attacked (i.e., about 33%), an alarm by the same detector should raise our belief in an attack to odds of 50:1, or about 98%.

Thus, the same alarm under different circumstances can lead to extremely different conclusions. The reasonableness of this can be seen by considering the following *contingency tables*, which show the number of times that we might expect an attack if we repeat each scenario 1,000,000 times and how often we would expect to see an alarm under those circumstances. The assumptions are as given in the text above. To get the likelihood ratio of 100:1 for the detector, we assume that $P(D \mid H) = 1.0$ and $P(D \mid \overline{H}) = 0.01$.

Dial-up scenario			
	Detector alarms	Does not alarm	
Actual attack	100	0	100
No attack	9,999	989,901	999,900
	10,099	989,901	1,000,000

IT Staff scenario			
	Detector alarms	Does not alarm	
Actual attack	333,333	0	333,333
No attack	6,667	660,000	666,667
	340,000	660,000	1,000,000

In the dial-up scenario, attacks occur very rarely, so even with a perfectly sensitive detector, most alarms are false positives from the overwhelming number of times when no attack is taking place. Only 100 times in 10,099 (about 1%) is there an actual intrusion when the detector goes off. In the staff scenario, by contrast, most occurrences of alarms (333,333/340,000) come from actual attacks, hence the probability that an attack actually occurred after the detector is triggered is about 98%. Clearly, the identical detector deployed in different circumstances leads to radically different results. No single "Certainty: xxx" judgment can capture this real phenomenon.

Public health workers in fact understand that choosing a test (detector) to use can very much depend on the prevalence (a priori likelihood) of the condition being sought. If a disease is rare, for example, then lowering the false positive rate of a test has a much greater impact on its ability to recognize disease than lowering its false negative rate. For conditions that are a priori more likely, just the opposite is the case. Contingency tables such as the ones above will demonstrate this point. Thus, the likelihood that a condition being studied is present will determine the nature of the most appropriate detectors that should be used to try to confirm or deny its presence.

Note B—Combining Posterior Judgments

Suppose you go to a doctor with a serious cough and chest pain, and your physician is worried that you have either lung cancer or valley fever (a fungal infection prevalent in California's San Joaquin Valley).. He sends you to get an x-ray, and the radiologist reports back that he believes you have lung cancer with probability 80% and valley fever with probability 20%. Independently, the lab analyzes your blood sample and reports to your doctor that you are 40% likely to have tuberculosis and 60% lung cancer. What is your doctor to make of these results? How can he combine them into one aggregate judgment? How can he incorporate his subjective judgment of what ails you, based on investigating your recent medical history, doing a physical examination, etc.? This formulation, which is equivalent to Dan's formulation of the detector problem, has no sensible solution, because both the radiologist and the lab yield posteriors, which happen to be inconsistent.

If, instead, radiology reported a likelihood ratio – how much more likely would one be to see this particular x-ray in someone with lung cancer than valley fever – and the lab did likewise, then your physician could use the odds formula above to combine these measures of the impact of each test with his own assessment of the likelihood of your disease possibilities based on everything else he knows to yield the best overall estimate. Thus, for example, if he knows you have been living in or traveling in the right part of California, his prior estimate for valley fever will be much higher than if you have not been there. Trying to make this sort of adjustment is rightly his job, not that of the radiologist or lab tech, who may know nothing of your travel history. Similarly, if the lab result raises the possibility of tuberculosis seriously, then your physician needs to estimate how likely or unlikely you are to have been exposed to TB, along with valley fever and lung cancer. Unless the evidence is absolutely certain, the test cannot make the decision on its own.