Note on Bayesian Screening of Possibly Rare Features for Predictive Value

William H. Press

February 16, 2005

1 Introduction

In the situation of interest, we want to classify an object into one of two categories, call them A and B, on the basis of the presence or absence of multiple features, call them $F_1, F_2, \ldots$. A specific feature F is characterized by the probabilities with which it occurs (denoted $+F$) or does not occur (denoted $-F$) when the truth is actually A or B, respectively. In other words, we have the ordinary 2×2 contingency table,

	A	B
$+F$	p_A	p_B
$-F$	$1-p_A$	$1-p_B$

(1)

A familiar example in the biomedical literature would have A be the presence of a disease, B its absence, and the F_i's a list of screening questions or laboratory test results. In such cases it is conventional to call p_A the feature's *sensitivity*, $1-p_B$ the feature's *specificity*. A perfect test has $p_A = 1$ and $p_B = 0$.

Also conventional are the definitions of *positive predictive value (+PV)* and *negative predictive value (-PV)*,

$$\begin{aligned} +\mathrm{PV} &= \frac{p_A}{p_A + p_B} \\ -\mathrm{PV} &= \frac{1-p_B}{(1-p_B)+(1-p_A)} \end{aligned} \tag{2}$$

Evidently these are equivalent to the Bayesian *odds ratios*

$$\begin{aligned} +\mathrm{O.R.} &= \frac{p_A}{p_B} \\ -\mathrm{O.R.} &= \frac{1-p_B}{(1-p_A)} \end{aligned} \tag{3}$$

with the relationship being, in each case,

$$\text{O.R.} = \frac{\text{PV}}{1 - \text{PV}} \tag{4}$$

The odds ratios tell us how to adjust our belief in A versus B when we are given the outcome of some specific test or feature, i.e., given either $+F$ or $-F$. The Bayesian "log odds" factors are of course the logarithms of the odds ratios.

2 Selecting Features by Odds Ratio Performance

Specifically, given multiple features that are independent (a big assumption!), and given the prior probabilities $P_0(A)$ and $P_0(B)$, the overall Bayesian posterior probability for the $A : B$ odds ratio can be written as

$$\mathcal{L} \equiv \log\left(\frac{P(A)}{P(B)}\right) = \log\left(\frac{P_0(A)}{P_0(B)}\right) + \sum_i \text{Choose}\left\{F_i, \log\left(\frac{p_{Ai}}{p_{Bi}}\right), \log\left(\frac{1-p_{Ai}}{1-p_{Bi}}\right)\right\} \tag{5}$$

where i now labels the different features, and the Choose function chooses between its second and third arguments depending on whether its first argument is $+F$ or $-F$ respectively. In words: We get the posterior log odds by starting with the prior log odds, then adding either $\log(p_A/p_B)$ when a feature is $+$, or else $\log[(1-p_A)/(1-p_B)]$ when a feature is $-$.

If we have many features available, we don't necessarily need to try them all. We can stop when $\mathcal{L}$ crosses a predetermined threshhold either positive or negative, corresponding (e.g.) to a 99.9% or 0.01% posterior probability for A.

Now suppose that we have an essentially unlimited number of possible features. We want to screen a large number of them on a set of learning data and select the "best" ones to use on production data, or for some further statistical process. This situation occurs if we are looking for features in DNA sequence among an essentially infinite combinatorial set of possibilities.

What is the "figure of merit" by which we should select features in our screen?

Evidently, we want features that will, on average, make $\mathcal{L}$ reach its positive or negative theshhold value as quickly as possible (i.e., with the smallest number of features). Suppose, first, that A is true. Then the expectation value of the change in $\mathcal{L}$ from a given feature is

$$E(\Delta\mathcal{L}|A) = p_A \log\left(\frac{p_A}{p_B}\right) + (1-p_A)\log\left(\frac{1-p_A}{1-p_B}\right) \tag{6}$$

because the result $+F$ occurs with probability p_A, $-F$ with probability $1-p_A$. Using the inequality

$$\log(1+x) \le x \tag{7}$$

it is easy to show that the expression (6) is nonnegative for all p_A and p_B in the range $(0, 1)$:

$$\begin{aligned} & p_A \log\left(\frac{p_A}{p_B}\right) + (1-p_A)\log\left(\frac{1-p_A}{1-p_B}\right) \\ & = -p_A \log\left(1 + \frac{p_B - p_A}{p_A}\right) - (1-p_A)\log\left(1 + \frac{p_A - p_B}{1-p_A}\right) \\ & \geq -p_A \left(\frac{p_B - p_A}{p_A}\right) - (1-p_A)\left(\frac{p_A - p_B}{1-p_A}\right) = 0 \end{aligned} \tag{8}$$

The equality holds only when $p_A = p_B$. This shows that *any* nondegenerate contingency table moves $\mathcal{L}$ in the right direction when A is true; and by symmetry this must be true for B as well. In other words, there are no "deceptive" contingency tables, there are only more- or less-good ones.

Since we face a mixture of A and B cases, our overall figure of merit (F.M.) must average over both contingencies. It is therefore

$$\begin{aligned} \text{F.M.} = P(A) & \left[p_A \log\left(\frac{p_A}{p_B}\right) + (1-p_A)\log\left(\frac{1-p_A}{1-p_B}\right)\right] \\ + P(B) & \left[p_B \log\left(\frac{p_B}{p_A}\right) + (1-p_B)\log\left(\frac{1-p_B}{1-p_A}\right)\right] \end{aligned} \tag{9}$$

Here $P(A)$ and $P(B) \equiv 1 - P(A)$ are the prior probabilities on A and B. These will generally be $P_0(A)$ and $P_0(B)$, respectively. However, there could also be situations in which a second batch of features are to be screened after the application of a first batch, in which case $P(A)$ and $P(B)$ can be posterior probabilities after the first batch.

Another way of looking at the priors $P(A)$ and $P(B)$ is that they define a trade-off curve between false positives and false negatives. As $P(A)$ tends to 1, the figure of merit will select features with low false positive rates (B misclassified as A) over features with low false negative rates (A misclassified as B). As $P(A)$ tends to 0, it is the opposite. In applications with different consequences for false positives and false negatives, $P(A)$ can be (arbitrarily) set accordingly.

In principle, equation (9) solves our problem. We calculate F.M. for some vast number of features, and subsequently use the ones with the largest F.M. In practice two issues arise: (1) The features are not independent. This is a well-known issue, and we have nothing to say about it in this note. (2) We don't actually know the probabilities p_A and p_B for each feature. Rather, we just have counts of how often the feature occurs in learning data of class A and B. This is the issue that we focus on in the rest of this note.

3 Bayesian Estimate of F.M. from Counts

A finite set of learning data replaces the contingency table (1) with a table of counts,

	A	B
$+F$	n	m
$-F$	$N-n$	$M-m$

(10)

Here N is the total number of A's in the learning data, M, the total number of B's. The values of n and m may be small, or even zero. We want to estimate p_A and p_B from these counts, and from priors on their probabilities. Note that when we speak of the probability of a value p_A, we mean the probability of a probability, a very Bayesian concept. Lacking other information we take the priors as uniform,

$$\begin{aligned} p(p_A) &= 1, \qquad 0 < p_A < 1 \\ p(p_B) &= 1, \qquad 0 < p_B < 1 \end{aligned} \tag{11}$$

Now, dealing with just the A case (the B case being analogous), and applying Bayes theorem,

$$p(p_A|N,n) \propto p(N,n|p_A) \propto p_A^n (1-p_A)^{N-n} \tag{12}$$

i.e., $p_A \sim \text{Beta}(n+1, N-n+1)$, or

$$p(p_A|N,n) = \frac{\Gamma(N+2)}{\Gamma(n+1)\Gamma(N-n+1)} p_A^n (1-p_A)^{N-n} \tag{13}$$

From the properties of the beta distribution we have the relations, should we need them,

$$\begin{aligned} E(p_A) &= \frac{n+1}{N+2} \\ \text{Var}(p_A) &= \frac{(n+1)(N-n+1)}{(N+2)^2(N+3)} \end{aligned} \tag{14}$$

In fact, even more ambitious integrals over equation (13) can be done analytically. In particular we can evaluate the expectation of our previous figure of merit, F.M., over the probability distributions of p_A and p_B. This turns out to be

$$
\begin{aligned}
E(\text{F.M.}) &= \int\int dp_A dp_B [\text{F.M.}] p(p_A) p(p_B) \\
&= P(A) \Big[\langle p_A \log p_A \rangle_{n,N} - \langle p_A \rangle_{n,N} \langle \log p_B \rangle_{m,M} \\
&\qquad + \langle p_A \log p_A \rangle_{N-n,N} - \langle p_A \rangle_{N-n,N} \langle \log p_B \rangle_{M-m,M} \Big] \\
&+ P(B) \Big[\langle p_B \log p_B \rangle_{m,M} - \langle p_B \rangle_{m,M} \langle \log p_A \rangle_{n,N} \\
&\qquad + \langle p_B \log p_B \rangle_{M-m,M} - \langle p_B \rangle_{M-m,M} \langle \log p_A \rangle_{N-n,N} \Big]
\end{aligned}
\tag{15}
$$

where the different expectation values evaluate as

$$
\begin{aligned}
\langle x \rangle_{j,J} &= \frac{j+1}{J+2} \\
\langle \log x \rangle_{j,J} &= H(j) - H(J+1) \\
\langle x \log x \rangle_{j,J} &= \frac{j+1}{J+2} \left[H(j+1) - H(J+2) \right]
\end{aligned}
\tag{16}
$$

Here x is either p_A or p_B, j and J are integers, and $H(n)$ is the harmonic sum,

$$
H(n) \equiv \sum_{i=1}^{n} \frac{1}{i} \sim \log n + \gamma \tag{17}
$$

The asymptotic equality holds for large n with γ being Euler's constant, $0.57721\ldots$. Note that $H(0) \equiv 0$.

Because of its asymptotic form, it is convenient to define the notation

$$
H\binom{J}{j} \equiv H(J) - H(j) \sim \log \left(\frac{J}{j} \right) \tag{18}
$$

in terms of which

$$
\begin{aligned}
E(\text{F.M.}) = P(A) \Bigg\{ & \frac{n+1}{N+2} \left[H\binom{M+1}{m} - H\binom{N+2}{n+1} \right] \\
& + \frac{N-n+1}{N+2} \left[H\binom{M+1}{M-m} - H\binom{N+2}{N-n+1} \right] \Bigg\} \\
+ P(B) \Bigg\{ & \frac{m+1}{M+2} \left[H\binom{N+1}{n} - H\binom{M+2}{m+1} \right] \\
& + \frac{M-m+1}{M+2} \left[H\binom{N+1}{N-n} - H\binom{M+2}{M-m+1} \right] \Bigg\}
\end{aligned}
\tag{19}
$$

In this form it is evident how (19) becomes (9) in the limit of m, n, M, N all large. Equation (19) is, however, the better figure of merit to use for screening features, since it incorporates the appropriate corrections to (9) for small number statistical fluctuations in the counts m, n, M, N.

To use features selected in this manner, one wants to add not

$$\text{Choose}\left\{F_i, \log\left(\frac{p_{Ai}}{p_{Bi}}\right), \log\left(\frac{1-p_{Ai}}{1-p_{Bi}}\right)\right\} \tag{20}$$

to the log odds, since p_A and p_B are now not known, but rather

$$\begin{aligned}&\text{Choose}\left\{F_i, E\left[\log\left(\frac{p_{Ai}}{p_{Bi}}\right)\right], E\left[\log\left(\frac{1-p_{Ai}}{1-p_{Bi}}\right)\right]\right\}\\ &= \text{Choose}\left\{F_i,\ H\binom{M+1}{m} - H\binom{N+1}{n},\ H\binom{M+1}{M-m} - H\binom{N+1}{N-n}\right\}\end{aligned} \tag{21}$$

Note that this equation gives sensible (Bayesian) results, using $H(0) = 0$, even when m or n is zero, that is, when a feature is not observed at all in the A or B learning set.

4 Bayesian Correction for Multiple Hypotheses

Although equations (19) and (21) are correct for any single feature in the absence of prior knowledge of its performance, they are not quite correct for features that have been selected *on the basis of performance* on a learning set. The reason is that features with favorable fluctuations in their counts on the learning set will be preferentially selected over features with unfavorable fluctuations, so that both the F.M. and the log odds increment are slightly overestimated for selected features.

The most important manifestation of this bias occurs when we are screening very large numbers of rare features, most of which are expected to be causally unrelated to the classification of A versus B. Such features have $p_A = p_B$, but, because of fluctuations in the counts, not necessarily $n/N = m/M$. If selected, they will contribute spuriously to the log odds score.

A Bayesian way of looking at this is to assign to each feature i a probability Q_i of being causal ($p_A \neq p_B$), and a probability $1 - Q_i$ of being spurious ($p_A = p_B$). Then the expectation value of the *causal* F.M. is equation (19) multiplied by Q_i (plus zero multiplied by $1 - Q_i$), and similarly for equation (21).

So the question becomes: how do we estimate Q (omitting, for now, the index i) from the data?

Start with a prior Q_0 and convert it to a prior odds ratio $Q_0/(1-Q_0)$. Next multiply the prior odds ratio by the odds ratio of the evidence,

$$\begin{aligned}&\frac{\int\int P(\text{ Data }|p_A, p_B)dp_A dp_B}{\int P(\text{ Data }|p_A)dp_A}\\ &= \frac{\int\int \binom{N}{n}\binom{M}{m}p_A^n(1-p_A)^{N-n}p_B^m(1-p_B)^{M-m}dp_A dp_B}{\int \binom{N}{n}\binom{M}{m}p_A^{m+n}(1-p_A)^{N+M-n-m}dp_A}\\ &= \frac{B(n+1, N-n+1)B(m+1, M-m+1)}{B(n+m+1, N+M-n-m+1)}\end{aligned} \tag{22}$$

where B is a beta function.

While integrating over different numbers of parameters (p_A versus p_A and p_B) may seem odd to the uninitiated, this is in fact the standard Bayesian technique for comparing models with different numbers of parameters, yielding what are often called "Ockham factors". See, e.g., Sivia (1996), Chapter 4.

Finally, convert the resulting odds ratio back to a probability Q,

$$Q = \frac{\text{O.R.}}{\text{O.R.} + 1} \tag{23}$$

When M and N are large, and m/M and n/N are sensibly different, then Q is exponentially close to 1, even as the prior ranges over many orders of magnitude. This indicates that the feature is surely causal. When M and N are moderate or small, then nontrivial values of Q are obtained. For example, when $M = N = 30$ and $m = n = 15$, the evidence odds ratio (22) is 0.312, giving a modest discounting of the causal case – unless, of course, it is overwhelmed by the prior. When $M = N = 30$, $m = 2$, $n = 20$, on the other hand, (22) is 6.9×10^4, indicating strong evidence in favor of the causal case.

When one is screening combinatorially large numbers of possible features, it is reasonable to set the prior Q_0 quite small, for example on the order of the ratio of the number of features sought to the number of features screened. Strong features will survive, with $Q \approx 1$, even such a small prior. And, if there are no such strong features, then weak features will be appropriately ordered by their relative evidence factors, sharing the common prior Q_0.

One can also think of applications where it is logical to set the prior Q_0 close to unity, for example in evaluating features that are known on other experimental grounds to be associated with the classification of A versus B. Then, one wants to use the evidence ratio only to "knock out" features whose counts strongly imply $p_A = p_B$ in a Bayesian sense.

5 References

Sivia, D.S. 1996, *Data Analysis: A Bayesian Tutorial* (Oxford: Clarendon Press).

6 Acknowledgments

Thanks to Harlan Robbins for pointing out the need for a multiple hypothesis correction; and to Nick Hengartner for useful discussions.