Classification Rules that Include Neutral Zones and their Application to Microbial Community Profiling

Daniel R. Jeske[1], Zheng Liu[2], Elizabeth Bent[3] and James Borneman[3]

Abstract: We extend the classical one-dimensional Bayes binary classifier to create a classification rule that has a region of neutrality, to account for cases where the implied weight of evidence is too weak for a confident classification. The new classifier is illustrated using a microorganism community profiling application. In this application, ribosomal ribonucleic acid (rRNA) genes are hybridized with a series of oligonucleotide probes to determine if specific sequences of nucleotides are present. The outcome of the probe experiments provides a DNA fingerprint for the gene, which is subsequently used to predict the identity of the microorganism. The available measurement for determining presence or absence of a particular nucleotide sequence is a measured intensity level.

The traditional Bayes binary classifier uses the measured intensity level to predict presence or absence of the nucleotide sequence. Our proposed rule allows a "No Prediction" when the measured intensity level is too ambiguous to have confidence in an otherwise definite prediction. The motivation for making "No Prediction" is that in this application, a wrong prediction with one or more of the probes can be worse than making no prediction at all. On the other hand, too many "No Predictions" mutes the value of the DNA fingerprint for identifying the microorganism. Our proposed rule incorporates this trade-off using a cost structure that weighs the penalty for not making a definite prediction against the penalty for making an incorrect definite prediction. We demonstrate that our proposed rule outperforms a naive neutral-zone rule that has been routinely used in this type of application.

Keywords: Classification, Bayes Rule, Macroarray analysis

[1]Department of Statistics, University of California, Riverside, 92521 USA
[2]Department of Computer Science, University of California, Riverside, 92521 USA
[3]Department of Plant Pathology, University of California, Riverside, 92521 USA

1. Introduction

It is estimated that there are 160 different taxa of bacteria and other prokaryotes (single-celled organisms without a nucleus) in every milliliter of ocean water, while in a gram of soil there could be in the range of 6,400 – 38,000 different taxa [Curtis et al., (2002)]. The majority of microorganisms present in nature cannot be cultured for study under standard laboratory conditions [Amann et al., (1995)]. It is possible, however, to study these microorganisms via analysis of the molecules they produce. Ribosomal ribonucleic acid (rRNA) genes, which are found in the deoxyribonucleic acid (DNA) of every microorganism, are the most commonly used molecule for such studies. Comparative nucleotide sequence analysis of these genes have been used to provide an evolutionary basis for prokaryotic taxonomy [Woese and Fox (1995)], rRNA genes can be obtained from DNA extracted from environmental samples, and so an approach for identifying the microorganisms present in environmental samples is to identify particular rRNA gene sequences obtained from those samples. Although several rapid and cost-effective molecular methods for the analysis of rRNA genes in environmental samples exist [see, for example, Borneman and Triplett (1997), Liu et al. (1997), Muyzer et al. (1993) and Schwieger and Tebbe (1998)], these methods typically generate only superficial descriptions of microbial community composition.. Thorough descriptions of microbial communities based on rRNA gene profiles require a method that can identify thousands of genes simultaneously. One method for looking at many rRNA gene sequences at once in a cost-effective way is known as Oligonucleotide Fingerprinting of rRNA Genes, or OFRG [Valinsky et al. (2002a, 2002b, 2004)].

An OFRG experiment consists of PCR (polymerase chain reaction) amplifying rRNA genes from environmental DNA, separating this mixture of rRNA genes into individuals by the gene cloning process, arraying the gene inserts from the clones in a grid pattern on a nylon membrane (or macroarray) and subsequently hybridizing the arrayed genes with a series of oligonucleotide probes. Each probe consists of a particular sequence of ten nucleotides (A, C, G, T). A probe will usually bind to an rRNA gene sequence when a complimentary nucleotide sequence is present, and usually will not bind otherwise. Each probe is labeled with a radioisotope, and probe binding can be measured by the amount of radioactivity associated with each position in the macroarray. The amount of radioactivity across the macroarray is translated into an image, where more radioactivity becomes a darker area of the image. When the visual image is quantified using image analysis software, darker areas will have larger intensity values, and so large intensity levels give stronger evidence of probe binding. In typical applications, 30-40 different probes are used, each of which gives some information on the presence or absence of a particular 10-nucleotide long sequence in each rRNA gene. The result of all the probe hybridization trials is a fingerprint, which consists of a sequence of 0/1 indicators corresponding to the absence or presence of these specific 10-nucleotide long sequences.

The rRNA gene sequences from similar organisms will have similar fingerprints, while different organisms will tend to have different fingerprints. By determining which macroarray positions have fingerprints that are similar to each other (e.g., through the use of an unsupervised clustering algorithm) we can identify positions containing genes that are likely to be from the same or very similar microorganisms. We can then choose one of these positions, obtain the full sequence the rRNA gene in the clone associated with it, and match the gene with genes of known

organisms in public databases. In this way we can get useful information about a large number of rRNA genes from a sample without having to resort to obtaining the nucleotide sequence of each gene, an option that is still too costly for most laboratories to consider for large numbers of genes.

The aspect of OFRG that we focus on in this paper is the rule used to determine whether or not a probe successfully bound to the rRNA genes present in the macroarray. To assist in this decision, it is typical to calibrate the probes by analyzing a set of control clones, where the full nucleotide sequence, and therefore the binding result for each probe, is already known for the gene sequence in each clone. The analysis of the control clones consists of hybridizing them to each probe to generate a set of intensity values as depicted in Table 1. For each probe, the a-priori knowledge of whether binding should have occurred or not is shown along with the measured intensity level Y_{ij} obtained from hybridizing the probe with the control microorganism.

	Probe 1		Probe 2		…	Probe *K*	
Control Clone	Bind	Intensity	Bind	Intensity		Bind	Intensity
1	0	Y_{11}	1	Y_{12}	…	1	Y_{1K}
2	1	Y_{21}	1	Y_{22}	…	0	Y_{2K}
⋮	⋮	⋮	⋮	⋮		⋮	⋮
N	0	Y_{N1}	0	Y_{N2}	…	1	Y_{NK}

Table 1. Schematic Layout of the Data Available from Control Clones

The Y_{ij} observations in Table 1 can be used to build reference distributions for each probe of the measured intensity levels, given that binding occurs and given that it does not. The conditional distributions are obtained by separating the measured intensity levels for each probe into two groups, corresponding to the known a-priori binding status (0 or 1) of the control clones. Once

the conditional distributions for a given probe are obtained, the binding status of an experimental clone with respect to that probe can be predicted by comparing its measured intensity to each of the conditional distributions associated with that probe. Based on such a comparison, a statistical prediction can be made about whether or not the probe bound to the clone. Proceeding this way for all K probes, a predicted OFRG fingerprint is obtained for the unknown rRNA gene sequence.

In the experiment reported in Bent et al. (2005), there are 9,600 clones, of which there are $N = 432$ control clones. For the j-th probe, let $\{U_{jk}\}_{k=1}^{m_j}$ and $\{V_{jk}\}_{k=1}^{n_j}$ denote the intensity measurements corresponding to the control clones where no binding is expected and where binding is expected, respectively. Intensity measurements can occasionally be negative when the radioactivity in the clone sample is especially weak and background shade of the nylon membrane is darker than the spotted sample itself. It is customary to remove negative values from the analysis so that transformations such as log can be used when analyzing the data [see, for example, Edwards (2003)]. As a consequence, the number of intensity measurements is not the same for all probes.

It is not uncommon to have a large number of control clones in an OFRG experiment, and in what follows we assume that both n_j and m_j are sufficiently large to adequately characterize the distribution of intensity levels for both populations. In the present mode of operation, the OFRG fingerprint prediction for an experimental clone proceeds by considering separately each probe and referencing the measured intensity levels to the two sets of data $\{U_{jk}\}_{k=1}^{m_j}$ and $\{V_{jk}\}_{k=1}^{n_j}$. Let $U_{j(m_j)}$ denote the largest of the $\{U_{jk}\}_{k=1}^{m_j}$ values, and let $V_{j(1)}$ denote the smallest of the $\{V_{jk}\}_{k=1}^{n_j}$

values. Let the measured intensity levels with each probe for an experimental clone be denoted by $Y_1, \ldots, Y_K$ respectively. The currently used classification rule declares no binding (0) for the *j*-th probe if $Y_j \leq U_{j(m_j)}$, declares binding (1) if $Y_j \geq V_{j(1)}$, and makes no definitive declaration (N) if $\text{Min}\{V_{j(1)}, U_{j(m_j)}\} < Y_j < \text{Max}\{V_{j(1)}, U_{j(m_j)}\}$. In what follows, we refer to this rule as the min-max rule. The rationale for use of N is that there is too much 'noise' in the observation to get a clear indication of whether binding occurred or not, and rather than forcing a hard determination of 0 or 1 and risking a misclassification that could confuse the clustering algorithm, the result from the probe experiment will just be ignored.

Using the min-max rule for all probes $j = 1, \ldots, K$, results in a predicted fingerprint for an unknown rRNA gene sequence which can be satisfactorily utilized by the clustering algorithm provided the overlap between the $\{U_{jk}\}_{k=1}^{m_j}$ and $\{V_{jk}\}_{k=1}^{n_j}$ values is not substantial. The ideal situation for the min-max algorithm is when $U_{j(m_j)} < V_{j(1)}$. As the amount of overlap increases, the min-max rule becomes more likely to classify a clone as N. When a predicted OFRG fingerprint contains a lot of Ns, the subsequent clustering algorithm loses sensitivity and accuracy. Practioners of the min-max rule typically screen the intensity data from the analyses of the control microorganisms for outliers in an attempt to reduce the overlap problem. However, the identification of outlier points is largely subjective, and their removal in a data analysis procedure can be questionable.

In this paper, we develop an improved classification rule for use with OFRG experiments that has a rigorous interpretation and removes the subjectivity associated with the min-max rule. Our

proposed rule has the same form as the min-max rule, in that it makes 0, 1 and N predictions based on the observed intensity level. However, the region corresponding to N predictions is derived through a more formal approach where the relative cost of misclassifying the binding status is traded off against the cost of making an N classification. The cost structure of the new classification is a natural mechanism to control for the frequency of N classifications.

The rest of this paper is organized as follows. In Section 2 we review the formulation of a standard Bayesian classification rule for deciding which of two specified distributions an observation is most likely to have come from. Bayes rules always make a classification, and thus would not ever classify an observation as N. Section 3 modifies the Bayes rule in a very natural way to allow observations to be "too close to call" and thereby classify them as N. A cost function and the minimum cost modified Bayes rule is derived in Section 4. In Section 5, the modified Bayes rule is detailed for the special case of normal distributions. In Section 6, we return the OFRG application and develop the modified Bayes rule for use on the Bent et al. (2005) data set. We also present a comparative analysis that demonstrates the improvement afforded by the modified Bayes rule over the min-max rule.

2. Bayes Classification Rule

Suppose it is known that an object belongs to one of two classes, say $C = 0$ or $C = 1$, and that an observable continuous response variable Y is correlated to the value of C. The classification problem is to classify the object as either $C = 0$ or $C = 1$, based on the value of Y. Conditional on the value of C, we denote probability distribution functions (pdf's) of Y as $f_0(y) = f_{Y|C=0}(y)$ and $f_1(y) = f_{Y|C=1}(y)$, respectively, and assume these pdf's are either known or can be estimated with negligible error from a training sample. Although we do not consider it here, the Bayesian

approach to classification can be extended to cases where the underlying pdf depends on an unknown parameter vector θ by assuming a prior distribution for θ. See, for example, Jampachaisri et al. (2005) for an illustration that is related to our motivating application.

Let p_0 and p_1 represent the (known) a-priori probability that the object belongs to $C = 0$ and $C = 1$, respectively. Here, $0 < p_0$, $p_1 < 1$ and $p_0 + p_1 = 1$. In practice, p_0 (and thus $p_1 = 1 - p_0$) can be estimated by the proportion of objects in a random sample that belong to $C = 0$. In some cases, but not always, the training sample can be regarded as a random sample. In cases where a random sample is not available, values for p_0 and p_1 can be derived through the specification of subjective prior information (i.e., a Bayesian formulation). In cases where no subjective prior information is available, $p_0 = p_1 = 0.5$ can be used (i.e., a uniform prior).

Classification rules are mappings from *Y* to the set $\{0,1\}$. That is, a classification rule is a function d whose input is *Y* and whose output is 0 or 1. The outputted value for an object is a prediction for which class *C* that the object belongs too. The quality of a classification rule can be measured by the probability that it gives a correct prediction. In particular, $R(d) = 1 - Pr[d(Y) = C]$ is defined as the risk associated with the classification rule d. It is well known [see, for example, Press (1989)] that the classification rule that has the smallest risk is Bayes rule that is defined as follows

$$d_B(y) = \begin{cases} 0 & \text{if } \dfrac{f_1(y)p_1}{f_0(y)p_0 + f_1(y)p_1} < 0.5 \\ 1 & \text{if } \dfrac{f_1(y)p_1}{f_0(y)p_0 + f_1(y)p_1} \geq 0.5 \end{cases} \quad (1)$$

Note that the decision to classify as 1 when equality holds is arbitrary, but also inconsequential as the probability of equality is zero since *Y* is assumed to be continuous. Equivalently, $d_B(y)$ can be written as

$$d_B(y) = \begin{cases} 0 & \text{if } \dfrac{f_1(y)}{f_0(y)} < \dfrac{p_0}{p_1} \\ 1 & \text{if } \dfrac{f_1(y)}{f_0(y)} \geq \dfrac{p_0}{p_1} \end{cases} \tag{2}$$

Note that in the case of equal priors $p_0 = p_1 = 0.5$, Bayes rule reduces to classifying the object as $C = 1$ if and only if $f_1(y) \geq f_0(y)$ and becomes equivalent to the so-called maximum likelihood rule.

3. Modified Bayes Rule

Bayes rule have a probabilistic interpretation that points us toward a natural way to modify the rule for cases where the information in *Y* about which class the object belongs too is ambiguous. In particular, a straightforward application of Bayes rule shows that the posterior probability of the event $C = 1$ is given by $Pr(C = 1 | Y = y) = \dfrac{f_1(y)p_1}{f_0(y)p_0 + f_1(y)p_1}$ and thus (1) can be expressed as

$$d_B(y) = \begin{cases} 0 & \text{if } \Pr(C = 1 | Y = y) < 0.5 \\ 1 & \text{if } \Pr(C = 1 | Y = y) \geq 0.5 \end{cases} \tag{3}$$

The representation in (2) suggests that when the posterior probability of the event $C = 1$ is close to 0.5 (on either side), then there is relatively weak evidence in *Y* to distinguish between $C = 0$ and $C = 1$. An intuitively appealing modifed Bayes Rule is the rule of the form

$$d_B(y\,;L_0\,,L_1)=\begin{cases}0 & \text{if}\quad \Pr(\,C=1\,|\,Y=y\,)\;\le\;L_0\\ \text{N} & \text{if}\quad L_0<\;\Pr(\,C=1\,|\,Y=y\,)\;<\;L_1\\ 1 & \text{if}\quad \Pr(\,C=1\,|\,Y=y\,)\;\ge\;L_1\end{cases} \tag{4}$$

where $(L_0\,,L_1)\in D$, with $D=\{(L_0\,,L_1):\;0\le L_0\le L_1\le 1\}$. The idea with (4) is that if the posterior probability of the event $C=1$ lies in the interval $(L_0\,,L_1)$ [e.g., (.45, .55)], then the evidence in *Y* for the object to be classified as either $C=0$ and $C=1$ is too weak to make a definitive prediction. Rather than making a definitive classification based on weak evidence, the rule $d_B(y\,;L_0\,,L_1)$ instead classifies the object as, N for "No Prediction." An alternative formulation for $d_B(y\,;L_0\,,L_1)$ is

$$d_B(y\,;L_0\,,L_1)=\begin{cases}0 & \text{if}\quad \dfrac{f_1(y)}{f_0(y)}\;\le\;\dfrac{p_0L_0}{p_1(1-L_0)}\\ \text{N} & \text{if}\quad \dfrac{p_0L_0}{p_1(1-L_0)}\;<\;\dfrac{f_1(y)}{f_0(y)}\;<\;\dfrac{p_0L_1}{p_1(1-L_1)}\\ 1 & \text{if}\quad \dfrac{f_1(y)}{f_0(y)}\;\ge\;\dfrac{p_0L_1}{p_1(1-L_1)}\end{cases} \tag{5}$$

Clearly if $L_0=L_1=0.5$, the modified Bayes rule in (5) reduces to the Bayes rule in (2). Selection of specific values for L_0 and L_1 is based upon how much certainty in the evidence represented by *Y* is required by the user before they feel comfortable about making a classification decision. In the following section, we offer further guidance in selecting L_0 and L_1 by relating their values to the Type-1 and Type-2 misclassification rates of $d_B(y\,;L_0\,,L_1)$.

4. Minimum Cost Modified Bayes Rule

Because modified Bayes rules have three possible values (0, 1 and N), there are four types of error. Table 2 shows a symmetric cost function for each of the error types. For what follows, we

define the ratio $\rho = C_1 / C_2$. For the OFRG fingerprinting application we introduced in Section 1, $\rho > 1$ are the cases of interest.

Classified Population	True Population	
	0	1
0	0	C_1
1	C_1	0
N	C_2	C_2

Table 2. Costs of Misclassifications

Denote the classification regions of (5) as

$$D_0 = \left\{ y : \frac{f_1(y)}{f_0(y)} \le \frac{p_0 L_0}{p_1(1-L_0)} \right\}$$

$$D_{\mathrm{N}} = \left\{ y : \frac{p_0 L_0}{p_1(1-L_0)} < \frac{f_1(y)}{f_0(y)} < \frac{p_0 L_1}{p_1(1-L_1)} \right\}$$

$$D_1 = \left\{ y : \frac{f_1(y)}{f_0(y)} \ge \frac{p_0 L_1}{p_1(1-L_1)} \right\}$$

Letting C and $\hat{C}$ denote the true and predicted class membership for an object, the probability of each type of error can be found as

$$\begin{aligned} Pr(\hat{C} = 1 \mid C = 0) &= \int_{D_1} f_0(y)\, dy \\ Pr(\hat{C} = 0 \mid C = 1) &= \int_{D_0} f_1(y)\, dy \\ Pr(\hat{C} = \mathrm{N} \mid C = 0) &= \int_{D_{\mathrm{N}}} f_0(y)\, dy \\ Pr(\hat{C} = \mathrm{N} \mid C = 1) &= \int_{D_{\mathrm{N}}} f_1(y)\, dy \end{aligned} \tag{6}$$

The conditional expected costs of misclassifications are

$$\begin{aligned} \mathrm{E}\,[\,\mathrm{Cost} \mid C = 0\,] &= C_1 Pr(\hat{C} = 1 \mid C = 0) + C_2 Pr(\hat{C} = \mathrm{N} \mid C = 0) \\ \mathrm{E}\,[\,\mathrm{Cost} \mid C = 1\,] &= C_1 Pr(\hat{C} = 0 \mid C = 1) + C_2 Pr(\hat{C} = \mathrm{N} \mid C = 1)\ , \end{aligned} \tag{7}$$

each depending on L_0 and L_1 through the regions D_0, D_1 and D_N. Consequently, $\mathrm{E}[\text{ Cost }] \propto f(L_0, L_1)$ where

$$\begin{aligned} f(L_0, L_1) \;=\; & p_0\left[\rho Pr(\hat{C}=1 \mid C=0) \;+\; Pr(\hat{C}=\mathrm{N} \mid C=0)\right] \\ & +\; p_1\left[\rho Pr(\hat{C}=0 \mid C=1) \;+\; Pr(\hat{C}=\mathrm{N} \mid C=1)\right] \end{aligned} \tag{8}$$

We can then seek the value $(L_0^*, L_1^*) \in D$ that minimizes $f(L_0, L_1)$, and the corresponding rule $d_B(y; L_0^*, L_1^*)$ is the minimum expected cost modified Bayes rule. We denote the corresponding minimum cost modified Bayes rule by $d_B^*(y; \rho)$ to emphasize its dependence on ρ. In some applications, it may not be clear what particular value of ρ is appropriate, and an application-specific approach will be needed to determine a suitable value. In Section 6, we describe the approach we employed for the OFRG fingerprinting application described in Section 1.

The uniqueness of the minimum cost modified Bayes result has not yet been proved, but our application (Section 6) does provide some insight on the issue. In particular, we found that when there were two values (L_0, L_1) and (L_0', L_1') for which $f(L_0, L_1) = f(L_0', L_1')$, then $d_B(L_0, L_1) = d_B(L_0', L_1')$. That is, the form of the modified Bayes rule is the same for any two choices of (L_0, L_1) that have the same expected cost, a condition that is sufficient for the minimum expected cost modified Bayes rule to be unique. More generally, the issue of uniqueness is a topic of on-going work, but certainly in all applications the uniqueness of the rule can be easily checked empirically.

5. Modified Bayes Rule with Normal Distributions

In this section, we develop explicit forms of the modified Bayes rule $d_B(y\,;L_0\,,L_1)$ for the case where the probability distribution functions $f_0(y)$ and $f_1(y)$ are normal with means and variances given by $\{\mu_i\}_{i=0}^1$ and $\{\sigma_i^2\}_{i=0}^1$, respectively. We note that through a natural log transformation, our results also apply to contexts where the response has one of two lognormal distributions. The use of normal and lognormal distributions for analysis of microarray intensity measurements is common in the literature. Applications using the normal model can be found Li and Wong (2001) and Giles and Kipling (2003), while applications using the lognormal model can be found in Wolfinger et al. (2001), Baldi and Long (2001), and Hackl et al. (2004). In section 5.1 we assume the two Gaussian distributions have the same variance. The cases of unequal variances are addressed in sections 5.2 and 5.3.

5.1 $\sigma_0^2 = \sigma_1^2 = \sigma^2$

5.1.1 Classification Rule

Suppose $f_i(y) = e^{-\frac{(y-\mu_i)^2}{2\sigma^2}}/(2\sigma\sqrt{\pi})$ for i = 0, 1. Without loss of generality, we suppose that $\mu_1 > \mu_0$. It is straightforward to show that (5) evaluates to

$$d_B(y\,;L_0\,,L_1) = \begin{cases} 0 & \text{if } y \le r(L_0) \\ \text{N} & \text{if } r(L_0) < y < r(L_1) \\ 1 & \text{if } y \ge r(L_1) \end{cases} \tag{9}$$

where

$$r(L) = \frac{\mu_0+\mu_1}{2} + \frac{\sigma^2 \log\left[\dfrac{p_0 L}{p_1(1-L)}\right]}{\mu_1-\mu_0} \tag{10}$$

To ease notation, we will at times write r_i rather than $r(L_i)$, $i = 0, 1$. In applications, the training data sets would be used to estimate the parameters μ_0, μ_1 and σ^2. Figure 1 is a partition of the values for *Y* according to the rule specified by (9).

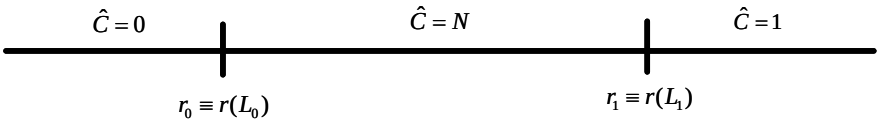

Figure 1. Classification Boundaries of the Modified Bayes Rule for Equal Variance Gaussian Distributions

5.1.2 Minimum Cost Modified Bayes Rule

Letting C and $\hat{C}$ denote the true and predicted class membership for an object, and referring to Figure 1, it is easily seen that

$$\begin{aligned}
Pr(\hat{C} = 1 \mid C = 0) &= 1 - \Phi\left(\frac{r_1 - \mu_0}{\sigma}\right) \\
Pr(\hat{C} = 0 \mid C = 1) &= \Phi\left(\frac{r_0 - \mu_1}{\sigma}\right) \\
Pr(\hat{C} = \mathrm{N} \mid C = 0) &= \Phi\left(\frac{r_1 - \mu_0}{\sigma}\right) - \Phi\left(\frac{r_0 - \mu_0}{\sigma}\right) \\
Pr(\hat{C} = \mathrm{N} \mid C = 1) &= \Phi\left(\frac{r_1 - \mu_1}{\sigma}\right) - \Phi\left(\frac{r_0 - \mu_1}{\sigma}\right)
\end{aligned} \tag{11}$$

The expressions in (11) characterize the four types of conditional misclassification rates of the modified Bayes rule. Inserting (11) into (8) gives

$$\begin{aligned}
f(L_0, L_1) = {} & p_0\left[\rho\left\{1 - \Phi\left(\frac{r_1 - \mu_0}{\sigma}\right)\right\} + \Phi\left(\frac{r_1 - \mu_0}{\sigma}\right) - \Phi\left(\frac{r_0 - \mu_0}{\sigma}\right)\right] \\
& + p_1\left[\rho\,\Phi\left(\frac{r_0 - \mu_1}{\sigma}\right) + \Phi\left(\frac{r_1 - \mu_1}{\sigma}\right) - \Phi\left(\frac{r_0 - \mu_1}{\sigma}\right)\right]
\end{aligned} \tag{12}$$

That is, $\mathrm{E}[\,\text{Cost}\,] \propto f(L_0, L_1)$, and we find optimal values of L_0 and L_1 by minimizing $f(L_0, L_1)$ over the set $D = \{(L_0, L_1):\ 0 \le L_0 \le L_1 \le 1\}$. The minimum of $f(L_0, L_1)$ over the region D occurs either at a stationary point located in the interior of *D*, or on a point located on the boundary of *D*. It can be shown that the minimum of $f(L_0, L_1)$ occurs at the candidate points shown in Table 3. Hence, depending on the value of ρ, $f(L_0, L_1)$ can simply be evaluated at the points shown in Table 3 to identify where the minimum occurs. Letting (L_0^*, L_1^*) denote the location of the minimum, the minimum expected cost modified Bayes rule is then $d_B^*(y;\rho) \equiv d_B(y; L_0^*, L_1^*)$.

Range of ρ	Possible Minima (L_0, L_1) for Expected Cost $f(L_0, L_1)$
$\rho \le 1$	$(0.5, 0.5)$, $(0, 0)$ or $(1, 1)$
$1 < \rho < 2$	$(0.5, 0.5)$, $(0, (\rho-1)/\rho)$ or $(1/\rho, 1)$
$\rho \ge 2$	$(1/\rho, (\rho-1)/\rho)$, $(0.5, 0.5)$, $(0, (\rho-1)/\rho)$ or $(1/\rho, 1)$

Table 3. Candidates for Minimum Expected Cost (L_0, L_1) in Equal Variance Normal Case

5.2 $\sigma_0^2 > \sigma_1^2$

5.2.1 Classification Rule

Suppose that $f_i(y) = e^{-\frac{(y-\mu_i)^2}{2\sigma_i^2}}/(2\sigma_i\sqrt{\pi})$ and $\sigma_0^2 > \sigma_1^2$. As before we assume without loss of generality that $\mu_1 > \mu_0$. In this case (5) evaluates to

$$d_B(y; L_0, L_1) = \begin{cases} 0 & \text{if } Q_0(y) \ge 0 \\ \text{N} & \text{if } Q_0(y) < 0 \text{ and } Q_1(y) > 0 \\ 1 & \text{if } Q_1(y) \le 0 \end{cases} \tag{13}$$

where

$$Q_i(y) \;=\; (\sigma_0^2-\sigma_1^2)y^2 \;+\; 2(\mu_0\sigma_1^2-\mu_1\sigma_0^2)y \;+\; (\sigma_0^2\mu_1^2-\sigma_1^2\mu_0^2) \;-\; 2\sigma_1^2\sigma_0^2\log\left[\frac{\sigma_0 p_1(1-L_i)}{\sigma_1 p_0 L_i}\right] \quad , \quad i=0,1 \tag{14}$$

Denote the ordered roots of $Q_0(y)=0$ and $Q_1(y)=0$ by $\{R^*_{(k)}\}_{k=0}^1$ and $\{R_{(k)}\}_{k=0}^1$, respectively. A necessary and sufficient condition for all four of the roots to be real is

$$L_i \;\le\; \left[\;1 \;+\; \frac{\sigma_1 p_0}{\sigma_0 p_1}\, e^{-\frac{(\mu_1-\mu_0)^2}{2(\sigma_0^2-\sigma_1^2)}}\right]^{-1} \;\equiv\; L^* \quad , \quad i=0,1 \tag{16}$$

The rule (13) can be written explicitly as a function of *y*, though the form varies as a function of the partition set that the point $(L_0\,,L_1)$ falls into. Figure 2 shows a partition of the region $D=\{(L_0\,,L_1):\;0\le L_0\le L_1\le 1\}$ into three sets S_1, S_2, S_3, eight line segments (*a*) thru (*h*), and five points P_0 thru P_4. Table 4 shows the different forms that the rule $d_B(y\,;L_0\,,L_1)$ takes on each partition set. We note that in the particular case of S_1, it can be shown that $R^*_{(0)}<R_{(0)}<R_{(1)}<R^*_{(1)}$.

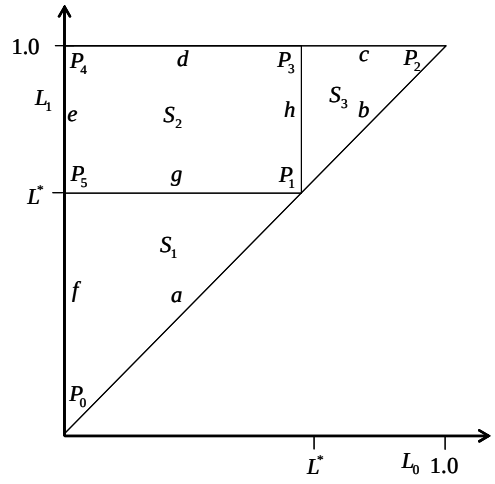

Figure 2. Partition of $(L_0\,,L_1)$ Domain

Partition Set	Form of Rule
P_3 , c , P_2 , b , P_1 , S_3 , h	$d_B(y\,;L_0\,,L_1) \equiv 0$
P_0	$d_B(y\,;L_0\,,L_1) \equiv 1$
P_5 , e , P_4	$d_B(y\,;L_0\,,L_1) \equiv N$
a	$d_B(y\,;L_0\,,L_1) \equiv \begin{cases} 0 & \text{if } y \le R^*_{(0)} \text{ or } y \ge R^*_{(1)} \\ 1 & \text{if otherwise} \end{cases}$
d , S_2 , g	$d_B(y\,;L_0\,,L_1) \equiv \begin{cases} 0 & \text{if } y \le R^*_{(0)} \text{ or } y \ge R^*_{(1)} \\ \text{N} & \text{if otherwise} \end{cases}$
f	$d_B(y\,;L_0\,,L_1) \equiv \begin{cases} 1 & \text{if } R_{(0)} \le y \le R_{(1)} \\ \text{N} & \text{if otherwise} \end{cases}$
S_1	$d_B(y\,;L_0\,,L_1) \equiv \begin{cases} 0 & \text{if } y \le R^*_{(0)} \text{ or } y \ge R^*_{(1)} \\ \text{N} & \text{if } R^*_{(0)} < y < R_{(0)} \text{ or } R_{(1)} < y < R^*_{(1)} \\ 1 & \text{if } R_{(0)} \le y \le R_{(1)} \end{cases}$

Table 4. Modified Bayes Rule for the Case $\sigma_0^2 > \sigma_1^2$

Figure 3 is a pictorial representation of the classification rule on the set S_1. Comparing Figure 3 with Figure 1, some insight is gained on the effect of unequal variances.

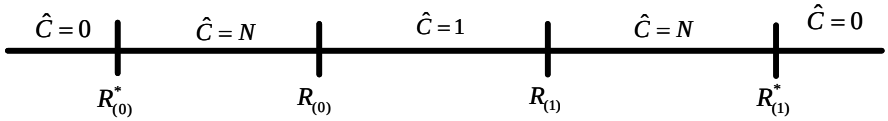

Figure 3. Classification Boundaries of the Modified Bayes Rule for Gaussian Distributions with $\sigma_0^2 > \sigma_1^2$ and $(L_0\,,L_1) \in S_1$

5.2.2 Minimum Cost Modified Bayes Rule

Evaluating (8) requires consideration of the varying form of the rule when evaluating the conditional error probabilities $Pr(\hat{C}=1\,|\,C=0)$, $Pr(\hat{C}=0\,|\,C=1)$, $Pr(\hat{C}=\text{N}\,|\,C=1)$ and $Pr(\hat{C}=\text{N}\,|\,C=0)$. Tables 5a and 5b give the required expressions. For a given cost ratio ρ, a numerical algorithm for finding the optimal modified Bayes rule is as follows.

1. Define a fine grid over the region shown in Figure 2, including the boundaries.

2. For each point $(L_0\,,\,L_1)$ of the grid:
   a. Identify which partition set $(L_0\,,\,L_1)$ belongs to.
   b. Compute the four conditional error probabilities from Tables 5a and 5b, using (15) to compute the roots $\{R^*_{(k)}\}_{k=0}^{1}$ and $\{R_{(k)}\}_{k=0}^{1}$, as needed.
   c. Combine the conditional error probabilities with equation (8) to compute the expected cost of the modified Bayes rule corresponding to $(L_0\,,\,L_1)$.
3. From the result of step (2), identify the grid point, $(L_0^*\,,\,L_1^*)$ which has the minimum expected cost.
4. Identify which partition set $(L_0^*\,,\,L_1^*)$ belongs to and use Table 4 to define the minimum expected cost modified Bayes rule $d_B^*(y\,;\rho) \equiv d_B(y\,;L_0^*\,,\,L_1^*)$.

Partition Set	$Pr(\hat{C}=1\,\vert\,C=0)$	$Pr(\hat{C}=0\,\vert\,C=1)$
$P_3\,,\,c\,,\,P_2\,,\,b\,,\,P_1\,,\,S_3\,,\,h$	0	1
P_0	1	0
$P_5\,,\,e\,,\,P_4$	0	0
a	$\Phi\left(\frac{R^*_{(1)}-\mu_0}{\sigma_0}\right) - \Phi\left(\frac{R^*_{(0)}-\mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R^*_{(0)}-\mu_1}{\sigma_1}\right) + 1 - \Phi\left(\frac{R^*_{(1)}-\mu_1}{\sigma_1}\right)$
$d\,,\,S_2\,,\,g$	0	$\Phi\left(\frac{R^*_{(0)}-\mu_1}{\sigma_1}\right) + 1 - \Phi\left(\frac{R^*_{(1)}-\mu_1}{\sigma_1}\right)$
f	$\Phi\left(\frac{R_{(1)}-\mu_0}{\sigma_0}\right) - \Phi\left(\frac{R_{(0)}-\mu_0}{\sigma_0}\right)$	0
S_1	$\Phi\left(\frac{R_{(1)}-\mu_0}{\sigma_0}\right) - \Phi\left(\frac{R_{(0)}-\mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R^*_{(0)}-\mu_1}{\sigma_1}\right) + 1 - \Phi\left(\frac{R^*_{(1)}-\mu_1}{\sigma_1}\right)$

Table 5a. Conditional Error Probabilities for the Case $\sigma_0^2 > \sigma_1^2$

Partition Set	$Pr(\hat{C} = \mathrm{N} \mid C = 0)$	$Pr(\hat{C} = \mathrm{N} \mid C = 1)$
P_3 , c , P_2 , b , P_1 , S_3 , h	0	0
P_0	0	0
P_5 , e , P_4	1	1
a	0	0
d , S_2 , g	$\Phi\left(\frac{R^*_{(1)} - \mu_0}{\sigma_0}\right) - \Phi\left(\frac{R^*_{(0)} - \mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R^*_{(1)} - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{R^*_{(0)} - \mu_1}{\sigma_1}\right)$
f	$\Phi\left(\frac{R_{(0)} - \mu_0}{\sigma_0}\right) + 1 - \Phi\left(\frac{R_{(1)} - \mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R_{(0)} - \mu_1}{\sigma_1}\right) + 1 - \Phi\left(\frac{R_{(1)} - \mu_1}{\sigma_1}\right)$
S_1	$\Phi\left(\frac{R_{(0)} - \mu_0}{\sigma_0}\right) - \Phi\left(\frac{R^*_{(0)} - \mu_0}{\sigma_0}\right)$ $+ \Phi\left(\frac{R^*_{(1)} - \mu_0}{\sigma_0}\right) - \Phi\left(\frac{R_{(1)} - \mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R_{(0)} - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{R^*_{(0)} - \mu_1}{\sigma_1}\right)$ $+ \Phi\left(\frac{R^*_{(1)} - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{R_{(1)} - \mu_1}{\sigma_1}\right)$

Table 5b. Conditional Error Probabilities for the Case $\sigma_0^2 > \sigma_1^2$

5.3 $\sigma_1^2 > \sigma_0^2$

5.3.1 Classification Rule

The classification rule for this case also takes the form (13) , with $\{R^*_{(k)}\}_{k=0}^{1}$ and $\{R_{(k)}\}_{k=0}^{1}$, defined the same way as they were for the case $\sigma_0^2 > \sigma_1^2$. An important difference, however, is that the necessary and sufficient condition for the roots to be real is

$$L_i \geq \left[1 + \frac{\sigma_1 p_0}{\sigma_0 p_1} e^{-\frac{(\mu_1 - \mu_0)^2}{2(\sigma_0^2 - \sigma_1^2)}} \right]^{-1} \equiv L^* \quad , \quad i = 0,1 \tag{17}$$

which is the reverse of (16). Again referring to Figure 2, the form of the modified Bayes rule depends on which partition set (L_0 , L_1) fall into. Table 6 shows the different forms of the rule for each partition set. We note that in the particular case of S_3, it can be shown that $R_{(0)} < R^*_{(0)} < R^*_{(1)} < R_{(1)}$.

Partition Set	Form of Rule
P_2	$d_B(y\,;L_0,L_1) \equiv 0$
P_0 , f , P_5 , P_1 , a , S_1 , g	$d_B(y\,;L_0,L_1) \equiv 1$
P_4 , d , P_3	$d_B(y\,;L_0,L_1) \equiv N$
b	$d_B(y\,;L_0,L_1) \equiv \begin{cases} 0 & \text{if } R_{(0)} \le y \le R_{(1)} \\ 1 & \text{if otherwise} \end{cases}$
c	$d_B(y\,;L_0,L_1) \equiv \begin{cases} 0 & \text{if } R^*_{(0)} \le y \le R^*_{(1)} \\ \text{N} & \text{if otherwise} \end{cases}$
e , S_2 , h	$d_B(y\,;L_0,L_1) \equiv \begin{cases} 1 & \text{if } y \le R_{(0)} \text{ or } y \ge R_{(1)} \\ \text{N} & \text{if otherwise} \end{cases}$
S_3	$d_B(y\,;L_0,L_1) \equiv \begin{cases} 0 & \text{if } R^*_{(0)} \le y \le R^*_{(1)} \\ \text{N} & \text{if } R_{(0)} < y < R^*_{(0)} \text{ or } R^*_{(1)} < y < R_{(1)} \\ 1 & \text{if } y \le R_{(0)} \text{ or } y \ge R_{(1)} \end{cases}$

Table 6. Modified Bayes Rule for the Case $\sigma_1^2 > \sigma_0^2$

Figure 4 is a pictorial representation of the classification rule on the set S_3.

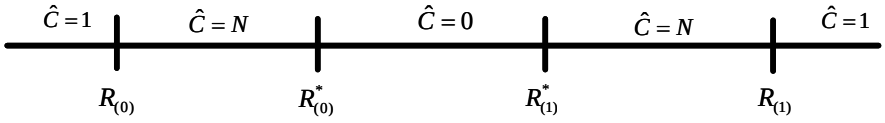

Figure 4. Classification Boundaries of the Modified Bayes Rule for Gaussian Distributions where $\sigma_1^2 > \sigma_0^2$ and $(L_0\,, L_1) \in S_3$

5.3.2 Minimum Expected Cost Modified Bayes Rule

The procedure for finding the optimal modified Bayes rule is similar to what is described in Section 5.2.2 for the case $\sigma_0^2 > \sigma_1^2$. For a given cost ratio ρ, the numerical algorithm described in steps (1)-(4) for finding the minimum expected cost modified Bayes rule $d_B^*(y\,;\rho)$ can be

followed, substituting Tables 7a -7b for Tables 5a -5b in step (2c) and substituting Table 6 for Table 4 in step (4).

Partition Set	$Pr(\hat{C}=1 \mid C=0)$	$Pr(\hat{C}=0 \mid C=1)$
P_2	0	1
P_0 , f , P_5 , P_1 , a , S_1 , g	1	0
P_4 , d , P_3	0	0
b	$\Phi\left(\frac{R_{(0)}-\mu_0}{\sigma_0}\right) + 1 - \Phi\left(\frac{R_{(1)}-\mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R_{(1)}-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{R_{(0)}-\mu_1}{\sigma_1}\right)$
c	0	$\Phi\left(\frac{R^*_{(1)}-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{R^*_{(0)}-\mu_1}{\sigma_1}\right)$
e , S_2 , h	$\Phi\left(\frac{R_{(0)}-\mu_0}{\sigma_0}\right) + 1 - \Phi\left(\frac{R_{(1)}-\mu_0}{\sigma_0}\right)$	0
S_3	$\Phi\left(\frac{R_{(0)}-\mu_0}{\sigma_0}\right) + 1 - \Phi\left(\frac{R_{(1)}-\mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R^*_{(1)}-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{R^*_{(0)}-\mu_1}{\sigma_1}\right)$

Table 7a. Conditional Error Probabilities for the Case $\sigma_1^2 > \sigma_0^2$

Partition Set	$Pr(\hat{C}=\mathrm{N} \mid C=0)$	$Pr(\hat{C}=\mathrm{N} \mid C=1)$
P_2	0	0
P_0 , f , P_5 , P_1 , a , S_1 , g	0	0
P_4 , d , P_3	1	1
b	0	0
c	$\Phi\left(\frac{R^*_{(0)}-\mu_0}{\sigma_0}\right) + 1 - \Phi\left(\frac{R^*_{(1)}-\mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R^*_{(0)}-\mu_1}{\sigma_1}\right) + 1 - \Phi\left(\frac{R^*_{(1)}-\mu_1}{\sigma_1}\right)$
e , S_2 , h	$\Phi\left(\frac{R_{(1)}-\mu_0}{\sigma_0}\right) - \Phi\left(\frac{R_{(0)}-\mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R_{(1)}-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{R_{(0)}-\mu_1}{\sigma_1}\right)$

S_3	$\Phi\left(\frac{R^*_{(0)}-\mu_0}{\sigma_0}\right) - \Phi\left(\frac{R_{(0)}-\mu_0}{\sigma_0}\right)$ $+ \Phi\left(\frac{R_{(1)}-\mu_0}{\sigma_0}\right) - \Phi\left(\frac{R^*_{(1)}-\mu_0}{\sigma_0}\right)$	$\Phi\left(\frac{R^*_{(0)}-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{R_{(0)}-\mu_1}{\sigma_1}\right)$ $+ \Phi\left(\frac{R_{(1)}-\mu_1}{\sigma_1}\right) - \Phi\left(\frac{R^*_{(1)}-\mu_1}{\sigma_1}\right)$

Table 7b. Conditional Error Probabilities for the Case $\sigma_1^2 > \sigma_0^2$

6. OFRG Fingerprinting Application

6.1 Application of Modified Bayes Rules

In this Section, we return to the OFRG application introduced in Section 1. Consistent with some of the other previously literature, we found that a log transformation made the normality assumption for Y_{ij} reasonable. In our particular application, there were $K = 28$ probes and $N = 432$ control clones. There were 68 control clones that did not bind properly to the probes, so the effective number of control clones had to be reduced to 364. Table 8 shows summary statistics for the log-intensity measurements for each probe. Columns 2-5 summarize the measurements for the control clones where no binding to the probe should occur, and the columns correspond to the sample size, the maximum value, the sample mean and the sample standard deviation, respectively. Columns 6-9 provide the same information for the control clones where binding to the probe should occur. The min-max rule for each probe utilizes the information in columns 3 and 7. Only fourteen of the probes (1, 2, 6, 8, 12, 21, 30, 40, 42, 43, 44, 45, 46, and 48) satisfy the non-overlapping property $U_{j(m_j)} < V_{j(1)}$, and in some cases the extent of overlap is quite significant, implying the min-max rule will have a high probability of classifying a clone as N.

Probe (j)	Control Clones							
	Without a-Priori Binding (0)				With a-Priori Binding (1)			
	m_j	$U_{j(m_j)}$	$\bar{U}_j$	S_j^U	n_j	$V_{j(1)}$	$\bar{V}_j$	S_j^V
1	212	-1.075	-4.913	1.563	127	-4.190	-1.259	0.797
2	97	-3.161	-6.424	1.646	231	-5.005	-3.484	0.369
6	186	0.008	-2.717	1.363	167	-2.501	-1.238	0.373
8	237	-0.713	-3.976	1.428	124	-3.329	-1.173	0.738
9	17	-1.647	-4.263	1.199	343	-6.885	-2.373	0.646
10	262	0.026	-1.625	1.009	98	-4.356	-0.124	0.657
11	19	-1.943	-3.732	1.235	342	-3.874	-1.352	0.519
12	113	-0.714	-3.348	1.552	230	-1.916	-0.579	0.311
13	240	1.755	-0.144	1.073	122	-0.957	0.986	0.442
15	103	0.354	-3.750	1.847	237	-4.111	-0.364	0.674
21	277	0.665	-4.039	1.384	79	-2.602	-0.686	0.764
23	16	-0.525	-1.368	0.405	346	-5.487	0.265	0.568
29	36	-0.654	-2.320	0.812	318	-3.503	-0.394	0.694
30	296	1.682	-2.246	1.668	60	-1.972	0.313	0.628
31	332	0.718	-2.449	0.815	29	-3.519	0.532	0.981
32	63	1.662	-3.491	2.133	279	-4.146	-0.745	0.508
33	67	-0.836	-2.943	1.470	289	-9.850	-1.825	0.635
35	190	-0.883	-5.207	1.535	153	-5.693	-1.696	0.832
36	335	-0.987	-4.987	0.982	29	-4.723	-0.838	0.843
37	287	-0.808	-4.776	1.381	70	-5.855	-1.369	0.927
40	66	-3.257	-5.924	1.259	288	-5.391	-3.498	0.538
41	70	-1.189	-4.867	1.816	271	-5.032	-2.069	0.682
42	333	1.312	-2.662	1.580	23	-0.627	0.216	0.465
43	221	0.309	-4.079	1.674	128	-3.842	-0.410	0.779
44	287	2.404	-2.236	1.577	72	-0.324	0.735	0.614
45	185	-0.526	-4.393	1.750	152	-3.793	-0.686	0.540
46	221	2.145	-1.390	1.590	140	-1.398	1.174	0.578
48	253	-0.809	-4.648	1.382	104	-3.949	-1.616	0.661

Table 8. Summary Statistics for the Control Clones

Modified Bayes rules for a probe would utilize the means and standard deviations in columns 4, 5, 8 and 9. Comparing columns 5 and 9, it is quite apparent that the standard deviations are not equal and thus the unequal variances cases considered in Section 5 are appropriate. To illustrate the construction of a modified Bayes rules, consider probe #1. Recalling the convention

$\mu_1 > \mu_0$, it follows from Table 8 that $(\mu_0\,,\,\sigma_0) = (-4.913\,,\,1.563)$ and $(\mu_1\,,\,\sigma_1) = (-1.259\,,\,0.797)$. Since $\sigma_0 > \sigma_1$, we follow the procedures outlined in Section 5.2 for constructing the modified Bayes rule.

For our first example with this probe, suppose a user selects $(L_0\,,\,L_1) = (0.45\,,\,0.55)$ on the basis of their subjective comfort levels with what the weight of evidence should be before classifying a clone as a 0 or 1. We find from (16) that $L^* = 0.987$, and thus all of the roots $\{R^*_{(k)}\}_{k=0}^{1}$ and $\{R_{(k)}\}_{k=0}^{1}$ are real. From (15) it follows that $(\,R^*_{(0)}\,,\,R_{(0)}\,,\,R_{(1)}\,,\,R^*_{(1)}\,) = (\,-2.775\,,\,-2.649\,,\,2.699\,,\,2.825\,)$. Figure 5 is a pictorial (not drawn to scale) representation of the modified Bayes rule.

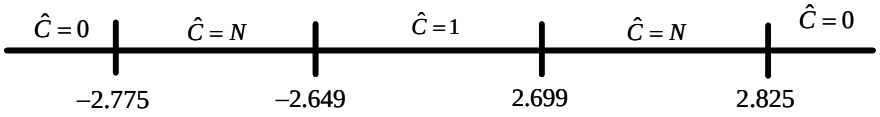

Figure 5. Classification Boundaries of the Modified Bayes Rule for Probe #1 Choosing $(L_0\,,\,L_1) = (0.45\,,\,0.55)$

Referring to Figure 2, this modified Bayes rule falls into Region S_1. Misclassification rates can be found from the last rows of Tables 5a and 5b. In particular, we have $Pr(\hat{C} = 1\,|\,C = 0) = 0.074$, $Pr(\hat{C} = 0\,|\,C = 1) = 0.029$, $Pr(\hat{C} = N\,|\,C = 0) = 0.012$ and $Pr(\hat{C} = N\,|\,C = 1) = 0.012$. As a comparison, the corresponding misclassification rates of the min-max rule (classify as 0 if the log-intensity level is less than -1.075, classify as 1 if the log-intensity is greater than 0.797, and classify as N otherwise) for this probe are 10^{-4}, 0.591, 0.007 and 0.404, respectively. The min-max rule performs well when the clone is a 0, but performs

poorly when the clone is 1. In contrast, the performance of the modified Bayes rule is more balanced and has satisfactory misclassification rates.

For our second example, suppose a user is interested in the minimum expected cost modified Bayes rule. To gain insight into how the rule depends upon ρ, the algorithm described by steps 1-4 in Section 5.2.2 for finding the minimum expected cost modified Bayes rule was used for values of $\rho \in \{1, 2, \ldots, 10\}$. Table 9 shows the optimal rules $d_B^*(y_1; \rho)$ along with their misclassification rates which are labeled as $e10 = Pr(\hat{C} = 1 | C = 0)$, $e01 = Pr(\hat{C} = 0 | C = 1)$, $eN0 = Pr(\hat{C} = N | C = 0)$ and $eN1 = Pr(\hat{C} = N | C = 1)$ for convenience.

ρ	L_0^*	L_1^*	Region	$R_{(0)}^*$	$R_{(0)}$	$R_{(1)}$	$R_{(1)}^*$	$e10$	$e01$	$eN0$	$eN1$	Cost
1.0	0.50	0.50	a	-2.713	-2.713	2.763	2.763	0.080	0.034	0.000	0.000	0.057
2.0	0.50	0.50	a	-2.713	-2.713	2.763	2.763	0.080	0.034	0.000	0.000	0.114
3.0	0.33	0.67	S1	-2.927	-2.481	2.531	2.977	0.060	0.018	0.042	0.044	0.160
4.0	0.25	0.75	S1	-3.038	-2.344	2.394	3.088	0.050	0.013	0.065	0.074	0.195
5.0	0.20	0.80	S1	-3.118	-2.237	2.287	3.168	0.043	0.010	0.082	0.100	0.224
6.0	0.17	0.83	S1	-3.172	-2.160	2.210	3.222	0.039	0.008	0.094	0.121	0.249
7.0	0.14	0.86	S1	-3.233	-2.068	2.118	3.283	0.034	0.007	0.107	0.148	0.271
8.0	0.13	0.87	S1	-3.255	-2.032	2.082	3.305	0.033	0.006	0.112	0.160	0.291
9.0	0.11	0.89	S1	-3.305	-1.952	2.002	3.354	0.029	0.005	0.123	0.187	0.309
10.0	0.10	0.90	S1	-3.332	-1.905	1.955	3.382	0.027	0.005	0.129	0.204	0.325

Table 9. Minimum Expected Cost Modified Bayes Rules for Probe #1

We can see from Table 9 that after $\rho > 2$, all the rules $d_B^*(y_1; \rho)$ correspond to (L_0, L_1) values that belong to the partition S_1 in Figure 2. It is also evident that as ρ increases, the misclassification rates $e10$ and $e01$ decrease while $eN0$ and $eN1$ increase. An interpretation of this last observation is that the intervals where the minimum expected cost modified Bayes classifier predicts N, namely $(R_{(0)}^*, R_{(0)})$ and $(R_{(1)}, R_{(1)}^*)$, get wider as ρ increases.

6.2 Selection of ρ for Minimum Cost Modified Bayes Rules

As discussed in Section 1, if a predicted OFRG fingerprint has too many N values, the subsequent clustering algorithm loses sensitivity and accuracy. As illustrated in Table 9, the probability of an N value increases with the value of ρ, suggesting we should try to select ρ as small as possible. On the other hand, the probability of a hard misclassification (classifying a 0 as 1 or vice-versa) is a decreasing function of ρ, suggesting we should try to select ρ as large as possible. Our solution to this tradeoff is to select the set of $\{\rho_j\}_{j=1}^{K}$ values such that the probability of having more N values than the maximum number that can be tolerated by the clustering algorithm is satisfactorily small, and that otherwise minimize the probability of a hard misclassification.

Assuming *K* probes, we can think of the process of constructing an OFRG fingerprint as a sequence of *K* independent trials where the *j*-th trial corresponds to classifying the *j*-th probe as either 0, 1 or N using the rule $d_B^*(y_j\,;\rho_j)$. The first step of our proposed approach is to select each ρ_j such that probability of an N when we use $d_B^*(y_j\,;\rho_j)$ is constant, say *p* (whose value is to-be determined), for all probes. The consequence of this first step is that the number of N values, say *S*, in a predicted OFRG fingerprint follows a binomial distribution with parameters *K* and *p*. The last step in our approach is to determine *p* by imposing $Pr(\,S > s_0\,) = \alpha$, where α is a (user-supplied) probability and s_0 is the (user-supplied) maximum number of N values that the clustering algorithm can satisfactorily work with. The required value of *p* is then the $p \in (0,1)$ that solves $\sum_{j=0}^{s_0}\binom{K}{j}p^j(1-p)^{K-j} = 1-\alpha$. Once *p* is obtained, each ρ_j is found individually by

finding the largest ρ that satisfies the inequality $Pr\,[\,d_B^*(y_j\,;\rho_j) = N\,] \le p$. Note that the left-hand-side of this last equation can be computed as $p_0 Pr(\hat{C} = N \mid C = 0) + p_1 Pr(\hat{C} = N \mid C = 1)$ with reference to either Table 5b or Table 7b, depending on which of the two variances is larger, for the appropriate conditional probability expressions.

For our application, we use values $\alpha = 0.1$ and $s_0 = 3$, which implies $p = 0.064$. To determine ρ_1, for example, we note from Table 8 that $(\mu_0\,,\sigma_0) = (-4.913\,, 1.563)$ and $(\mu_1\,,\sigma_1) = (-1.259\,, 0.797)$. Figure 6 shows a plot of $Pr\,[\,d_B^*(y_1\,;\rho) = N\,]$ versus ρ, from which it can be inferred that $\rho_1 = 3.77$.

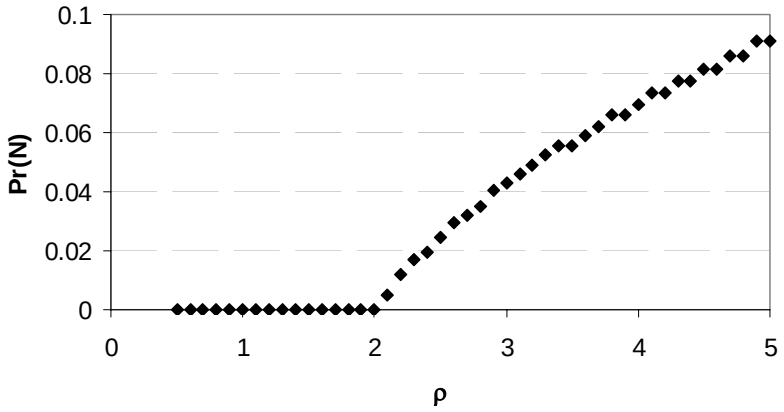

Figure 6. $Pr\,[\,d_B^*(y_1\,;\rho) = N\,]$ vs. ρ for finding $\rho_1 = 3.77$

A similar procedure was repeated for all $K = 28$ probes and Column 2 of Table 10 shows the resulting ρ_j values. Columns 3-4 of Table 10 show the corresponding $(L_0^*\,, L_1^*)$ values, and columns 5-8 show the roots needed to display the classification boundaries of the rule as a

function of the log-intensity y_j. Note that the roots are monotone, left to right, except for probes 23 and 31, which are different because for those two cases we have $\sigma_1 > \sigma_0$ (see Table 8).

Probe	ρ	L_0^*	L_1^*	$R_{(0)}^*$	$R_{(0)}$	$R_{(1)}$	$R_{(1)}^*$
1	3.77	0.27	0.73	-3.009	-2.381	2.431	3.059
2	5.4	0.19	0.81	-4.480	-4.032	-2.625	-2.177
6	2.66	0.38	0.62	-1.972	-1.777	-0.460	-0.265
8	2.89	0.35	0.65	-2.561	-2.143	1.840	2.258
9	2.46	0.41	0.59	-3.393	-3.136	-0.064	0.193
10	2.15	0.47	0.53	-0.943	-0.847	2.809	2.905
11	3.27	0.31	0.69	-2.443	-2.074	0.391	0.760
12	6.06	0.17	0.83	-1.452	-1.045	0.119	0.526
13	2.35	0.43	0.57	0.287	0.443	1.991	2.147
15	2.98	0.34	0.66	-1.789	-1.391	1.703	2.101
21	3.77	0.27	0.73	-2.340	-1.749	3.316	3.907
23	4.08	0.25	0.75	-5.316	-5.620	-0.494	-0.798
29	2.59	0.39	0.61	-1.456	-1.196	10.848	11.108
30	2.81	0.36	0.64	-0.903	-0.541	2.012	2.375
31	3.77	0.27	0.73	-16.865	-17.395	-0.786	-1.316
32	3.07	0.33	0.67	-1.866	-1.517	0.357	0.706
33	2.24	0.45	0.55	-2.745	-2.561	-0.576	-0.392
35	3.39	0.29	0.7	-3.432	-2.843	2.372	2.961
36	40	0.03	0.97	-3.460	-2.074	23.644	25.030
37	3.17	0.32	0.68	-3.150	-2.602	5.452	6.000
40	3.27	0.31	0.69	-4.623	-4.238	-1.674	-1.290
41	2.81	0.36	0.64	-3.393	-3.000	-0.219	0.174
42	4.25	0.24	0.76	-0.920	-0.456	1.434	1.898
43	3.64	0.27	0.73	-2.130	-1.492	2.700	3.338
44	3.39	0.29	0.7	-0.603	-0.129	2.660	3.135
45	5.71	0.18	0.82	-2.122	-1.470	0.878	1.530
46	2.98	0.34	0.66	0.001	0.372	2.757	3.128
48	3.6	0.28	0.72	-3.049	-2.534	1.100	1.616

Table 10. Modified Bayes Rules that Collectively Imply the Probability of Greater than 2 "N" Predictions is Less than 0.10

With Table 7 now in hand, an OFRG fingerprint can be predicted for a new environmental sample. Each probe is hybridized to the sample and the set of log-intensity measure $\{Y_j\}_{j=1}^{28}$ is

obtained. Each of the classifiers shown in Table 7 is then applied, one by one using the $\{Y_j\}_{j=1}^{28}$, to obtain the OFRG fingerprint of the sample. As described in Section 1, this process is repeated on all the experimental clones and a clustering algorithm is then used to group the clones based on similarities of their OFRG fingerprints. A taxonomic identity for each group is then made by selecting several of the clones from each group, obtaining the full sequence the rRNA gene from these clones, and finding the best match of the sequences within a public database containing known microorganism gene sequences [e.g., Benson et al. (2005)].

6.3 Comparison of Modified Bayes Rule and Min-Max Rule

In this Section, we compare the performance of the set of modified Bayes rules given in Table 10 to the set of min-max rules that follow from columns 3 and 7 of Table 8. Performance is assessed by applying both sets of rules to 130 validation clones. The validation clones are like the control clones in that the binding outcome to each of the 28 probes is theoretically known; however they are independent of the control clones and can be used to obtain unbiased estimates of the misclassification rates associated with using each set of rules. In particular, each set of rules makes $130 \times 28 = 3,640$ predictions and by direct comparison with the theoretically known outcomes the misclassification rates can be estimated. The results for the modified Bayes rules and the min-max rules are shown in Table 11 respectively.

True Binding Status	Predicted Binding Status					
	0		1		N	
	MBR	min-max	MBR	min-max	MBR	min-max
0	1680	1886	170	24	87	27
1	61	799	1587	149	55	755

Table 11. Results of Running the Modified Bayes Rules and min-max Rules on 130 Validation Clones

We see from Table 11 that the rate of a hard misclassification (i.e., classifying a 0 as a 1 or vice-versa) rate for the set of min-max rules is $(24+799)/3640 = 22.6\%$, whereas for the MBR rules it is $(170+61)/3640 = 6.35\%$. In addition, the rate of an N classification for the set of min-max rules is $(27+755)/3640 = 21.5\%$, while for the MBR rules it is $(87+55)/3640 = 3.9\%$. The MBR rules clearly outperform the min-max rules with respect to these two metrics. We can also see from Table 11 that the performance of the MBR rules is balanced in the sense that conditional error rates, given the true binding status, are comparable for both values of true binding status. In contrast, the performance of the min-max rules is very unbalanced in that they work well, given that the true binding status is 0, but they work poorly, given that the true binding status is 1.

7. Summary

In this paper, we have extended the classical one-dimensional Bayes binary classifier in a natural way to create a rule that has a region of neutrality, where there weight of evidence is too weak for a confident classification. The rule is derived under a cost structure that weighs the penalty for not making a definite classification against the penalty for making an incorrect definite classification. Explicit details for implementing the rule under assumed normal distributions were provided. The extended classifier was illustrated using a microbial community profiling application where the motivation and need for a region of neutrality was described. The new rule was shown to have superior performance for this application, compared to a commonly used alternative classifier.

References

Amann, R., Wolfgang L., and Schleifer, K.H. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* 59 (1):143-169.

Baldi, P. and Long, A. D. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-tests and statistical inferences of gene changes. *Bioinformatics*, Vol. 17, pp. 509-517.

Benson, D. A. , Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. 2005. GenBank. *Nucleic Acids Research*, Vol. 33, D34-D38.

Bent, E., Yin, B., Liu, Z., Figueroa, A., Jeske, D. R. and Borneman, J. (2005), Oligonucleotide fingerprinting of rRNA genes: a 9,600 clone macroarray for high-resolution studies of complex microbial communities. *Manuscript Under Preparation*.

Borneman, J., and E. W. Triplett. 1997. Molecular microbial diversity in soils from eastern Amazonia: Evidence for unusual microorganisms and microbial population shifts associated with deforestation. *Applied and Environmental Microbiology* 63 (7):2647-2653.

Curtis, T. P., W. T. Sloan, and J.W. Scannell. 2002. Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. USA* 99:10494-10499.

Edwards, D. (2003), Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, Vol. 19, pp. 825-833.

Giles, P. J, and Kipling, D. (2003), Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, Vol. 19, pp. 2254-2262.

Hackl et al. (2004), Analysis of DNA microarray data. *Current Topics in Medicinal Chemistry*, Vol. 4, pp. 1357-1370.

Jampachaisri, K., L. Valinsky, J. Borneman, and S. J. Press. 2005. Classification of Oligonucleotide Fingerprints: Application for Microbial Community and Gene Expression Analyses. Bioinformatics 21:3122-3130.

Li, C. and Wong, W. H. (2001), "Model based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proceedings of the National Academy of Sciences*, Vol. 98, pp. 31-36.

Liu, W. T., T. L. Marsh, H. Cheng, and L. J. Forney. 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and Environmental Microbiology* 63 (11):4516-4522.

Muyzer, Gerard, Ellen C. De Waal, and Andre G. Uitterlinden. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology* 59 (3):695-700.

Press, S. J. 1989. Bayesian Statistics: Principles, Models and Applications. John Wiley, New York.

Schwieger, F., and C.C Tebbe. 1998. A new approach to utilize PCR-single-strand conformation polymorphism for 16S rRNA gene-based microbial community analysis. *Applied and Environmental Microbiology* 64:4870-4876.

Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* 74:5088-5090.

Wolfinger, R. D., et al. (2001), "Assessing gene significance cDNA microarray expression data via mixed models," *Journal of Computational Biology*, Vol. 8, pp. 625-637.

Valinsky, L., G. Della Vedova, T. Jiang, and J. Borneman. 2002. Oligonucleotide fingerprinting of rRNA genes for analysis of fungal community composition. *Applied and Environmental Microbiology* 68 (12):5999-6004.

Valinsky, L., G. Della Vedova, A. J. Scupham, S. Alvey, A. Figueroa, B. Yin, J. Hartin, M. Chrobak, D. E. Crowley, T. Jiang, and J. Borneman. 2002. Analysis of bacterial community composition by oligonucleotide fingerprinting of rRNA genes. *Applied and Environmental Microbiology* 68 (7):3243-3250.

Valinsky, L., A. J. Scupham, G.D. Vedova, Z. Liu, A. Figueroa, K. Jampachaisri, B. Yin, E. Bent, J. Press, T. Jiang, and J. Borneman. 2004. Oligonucleotide fingerprinting of rRNA genes. In *Molecular Microbial Ecology Manual, 2nd. ed*., edited by G. A. Kowalchuk, J. J. de Bruijn, I. M. Head, A. D. L. Akkermans and J. D. van Elsas. New York NY: Kluwer Academic Publishers