Supplementary materials for this article are available at 10.1007/s13253-010-0034-6.

Neutral Zone Classifiers Using a Decision-Theoretic Approach With Application to DNA Array Analyses

Hua YU, Daniel R. JESKE, Paul RUEGGER, and James BORNEMAN

Two-class neutral zone classifiers were recently proposed for use in microbial community profiling applications. These classifiers allow a region of neutrality for cases where probe hybridization outcomes are too ambiguous to have adequate confidence in assigning a "binding" or "no binding" result. In this paper, we generalize the idea of neutral zone classifiers to an arbitrary number of classes and apply it to improve the process of microbial community profiling by considering a third class for the outcome of probe hybridization experiments, "partial binding." We introduce a family of class distributions that uses a mixture of Gaussian distributions as a model for a Box–Cox power transformation of the raw intensity measurements. Stratified cross-validation analyses are used to assess the efficacy of the proposed three-class neutral zone classifier. This article has supplementary material online.

Key Words: Classification; Bayes classifier; Gaussian Mixture model; Microbial Community Profiling.

1. INTRODUCTION

Classification is a procedure in which objects are assigned class labels based on values of attribute variables describing features of the objects. A classifier is an algorithm that realizes this mapping from the feature space to the label set. Based on a set of previously labeled training data, the classifier learns to predict class labels and finally operates on objects with unknown labels. Accuracy is essential to the success of classifiers. Especially in some circumstances, the accuracy of the classifier is so crucial that a wrong prediction may result in extraordinarily high costs. A good example of this is medical diagnosis of

Hua Yu is Research Statistician, Amgen, Inc., One Amgen Center Drive, Thousand Oaks, CA 91320-1799, USA. Daniel R. Jeske (✉) is Professor and Chair, Department of Statistics, Room 2605 STAT-COMP Building, University of California, Riverside, CA 92521, USA (E-mail: *daniel.jeske@ucr.edu*). Paul Ruegger is Post-doc and James Borneman is Professor, Department of Plant Pathology and Microbiology, University of California, Riverside, CA 92521, USA.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 15, Number 4, Pages 474–490
DOI: 10.1007/s13253-010-0034-6

thyroid dysfunction. There are three possible statuses for thyroid function: hypothyroidism (underactive), hyperthyroidism (overactive) and normal (Berardi and Zhang 1999). Misclassifying thyroid conditions may have a variety of potentially harmful results. Patients may be suffering from improper drug or radioactive iodine treatment or even have their thyroid surgically removed if misclassified as hyperthyroidism, while patients may unnecessarily experience a lifelong hormone supplement regiment if they are misdiagnosed as hypothyroidism. Those who actually have a thyroid dysfunction but are mislabeled as normal may suffer from long-term organ damage; in some cases, the error can even lead to death. There exist many other examples which are analogous to medical diagnosis where misclassification can bring serious consequences.

Misclassifications may come from the high similarities of the feature set between two objects, which makes it difficult to distinguish them precisely. For example, a blood test is the main feature used in diagnosing thyroid abnormalities. However, the test result features of under- or overactive thyroid are not consistent across different patients that have the same level of dysfunction. It would be helpful to manage this situation if there were a classifier which can more satisfactorily deal with this uncertainty and imprecision.

Jeske et al. (2007) proposed a classifier with an enlarged action space that includes "no classification" as a prediction outcome. The so-called neutral zone classifier was utilized in a microbial community profiling application. In that application, a classification rule is needed to predict whether a nucleotide probe successfully binds to an rRNA gene. For those binding experiments that are too ambiguous to show enough evidence for a confident prediction, the neutral zone classifier uses the "no classification" outcome because inaccurate prediction of the binding status can confuse a subsequent clustering analysis of rRNA gene fingerprints. "No classification" is a useable classification outcome since the profiling application utilizes multiple probes to obtain the gene fingerprint. The enlarged action space enables the user to minimize the risks associated with misinformation in the fingerprint and be prompted to potentially conduct further investigations that could lead to assigning a crisp label.

In this paper, we generalize the development of the neutral zone classifier to handle k classes, being motivated to do so by an intent to improve the microbial community profiling process via use of "partial binding" as a third possible outcome to the binding experiments. Intuitively, introducing the partial binding outcome should provide more discriminative information for the subsequent clustering analysis of the gene fingerprints, and therefore result in a more accurate microbial taxonomy. Within the context of our application, we quantitatively compare the advantage of the three-class neutral zone classifier.

The rest of this paper is organized as follows. In Section 2 we derive the general form of the k-class neutral zone classifier. In Section 3, we introduce a flexible family of class distributions whose need was motivated by our application. Details of applying the neutral zone classifier to our microbial community profiling application are described in Section 4. Included in Section 4 is a validation study where 5-fold stratified cross-validation analysis is used to evaluate the effectiveness of the three-class neutral zone classifier relative to a reduced two-class neutral zone classifier that merges the no binding and partial binding classes. In addition, the performance of the neutral zone classifier is compared to an

extended version (adapted to handle three classes) of a min-max classifier proposed in Valinsky et al. (2002a, 2002b). Finally, Section 5 summarizes the work presented in this paper.

2. NEUTRAL ZONE CLASSIFIERS

2.1. General Formulation

Suppose there are k possible classes for an object, say $C \in \{0, 1, \ldots, k-1\}$. A classifier based on a $p \times 1$ vector of attributes $\boldsymbol{Y}$ is to be built. We assume the conditional class probability density functions for $\boldsymbol{Y}$, written as $\{f_i(y)\}_{i=0}^{k-1}$, are known or otherwise can be estimated from training data with negligible error. Let $\{\pi_i\}_{i=0}^{k-1}$ denote the (known) a priori class probabilities. The corresponding posterior class probabilities are $p_i(y) = f_i(y)\pi_i / \sum_{j=0}^{k-1} f_j(y)\pi_j$, and the usual Bayes classifier with equal misclassification costs is $\hat{C}(y) = \text{argmax}_{0 \le i \le k-1}\, p_i(y)$.

When the two highest posterior probabilities are very close together, the evidence for a confident classification is weak. In this case, it might be preferable to avoid making a crisp label assignment and instead classify the object as "N," for no classification, with the interpretation being that follow-up is necessary in order to more reliably classify the object. Mathematically, instances when the N classification outcome is needed can be characterized by the condition $p_{(k)}(y) - p_{(k-1)}(y) \le L$, where $p_{(k)}(y)$ and $p_{(k-1)}(y)$ denote the two largest values among $\{p_i(y)\}_{i=0}^{k-1}$, and where $0 \le L \le 1$ is to be determined. The k-class neutral zone classifier can be formally written as

$$\hat{C}_k(y; L) = \begin{cases} \text{argmax}_{0 \le i \le k-1}\, p_i(y), & \text{if } p_{(k)}(y) - p_{(k-1)}(y) > L \\ N, & \text{if } p_{(k)}(y) - p_{(k-1)}(y) \le L. \end{cases} \tag{2.1}$$

2.2. Special Cases

It can be shown that for the two-class problem, $k = 2$, the neutral zone classifier defined by (2.1) can be simplified as

$$\hat{C}_2(y; L) = \begin{cases} 0 & \text{if } p_0(y) > 1/2 + L/2 \\ 1 & \text{if } p_0(y) < 1/2 - L/2 \\ N & \text{if } 1/2 - L/2 \le p_0(y) \le 1/2 + L/2, \end{cases} \tag{2.2}$$

which reveals that when $p_0(y)$ falls in the interval $(1/2 - L/2, 1/2 + L/2)$, there is not enough evidence in the data to make a confident decision about whether the object belongs to class 0 or class 1. The classifier $\hat{C}_2(y; L)$ was used in Jeske et al. (2007).

For the three-class problem, $k = 3$, the neutral zone classifier defined by (2.1) can be simplified as

$$\hat{C}_3(y; L) = \begin{cases} 0 & \text{if } p_1(y) > 1 - 2p_0(y) + L \text{ and } p_1(y) < p_0(y) - L \\ 1 & \text{if } p_1(y) > p_0(y) + L \text{ and } p_1(y) > -p_0(y)/2 + 1/2 + L/2 \\ 2 & \text{if } p_1(y) < 1 - 2p_0(y) - L \text{ and } p_1(y) < -p_0(y)/2 + 1/2 - L/2 \\ N & \text{otherwise.} \end{cases} \tag{2.3}$$

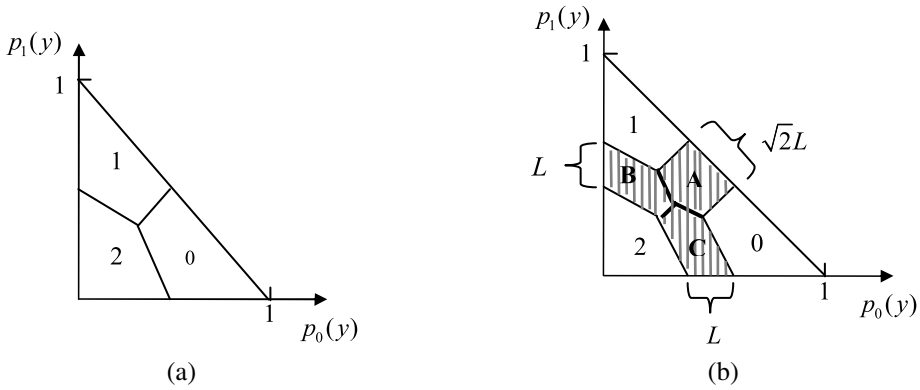

Figure 1. (a) Three-Class Bayes Classifier. (b) Three-Class Neutral Zone Classifier.

Figure 1 contrasts the equal cost Bayes classifier with the neutral zone classifier. Referring to Figure 1a, the point $[p_0(y), p_1(y)]$ maps to a region corresponding to which of $\{p_i(y)\}_{i=0}^{2}$ is the largest. Figure 1b similarly shows the classification regions for the neutral zone classifier. Together, regions A, B and C comprise the neutral zone region. Region A represents the neutral zone between class 0 and 1, region B represents the neutral zone between class 1 and 2, region C represents the neutral zone between class 0 and 2. The following two properties associated with Figure 1b are proved in Yu (2009):

Property 1. *The width of region A is* $\sqrt{2}L$*, while the width of regions B and C are each* L*.*

Property 2. *The area of all three regions* A, B, C *is equal to* $(2 - L)L/6$*.*

2.3. SELECTING L

According to (2.1), we will have k true class labels versus $k + 1$ predicted class labels. A misclassification cost matrix is shown in Table 1, defining the costs for each possible misclassification error. Let $A_0, A_1, \ldots, A_{k-1}$ and A_N denote the regions in the space of Y

Table 1. Costs of Misclassification Errors of k-Class Neutral Zone Classifier.

True Class Label	Predicted Class Label					
	0	1	2	$\cdots$	$k-1$	N
0	0	c_{01}	c_{02}	$\cdots$	$c_{0,k-1}$	$c_{0,N}$
1	c_{10}	0	c_{12}	$\cdots$	$c_{1,k-1}$	$c_{1,N}$
2	c_{20}	c_{21}	0	$\cdots$	$c_{2,k-1}$	$c_{2,N}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$k-1$	$c_{k-1,0}$	$c_{k-1,1}$	$c_{k-1,2}$	$\cdots$	0	$c_{k-1,N}$

corresponding to where the neutral zone classifier predicts the labels $0, 1, \ldots, k-1$ and N, respectively. That is,

$$A_j = \Big\{ y : p_{(k)}(y) - p_{(k-1)}(y) > L \text{ and } j = \underset{0 \le i \le k-1}{\operatorname{argmax}}\, p_i(y) \Big\},$$
$$j = 0, 1, \ldots, k-1,$$
$$A_N = \big\{ y : p_{(k)}(y) - p_{(k-1)}(y) \le L \big\}.$$

The corresponding misclassification probabilities of the neutral zone classifier are then

$$P\big[\hat{C}_k(y; L) = j \,\big|\, C = i\big] = \int_{A_j} f_i(y)\, dy,$$
$$i = 0, 1, \ldots, k-1,$$
$$j = 0, 1, \ldots, k-1, N.$$

The expected cost of misclassification is

$$EC_k(L) = \sum_{j=0}^{k-1} \sum_{i=0}^{k-1} P\big[\hat{C}_k(y; L) = j \,\big|\, C = i\big] \pi_i c_{ij} + \sum_{i=0}^{k-1} P\big[\hat{C}_k(y; L) = N \,\big|\, C = i\big] \pi_i c_{iN}$$

and the minimum expected cost-neutral zone classifier is defined by $\hat{C}_k(y; L^*)$, where $L^* \in [0, 1]$ is the value of L which minimizes $EC_k(L)$. Note that if all non-zero costs are equal, $\hat{C}_k(y; L^*)$ is the equal-cost Bayes classifier, since in this case the right-hand side of $EC_k(L)$ is proportional to the overall probability of misclassification.

3. UNIVARIATE CLASS DISTRIBUTION MODEL

In this section, motivated by the three-class microbial community profiling application which will be discussed in detail in Section 4, we introduce a flexible univariate class distribution model based on Gaussian mixtures. We also discuss evaluation of the misclassification probabilities of the three-class neutral zone classifier under this model.

It has been customary in the literature of DNA array data analysis to model gene intensity measurements using normal distributions (Giles and Kipling 2003) or to transform first with a log transformation and then use normal distributions (Hoyle et al. 2002; Jeske et al. 2007). We seek to extend these approaches by utilizing more general transformations and also Gaussian mixtures. In particular, we assume each of the three underlying class distributions is a two-component Gaussian mixture and prior to fitting that type of model we propose use of a Box–Cox power transformation (Box and Cox 1964) to further improve the goodness of fit.

Denoting the training intensity measurements by $\{x_{ij}\}$, for $i = 0, 1, 2$ and $j = 1, \ldots, n_i$, define

$$y_{ij} = \begin{cases} \frac{x_{ij}^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log x_{ij} & \text{if } \lambda = 0. \end{cases}$$

For $\lambda \in [-2, 2]$, the transformed training data in each class i is assumed to follow a two-component Gaussian mixture model

$$f_i(y; q_{i1}, \mu_{i1}, \sigma_{i1}, q_{i2}, \mu_{i2}, \sigma_{i2}) = q_i \phi\big(y; \mu_{i1}, \sigma_{i1}^2\big) + (1 - q_i)\phi\big(y; \mu_{i2}, \sigma_{i2}^2\big)$$
$$(i = 0, 1, 2)$$

where $\phi(y; \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2 and $q_i \in [0, 1]$ is the mixing proportion for the ith class. We denote the unknown parameters by the vectors $\theta_i = (q_i, \mu_{i1}, \sigma_{i1}, \mu_{i2}, \sigma_{i2})$, $i = 0, 1, 2$, and for a fixed λ we then use the EM algorithm to obtain their conditional MLEs $\tilde{\theta}_i(\lambda) = (\tilde{q}_i(\lambda), \tilde{\mu}_{i1}(\lambda), \tilde{\sigma}_{i1}(\lambda), \tilde{\mu}_{i2}(\lambda), \tilde{\sigma}_{i2}(\lambda))$. Finally, we seek the value $\hat{\lambda}$ which maximizes the profile likelihood for λ:

$$\log L(\lambda) = \sum_{i=0}^{2} \sum_{j=1}^{n_i} \log f_i\big(y_{ij}; \tilde{\theta}_i(\lambda)\big) + (\lambda - 1) \sum_{i=0}^{2} \sum_{j=1}^{n_i} \log x_{ij}.$$

Replacing λ by $\hat{\lambda}$ in the conditional MLEs, the corresponding vectors $\hat{\theta}_i = (\hat{q}_i, \hat{\mu}_{i1}, \hat{\sigma}_{i1}, \hat{\mu}_{i2}, \hat{\sigma}_{i2})$ are the unconditional MLEs of the model parameters.

Based on the Gaussian mixture model, each of the misclassification probabilities can be expressed as

$$\begin{aligned} P\big[\hat{C}_3(y; L) = j \,\big|\, C = i\big] &= \int_{A_j} f_i(y)\, dy \\ &= \int I_{A_j}(y) f_i(y)\, dy \\ &= q_i \int I_{A_j}(y) \phi\big(y; \mu_{i1}, \sigma_{i1}^2\big)\, dy \\ &\quad + (1 - q_i) \int I_{A_j}(y) \phi\big(y; \mu_{i2}, \sigma_{i2}^2\big)\, dy \end{aligned} \tag{3.1}$$

where $i = 0, 1, 2$, $j = 0, 1, 2$, N and $I_{A_j}(y)$ is the usual set indicator function. The means, variances and mixing probabilities in (3.1) can be replaced by their unconditional MLEs.

Analytical evaluation of the integrals in (3.1) is not easy due to the difficulty in inverting the regions of integration. However, rewriting (3.1) in an approximate form as

$$\begin{aligned} P\big[\hat{C}_3(y; L) = j \,\big|\, C = i\big] \approx \hat{q}_i &\int_{L_{i1}}^{U_{i1}} I_{A_j}(y) \phi\big(y; \hat{\mu}_{i1}, \hat{\sigma}_{i1}^2\big)\, dy \\ &+ (1 - \hat{q}_i) \int_{L_{i2}}^{U_{i2}} I_{A_j}(y) \phi\big(y; \hat{\mu}_{i2}, \hat{\sigma}_{i2}^2\big)\, dy \end{aligned} \tag{3.2}$$

where $U_{ik} = \hat{\mu}_{ik} + 5\hat{\sigma}_{ik}$ and $L_{ik} = \hat{\mu}_{ik} - 5\hat{\sigma}_{ik}$ $(k = 1, 2)$, enables the use of Gaussian Legendre quadrature (Givens and Hoeting 2005). For a fixed L, all the misclassification probabilities in $EC_3(L)$ can be calculated based on the quadrature approximations to (3.2), and we can search over $0 \leq L \leq 1$ to find the optimal value L^* which minimizes the expected cost. The corresponding rule $\hat{C}_3(y; L^*)$ is the proposed three-class neutral zone classifier.

4. APPLICATION TO MICROBIAL COMMUNITY PROFILING

4.1. INTRODUCTION TO APPLICATION

Microbial community profiling based on a fingerprinting strategy has been developed by an interdisciplinary team of researchers at the University of California, Riverside (Valinsky et al. 2002a, 2002b, 2004). It provides a novel cost-effective means for extensively analyzing microbial community composition. In what follows, we give a relatively brief sketch of the fingerprinting process, and refer the reader to Figure 1 in Valinsky et al. (2002a) for further details. After conducting DNA extraction and PCR amplification for samples taken from a host (e.g., soil, water, human, etc.), rRNA genes, which represent a specific group of microorganism such as fungal or bacteria, are available (a fungal example is used in this paper). Clone libraries of these rRNA genes are then constructed. The clones are then subjected to a series of hybridization experiments with different 10-base nucleotide probes whose sequences are known. The output of each hybridization experiment is a measured intensity level that carries evidence that binding has occurred between the probe and the rRNA gene clone. Complete binding indicates the whole probe sequence is contained within the full clone sequence. Partial binding would indicate a subset of the probe sequence is contained within the full clone sequence. A precise definition of partial binding is given in Section 4.2 below. In this paper, the intensity level is input into a three-class neutral zone classifier that tries to differentiate no binding (class 0) from either partial binding (class 1) or complete binding (class 2).

A clone fingerprint consists of a vector (one position for each probe) of 0, 1 or 2 elements, plus as many N elements as needed for the cases where a confident crisp label assignment is not possible. Here, the N elements in the fingerprint result from use of our three-class neutral zone classifier. Based on a clustering analysis of these fingerprints, the rRNA gene clones are then grouped into subsets to reflect their similar binding characteristics with respect to the probes. A full nucleotide sequence analysis of a representative clone from each group is then used to find the best match within a public database containing gene sequences of known microorganisms. In this way, information about the composition of the microbial community in the host can be inferred.

The classification problem corresponding to this application is the assignment of labels 0, 1, 2 or N based on the measured intensities of each gene from the probe hybridization experiments. The rationale for introducing the neutral zone classifier is that some of the measured intensities are too ambiguous to provide enough evidence to confidently assign 0, 1 or 2, and a wrong prediction with one or more of the probes for a specific gene can have

more of an adverse consequence on the clustering step than alternatively ignoring the result from those binding experiments. On the other hand, unnecessarily ignoring hybridization outcomes will weaken the fidelity of the clustering step. Consequently, our neutral zone classifier is needed to appropriately make N assignments.

4.2. DATA AND CLASS DISTRIBUTIONS

In the microbial community profiling data discussed in Jeske et al. (2007), two outcomes from the hybridization experiments with the control clones were defined. The binding outcome (denoted by 1) was declared when all 10 of the probe bases match (based on Watson–Crick base-pairing rules) a 10-base long fragment somewhere in the rRNA gene clone sequence. The non-binding outcome (denoted by 0) was declared when no such match could be found. Usually, intensity levels corresponding to binding outcomes are higher than intensity levels corresponding to non-binding outcomes. Our conjecture (substantiated by Figure 2) is that within the class of non-binding outcomes, the intensity level is largely proportional to the numbers of mismatches that occurred, although we do recognize that the position of the mismatch is also a factor. For the microbial profiling application presented in this paper, we define partial binding to be cases where complete binding does not occur but for which there exists a fragment in the rRNA sequence where just one base of the probe does not match. Non-binding cases are then correspondingly the cases where neither complete binding nor partial binding occurs.

In Figure 2, we take the training data (intensity levels for 344 control clones) for probe #3 as an example to illustrate the plausibility and intuition for the partial binding class. Figure 2a shows nonparametric density estimates of the two-class outcome data, complete binding and everything else as non-binding, and Figure 2b shows the effect of dividing the non-binding class to form the partial binding class and the residual non-binding class. It can be seen in Figure 2b that the distribution of the partial binding class sits between that of the residual non-binding class and the unchanged binding class.

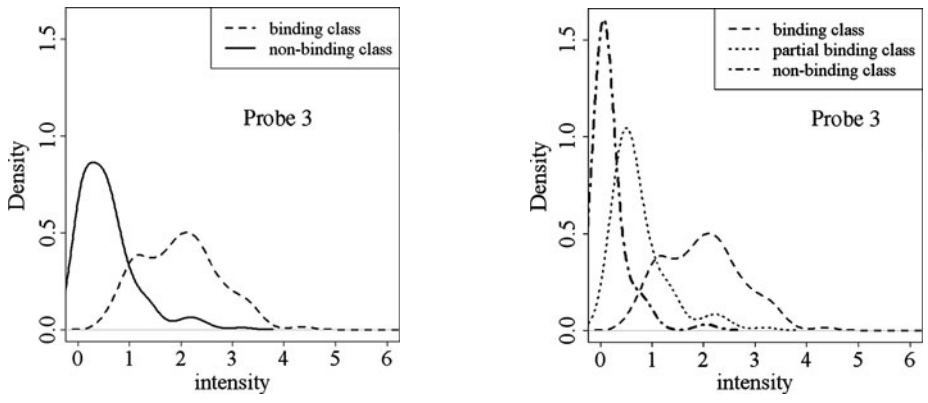

Figure 2. (a) Non-binding and Binding Classes before Introducing Partial Binding Class. (b) Non-binding, Partial Binding and Binding Classes.

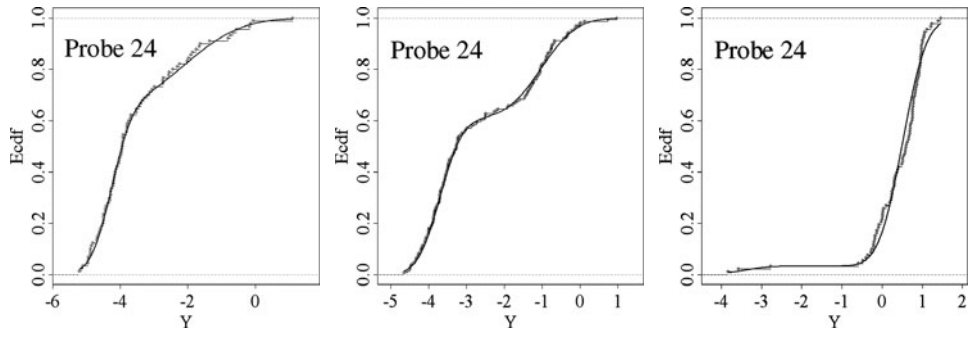

Figure 3. (a) Ecdf and Fitted Class Distribution for Non-binding Data. (b) Ecdf and Fitted Class Distribution for Partial Binding Data. (c) Ecdf and Fitted Class Distribution for Binding Data.

In this application, 33 probes were hybridized to 344 control clones. For each probe, the intensity level data divides into three classes according to the definition outlined above. A single-component Gaussian distribution following a Box–Cox transformation did not show a good fit to the data. For example, QQ-plot diagnostics revealed inadequate fit in the tails as well as bimodal characteristics for many of the probes. Different reasons such as array spot background noises or spot size variation from the printing process, as well as nature itself, may account for this. In a small percentage of cases, the background-subtracted binding intensities from the hybridization experiments on the control clones were negative. Since intensity measurements are inherently non-negative, it is believed that such outcomes reflect errors in the experimental procedure and they were therefore removed from the training data before fitting the class distributions. For future clones, outside the training data set, if negative intensity measurements are observed for a probe-binding event, the neutral zone classifier is defined to classify the outcome as N.

We use the two-component Gaussian mixture model combined with the Box–Cox transformation that was discussed in Section 3 to model the intensity measurements. The table in the Web Appendix shows the class size n_i, the Box–Cox parameter $\hat{\lambda}$, and $\hat{\theta}_i = (\hat{q}_i, \hat{\mu}_{i1}, \hat{\sigma}_{i1}, \hat{\mu}_{i2}, \hat{\sigma}_{i2})$, $i = 0, 1, 2$, for each of the three classes and for each of the 33 probes. Empirical cumulative distribution functions (Ecdfs) of the Box–Cox transformed intensity data with the overlaid fitted Gaussian mixture provide a visual display of the adequacy of the class distribution model. Figure 3 shows the set of Ecdfs for probe #24 and is illustrative of how well the proposed class distribution model fits the data. Similar figures for the remaining probes are included in the Web Appendix.

4.3. THREE-CLASS NEUTRAL ZONE CLASSIFIER

In our application, class labels 0, 1 and 2 correspond to no binding, partial binding and binding, respectively. Similarly to Jeske et al. (2007) we assume a symmetric cost structure. The cost of "hard" misclassification errors, where no binding is misclassified as binding (or vice versa), is denoted by c_1. We assume the cost of misclassifying no

binding as partial binding (or vice versa) is the same as the cost of misclassifying binding as partial binding (or vice versa) and denote these "medium" misclassification error costs as c_2. The cost of making an N classification will be denoted by c_3. We take $c_1 \geq c_2 \geq c_3$ to reflect the severity of hard errors relative to medium errors and the fact that an N classification can be viewed as a relatively "soft" error. With these simplifications, and introducing the shorthand notation $\hat{C}_3 \equiv \hat{C}_3(y; L)$, the expected cost of misclassification becomes

$$\begin{aligned} EC_3(L) = {} & P[\hat{C}_3 = 1 \mid C = 0]\pi_0 c_2 + P[\hat{C}_3 = 2 \mid C = 0]\pi_0 c_1 + P[\hat{C}_3 = N \mid C = 0]\pi_0 c_3 \\ & + P[\hat{C}_3 = 0 \mid C = 1]\pi_1 c_2 + P[\hat{C}_3 = 2 \mid C = 1]\pi_1 c_2 \\ & + P[\hat{C}_3 = N \mid C = 1]\pi_1 c_3 + P[\hat{C}_3 = 0 \mid C = 2]\pi_2 c_1 \\ & + P[\hat{C}_3 = 1 \mid C = 2]\pi_2 c_2 + P[\hat{C}_3 = N \mid C = 2]\pi_2 c_3. \end{aligned}$$

It is not necessary to explicitly specify c_1, c_2 and c_3. Rather, it will suffice to know the ratios $\rho_1 = c_1/c_3$ and $\rho_2 = c_2/c_3$ since we can alternatively write

$$\begin{aligned} EC_3(L) \propto {} & \pi_0\big(\rho_2 P[\hat{C}_3 = 1 \mid C = 0] + \rho_1 P[\hat{C}_3 = 2 \mid C = 0] + P[\hat{C}_3 = N \mid C = 0]\big) \\ & + \pi_1\big(\rho_2 P[\hat{C}_3 = 0 \mid C = 1] + \rho_2 P[\hat{C}_3 = 2 \mid C = 1] + P[\hat{C}_3 = N \mid C = 1]\big) \\ & + \pi_2\big(\rho_1 P[\hat{C}_3 = 0 \mid C = 2] + \rho_2 P[\hat{C}_3 = 1 \mid C = 2] + P[\hat{C}_3 = N \mid C = 2]\big) \\ \equiv {} & h(L). \end{aligned} \tag{4.1}$$

The relationship $c_1 \geq c_2 \geq c_3$ implies $\rho_1 \geq \rho_2 \geq 1$. The values of ρ_1 and ρ_2 can be specified by the user, or alternatively, can be determined by utilizing application-specific constraints. The next two subsections illustrate each of these cases.

4.3.1. Known (ρ_1, ρ_2)

As an illustrative example of this case, suppose the user selects $\rho_1 = 3$ and $\rho_2 = 2$ for probe #24. Following the approach outlined in Section 3, for each value of $L \in [0, 1]$ we first approximate the values of all nine misclassification probabilities based on 2000-point quadrature approximations to (3.2). We then calculate the corresponding value of $h(L)$ from (4.1) using specified values $\rho_1 = 3$ and $\rho_2 = 2$, and taking $\pi_0 = \pi_1 = \pi_2 = 1/3$. Finally, we search over the whole range of L to find the minimum value of $h(L)$. Using probe #24 to illustrate this approach, Figure 4 shows that $h(L)$ obtains its minimum value 0.530 at $L^* = 0.054$.

The discontinuities in the first derivative of $h(L)$ shown in Figure 4 are caused by corresponding discontinuities in some of the non-neutral misclassification probability functions involved in (4.1). We find that some of these functions have concave shapes and fall to zero at values of L between zero and one. Since $h(L)$ is a linear combination of the misclassification probabilities, it will have first derivative discontinuities at those values. Referring to (2.3), the corresponding three-class neutral zone classifier is thus $\hat{C}_3(y; 0.054)$

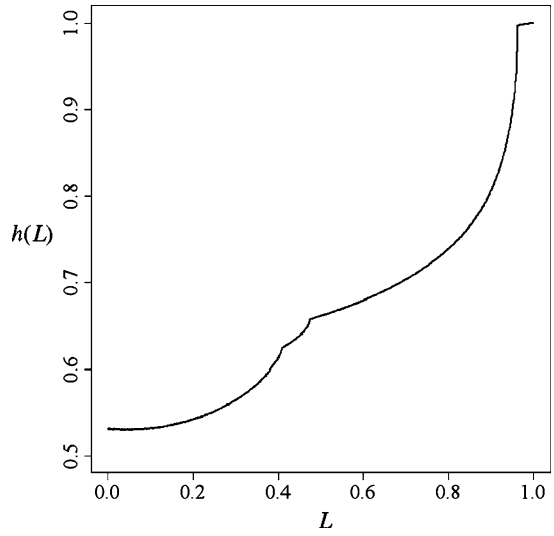

Figure 4. $h(L)$ vs. L for Probe 24 as $\rho_1 = 3, \rho_2 = 2$.

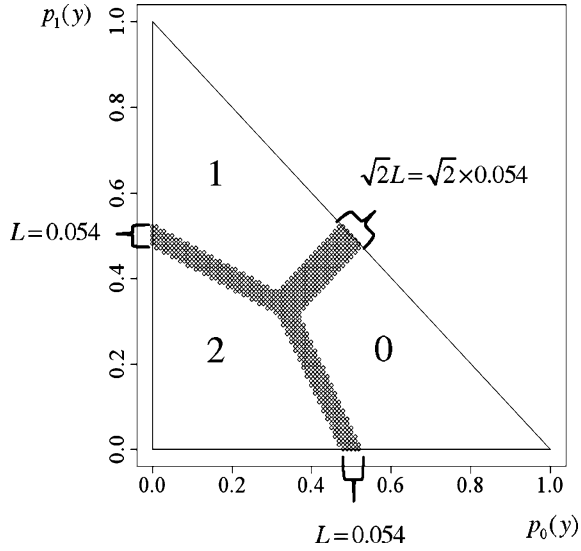

Figure 5. Posterior Representation of Three-Class Neutral Zone Classifier for Probe #24 for Case $\rho_1 = 3, \rho_2 = 2$.

and Figure 5 shows its corresponding two-dimensional posterior representation with the neutral zone represented by the shaded region.

4.3.2. Unknown (ρ_1, ρ_2)

Now suppose the user is unable to specify ρ_1 and ρ_2. Then the classifier can be viewed as a function of ρ_1 and ρ_2 since $L^* \equiv L^*(\rho_1, \rho_2)$ depends on ρ_1 and ρ_2. For this case

we denote the classifier by $\hat{C}_3(y; L^*(\rho_1, \rho_2))$ and, as before, for notational convenience at times we will write $\hat{C}_3 \equiv \hat{C}_3(y; L^*(\rho_1, \rho_2))$.

If a predicted fingerprint has too many N values, the clustering step in the overall profiling process loses sensitivity and accuracy. It can be seen by examining (4.1) that the effect of increasing ρ_1 or ρ_2 is to drive the classifier toward reducing the probability of making a medium or hard error and, simultaneously, increase the probability of making a soft error. Following Jeske et al. (2007), we can think of the process of constructing an OFRG fingerprint as a sequence of K independent trials where the trials correspond to classifying the hybridization outcome of the K probes as either 0, 1, 2 or N.

Suppose the tolerance for N values can be characterized in terms of the proportion α of fingerprints that are allowed to have more than a specified number s_0 of N values. For the jth probe let $L_j^*(\rho_{j1}, \rho_{j2})$ denote the value of L that minimizes $h(L)$ when the cost ratios are equal to (ρ_{j1}, ρ_{j2}). For a given p, define the locus

$$R_j(p) = \big\{(\rho_{j1}, \rho_{j2}) : P\big[\hat{C}_3\big(y; L_j^*(\rho_{j1}, \rho_{j2})\big) = N\big] = p\big\}$$

where we note that

$$P\big[\hat{C}_3\big(y; L_j^*(\rho_{j1}, \rho_{j2})\big) = N\big] = \sum_{i=0}^{2} \pi_i P(\hat{C}_3 = N \mid C = i).$$

Provided each $(\rho_{j1}, \rho_{j2}) \in R_j(p)$, it follows that the number of N values in a fingerprint is a binomial random variable with parameters K and p. Hence, if we take p_0 as the solution to $\sum_{j=0}^{s_0} \binom{K}{j} p_0^j (1 - p_0)^{K-j} = 1 - \alpha$, it would follow that the expected proportion of fingerprints having more than s_0 values equal to N will be α. Aside from this property, we would like to have each (ρ_{j1}, ρ_{j2}) as large as possible to otherwise minimize the probability of medium and hard errors. To this end, we define an aggregated expected cost due to these types of errors as follows:

$$\begin{aligned} EC_{\mathrm{MH}}(\rho_{j1}, \rho_{j2}) \propto \rho_{j2}\big[&\pi_0 P(\hat{C}_3 = 1 \mid C = 0) + \pi_1 P(\hat{C}_3 = 0 \mid C = 1) \\ &+ \pi_1 P(\hat{C}_3 = 2 \mid C = 1) + \pi_2 P(\hat{C}_3 = 1 \mid C = 2)\big] \\ &+ \rho_{j1}\big[\pi_0 P(\hat{C}_3 = 2 \mid C = 0) + \pi_2 P(\hat{C}_3 = 0 \mid C = 2)\big]. \end{aligned}$$

An optimal selection of (ρ_{j1}, ρ_{j2}) is then $(\rho_{j1}^*, \rho_{j2}^*) = \operatorname{argmin}_{(\rho_{j1}, \rho_{j2}) \in R_j(p_0)} EC_{\mathrm{MH}}(\rho_{j1}, \rho_{j2})$.

The set $R_j(p_0)$ can be constructed approximately by employing a numerical search over a set that captures the feasible (application-dependent) space for (ρ_{j1}, ρ_{j2}). In particular, let Υ_j denote a maximum feasible value for ρ_{j1} (we note here that in our application Υ_j will be the same for all j). Let (m, n) denote indices that vary to sweep out a discrete lattice $L_j(\delta) = \{(\rho_{j1,m}, \rho_{j2,n}), 1 \le \rho_{j2,m} \le \rho_{j1,n} \le \Upsilon_j\}$, where δ denotes the step size in each dimension. Define

$$G_j = \big\{(\rho_{j1,m}, \rho_{j2,n}) \in L_j(\delta) : P\big[\hat{C}_3\big(y; L_j^*(\rho_{j1,m}, \rho_{j2,n})\big) = N\big] \le p_0\big\}.$$

Table 2. Three-Class Neutral Zone Classifiers when (ρ_1, ρ_2) Unspecified.

Probe	ρ_1^*	ρ_2^*	$L^*(\rho_1^*, \rho_2^*)$	Probe	ρ_1^*	ρ_2^*	$L^*(\rho_1^*, \rho_2^*)$
1	3.0	2.1	0.136	25	4.3	1.9	0.006
2	2.1	2.1	0.050	26	4.9	1.9	0.153
3	2.1	2.1	0.084	27	2.1	2.1	0.069
4	3.2	2.0	0.092	29	2.4	2.4	0.167
6	2.2	2.2	0.086	30	2.1	2.1	0.045
8	2.1	2.1	0.052	31	3.5	2.0	0.028
9	2.8	2.8	0.302	32	2.2	2.2	0.130
10	2.1	2.1	0.078	33	2.9	2.9	0.307
13	2.7	2.0	0.122	34	3.6	3.6	0.450
15	2.0	2.0	0.028	36	4.7	1.8	0.069
16	2.2	2.2	0.081	37	2.1	2.1	0.062
17	2.0	2.0	0.014	38	2.3	2.3	0.114
19	2.3	2.0	0.054	39	3.7	2.0	0.037
21	1.9	1.9	0.002	40	4.8	2.1	0.084
22	2.1	2.1	0.038	43	3.0	1.9	0.023
23	2.1	1.7	0.006	44	2.7	2.1	0.087
24	3.0	2.0	0.054				

A discrete approximation to the locus $R_j(p_0)$ is the set

$$S_j(p_0) = \Big\{(\rho_{j1,m}, \rho_{j2,n}) \in G_j : (\rho_{j1,m}, \rho_{j2,n}) = \underset{(\rho_{j1,m},\rho_{j2,n})\in G_j}{\operatorname{argmax}} P\big[\hat{C}_3\big(y; L_j^*(\rho_{j1,m}, \rho_{j2,n})\big) = N\big]\Big\}$$

and a corresponding grid approximation to the optimal (ρ_{j1}, ρ_{j2}) is given by $(\rho_{j1}^*, \rho_{j2}^*)_G = \operatorname{argmin}_{(\rho_{j1,m},\rho_{j2,n})\in S_j(p_0)} EC_{\mathrm{MH}}(\rho_{j1,m}, \rho_{j2,n})$.

As an example, suppose $s_0 = 3$ and $\alpha = 0.1$. Since $K = 33$, it follows that $p_0 = 0.054$. Considering again probe #24, taking $\Upsilon_{24} = 5$, and using 0.1 for the grid step size, we find that $(\rho_{j1}^*, \rho_{j2}^*)_G = (3, 2)$ and correspondingly the three-class neutral zone classifier for probe #24 is defined by $L_{24}^* \equiv L_{24}^*(3, 2)$. Table 2 shows the results for all 33 probes.

4.4. Model Evaluation

4.4.1. Benefit of Partial Binding Class

In the previous section, training data from 344 control clones were used to build the classifiers shown in Table 2 for each probe. In this section, we use a 5-fold cross-validation analysis to evaluate the classifiers. For a given probe, the training data divides the control clones into three groups corresponding to no binding, partial binding and binding. Since stratification improves the performance of the regular cross-validation in terms of having lower bias and smaller variance (Kohavi 1995), 5-fold stratified cross-validation is used by equally dividing each of the three groups into five approximately equal-sized subsets to make the distribution of three classes in each fold as similar as possible. For example, according to the table in the Web Appendix, probe 1 has 131 non-binding clones, 93 partial

Table 3. Cross-Validation Estimated Costs of Classifiers for 33 Probes.

Probe	Three-Class Neutral Zone	Two-Class Neutral Zone	Three-Class Min-Max	Probe	Three-Class Neutral Zone	Two-Class Neutral Zone	Three-Class Min-Max
1	0.431	0.748	0.747	25	0.756	0.781	0.949
2	0.622	0.733	0.926	26	0.446	0.729	0.964
3	0.433	0.761	0.772	27	0.374	0.794	0.783
4	0.438	0.757	0.940	29	0.452	0.885	0.953
6	0.516	0.811	0.944	30	0.463	0.754	0.728
8	0.465	0.743	0.855	31	0.623	0.725	0.929
9	0.296	0.978	0.721	32	0.429	0.821	0.951
10	0.540	0.749	0.858	33	0.297	0.983	1.760
13	0.410	0.782	0.943	34	0.279	1.267	2.486
15	0.537	0.687	0.909	36	0.443	0.642	0.899
16	0.618	0.821	0.852	37	0.396	0.711	1.005
17	0.481	0.696	1.029	38	0.321	0.803	0.708
19	0.543	0.783	0.947	39	0.555	0.702	0.970
21	0.547	0.684	0.873	40	0.546	0.749	0.830
22	0.498	0.755	0.814	43	0.535	0.808	0.917
23	0.498	0.612	0.775	44	0.525	0.815	0.926
24	0.533	0.761	0.913				

binding clones, 120 binding clones. After the division, each of the 5-folds approximately has 69 clones in total with 26 non-binding cases, 19 partial binding cases, and 24 binding cases.

Consider now the jth probe. For ith fold ($i = 1, \ldots, 5$), we use the remaining four folds and the corresponding (known) ρ^*_{j1} and ρ^*_{j2} in Table 2 to build a three-class neutral zone classifier and apply this classifier to classify the data in the ith fold. In this way, for each probe, we use five different but similar classifiers to ultimately make an independent classification for each clone in the training data. Empirical misclassification rates are then calculated by comparing the predicted labels to the known class labels of the 344 training data control clones. The empirical misclassification rates can then be used in (4.1) to obtain an estimated cost associated with the classifier. The estimated cost can be used as a figure of merit for the classifier and it is shown in column 2 and column 6 of Table 3.

To get a perspective of how the addition of the partial binding class improves the classification performance, the three-class neutral zone classification model was compared to a two-class neutral zone classification model obtained by merging the training data for partial binding outcomes with non-binding outcomes to reflect what would have been done with the data analysis prior to our introduction of the partial binding class. We continue to assume the same family of Box–Cox mixture Gaussian models for the underlying class distributions, and when building the two-class neutral zone classifier, the reduced (comparing to Table 1) misclassification cost matrix is shown in Table 4 and cost ratio c_1/c_3 is set equal to the value of ρ_1^* that was derived in the context of the three-class model (refer to Table 2). A 5-fold stratified cross-validation analysis is again used to estimate the empirical misclassification rates. However, because there are really three underlying classes, there are seven misclassification rates that can be estimated. In turn, these can be used to

Table 4. Reduced Cost Structure for Two-Class Neutral Zone Classifier.

	Predicted Class Label		
True Class Label	0	2	*N*
0	0	c_1	c_3
2	c_1	0	c_3

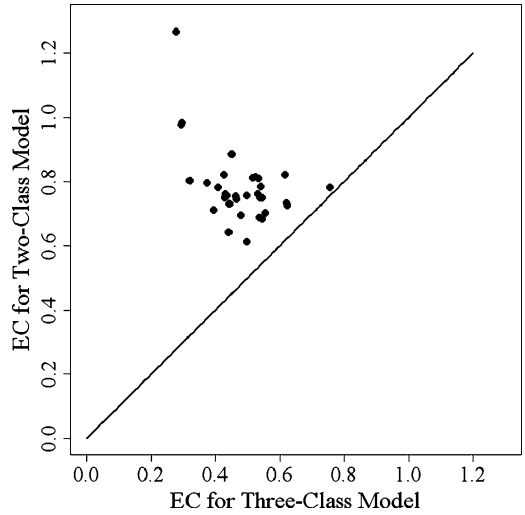

Figure 6. Estimated Costs of Neutral Zone Classifiers for 33 Probes.

obtain the estimated cost associated with use of the two-class neutral zone classifier. The results are shown in column 3 and column 7 of Table 3. A scatter plot of all 33 pairs of estimated costs is shown in Figure 6 where it can be seen that the reduction in cost that is realized by using the correct three-class neutral zone classifier rather than an incorrect two-class neutral zone classifier is appreciable in most cases.

4.4.2. Comparison with a Min-Max Classifier

Valinsky et al. (2002a, 2002b) introduced a naïve but intuitive classifier for assigning *N* to the outcome of hybridization experiments in the two-class context. We refer to this method as the min-max classifier. The intuition for the classifier is that potential intervals for the intensity measurement that include training data from neither or both populations are considered ambiguous and result in assigning the *N* outcome. The min-max classifier was compared to the two-class neutral zone classifier in the illustrative microbial profiling application discussed by Jeske et al. (2007). Here we generalize the min-max classifier to the case of three classes and correspondingly compare its performance to the three-class neutral zone classifier.

The three-class min-max classifier operates directly on the raw intensity measurement, x. Recall from Section 3 that the training data intensity measurements are denoted

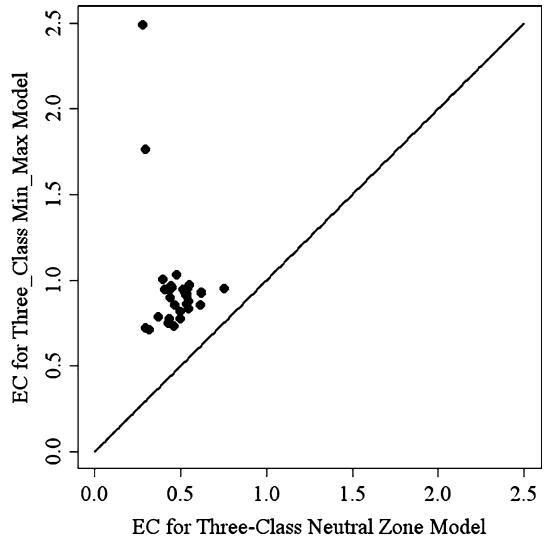

Figure 7. Estimated Costs of Min-Max and Neutral Zone Classifiers for 33 Probes.

as $\{x_{ij}\}$, for $i = 0, 1, 2$ and $j = 1, \ldots, n_i$. Extending the intuition for min-max classifiers to the three-class scenario gives the following three-class min-max classifier:

$$\hat{C}(x) = \begin{cases} 0 & \text{if } x \le \min(\max x_0, \min x_1, \min x_2) \\ 1 & \text{if } \max(\max x_0, \min x_1) \le x \le \min(\max x_1, \min x_2) \\ & \quad \text{and } \max(\max x_0, \min x_1) \le \min(\max x_1, \min x_2) \\ 2 & \text{if } x \ge \max(\max x_0, \max x_1, \min x_2) \\ N & \text{if otherwise.} \end{cases}$$

Columns 4 and 8 in Table 3 show the estimated costs for the 33 probes using the three-class min-max classifier. Figure 7 is the scatter plot of the estimated costs compared to the three-class neutral zone classifier, where again we see better performance from the neutral zone classifier.

5. SUMMARY

Neutral zone classifiers allow regions of neutrality to account for cases where the data is too ambiguous to have adequate confidence in assigning a specific predicted class. They can be used in many areas such as medical diagnosis, safety evaluation and biology. Before this paper, only a two-class neutral zone classifier had been proposed in the literature. Motivated by a DNA array analysis application where we have proposed a new partial binding class for probe hybridization experiments, this paper develops the general form of a k-class neutral zone classifier. Our application also motivated use of a new class distribution model obtained by combining a Gaussian mixture with a Box–Cox transformation. Cross-validation analyses were used to demonstrate superior performance of the three-class neutral zone classifier compared to practical alternative methods.

SUPPLEMENTARY MATERIALS

The Web Appendix referenced in Section 4.2 contains a table showing the fitted parameters of the class distributions for each probe and corresponding figures showing the overlay of empirical distribution functions and fitted models. Both of these supplemental materials are contained in a single.zip archive file.

ACKNOWLEDGEMENT

The research is supported in part by NIH grant 5R01AI078885.

[Accepted July 2009. Published Online June 2010.]

REFERENCES

Berardi, V. L., and Zhang, G. P. (1999), "The Effect of Misclassification Costs on Neural Network Classifiers," *Decision Science*, 30(3), 659–682.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, 26, 211–252.

Giles, P. J., and Kipling, D. (2003), "Normality of Oligonucleotide Microarray Data and Implications for Parametric Statistical Analyses," *Bioinformatics*, 19, 2254–2262.

Givens, G. H., and Hoeting, J. A. (2005), *Computational Statistics*, New York: Wiley-Interscience.

Hoyle, D. D., Rattray, M., Jupp, R., and Brass, A. (2002), "Making Sense of Microarray Data Distributions," *Bioinformatics*, 18, 576–584.

Jeske, D. R., Liu, Z., Bent, E., and Borneman, J. (2007), "Classification Rules that Include Neutral Zones and Their Application to Microbial Community Profiling," *Communication in Statistics—Theory and Methods*, 36 (10), 1965–1980.

Kohavi, R. (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection", in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1137–1143.

Valinsky, L., Vedova, G. D., Jiang, T., and Borneman, J. (2002a), "Oligonucleotide Fingerprinting of rRNA Genes for Analysis of Fungal Community Composition," *Applied and Environmental Microbiology*, 68 (12), 5999–6004.

Valinsky, L., Vedova, G. D., Scumpham, A. J., Alvey, S., Figueroa, A., Yin, B., Hartin, J., Chrobak, M., Crowley, D. E., Jiang, T., and Borneman, J. (2002b), "Analysis of Bacterial Community Composition by Oligonucleotide Fingerprinting of rRNA Genes," *Applied and Environmental Microbiology*, 68 (7), 3243–3250.

Valinsky, L., Scupham, A. J., Vedova, G. D., Liu, Z., Figueroa, A., Jampachaisri, K., Yin, B., Bent, E., Press, J., Jiang, T., and Borneman, J. (2004), "Oligonucleotide Fingerprinting of rRNA Genes," in *Molecular Microbial Ecology Manual* (2nd. ed.), eds. G. A. Kowalchuk, J. J. de Bruijn, I. M. Head, Akkermans, A. D. L., and J. D. van Elsas, New York, NY: Kluwer Academic.

Yu, H. (2009), "Neutral zone classifiers within a decision-theoretic framework," Ph.D. Dissertation, University of California, Riverside, CA.