Olfactory Classification via Interpoint Distance Analysis

Carey E. Priebe

Abstract—Detection of the presence of a single prespecified chemical analyte at low concentration in complex backgrounds is a difficult application for chemical sensors. This article considers a database of artificial nose observations designed specifically to allow for the investigation of chemical sensor data analysis performance on the problem of trichloroethylene (TCE) detection. We consider an approach to this application which uses an ensemble of subsample classifiers based on interpoint distances. Experimental results are presented indicating that our nonparametric methodology is a useful tool in olfactory classification.

Index Terms—Ensemble classifiers, combining classifiers, nonparametric, nearest-neighbor, interpoint distance, rank statistic, subsample statistic, functional data, artificial nose, electronic nose, analytical chemistry, chemometrics.

1 Introduction

It is known that some animals have very effective olfactory systems that can detect concentrations of odorants as low as parts per trillion. In the last two decades, there has been significant progress in the neuroscience of olfaction (see, for instance, the recent special issue of *Science* [29] and references contained therein), as well as in developing sensor arrays for odor detection, including efforts in designing vapor-sensitive instruments that mimic the effectiveness of the mammalian nose (see, for instance, the recent special issue of *IEEE Spectrum* [17] and references contained therein). Artificial "electronic noses" which can identify and measure odor have been developed for applications, such as environmental monitoring, health care, and quality assurance [19], [24]. The continuing development and evaluation of these devices requires what has been dubbed "olfactory signal processing and pattern recognition" [10]. This paper presents a novel nonparametric methodology for statistical olfactory analysis and experimental results indicating the utility of the proposed methodology.

We consider a cross-reactive optical-fiber sensor array that can identify individual vapors. Our task is to identify an odorant sample. (Ultimately, it may also be of interest to estimate concentration once the odorant sample is identified, but our simplified task is, nonetheless, the fundamental first step.) An available training database D_n consists of observations at various concentration levels for each odorant in a library. Our goal is to construct a classifier g so that, given an unidentified odorant observation z, $g(z|D_n)$ will be a statistically reliable estimate of the associated odorant class.

• *The author is with the Department of Mathematical Sciences, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21218-2682. E-mail: cep@jhu.edu.*

Manuscript received 28 Dec. 1999; revised 2 Nov. 2000; accepted 28 Nov. 2000.
Recommended for acceptance by T.K. Ho.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 111138.

In Section 2, we describe the character of the artificial nose data under consideration and the associated "needle in the haystack" detection task. Section 3 details the statistical classification methodology we employ—an ensemble of subsample classifiers based on interpoint distances. Example results are presented in Section 4. Section 5 provides a short discussion of the implications of our results to chemical sensor data analysis.

2 Data

There are numerous technological approaches to artificial nose sensor development ([24], Table 1, p. 29) and there are commercially available instruments based on some of these technologies ([24], Table 2, p. 31). Optical-fiber technology for artificial noses, currently in the research stage, offers cheap and easily fabricated sensors which can be arrayed for simultaneous sensitivity to a wide range of specific molecules [24].

2.1 Tufts Optical-Fiber Artificial Nose Data

We consider data taken from an optical system constructed at Tufts University. The sensor fabrication and preparation [33] results in a sensor for which a change in fluorescence intensity is in response to a change in the molecular environment of a solvatochromic dye due to interactions of a polymer matrix with an analyte present [7].

The Tufts data are obtained from a 19-fiber bundle. An observation is obtained by passing an analyte (a single compound or a mixture) over the fiber bundle in a four second pulse, or "sniff." The information of interest is the change over time in emission fluorescence intensity of the dye molecules for each of the 19 fiber-optic sensors (see Fig. 1).

Data collection consists of recording sensor responses to various analytes at various concentrations. Each observation is a measurement of the fluorescence intensity response at each of two wavelengths (620 nm and 680 nm) for each sensor in the 19-fiber bundle as a function of time. Thus, the sensor produces highly multivariate (38-dimensional) *functional*

0162-8828/01/$10.00 © 2001 IEEE

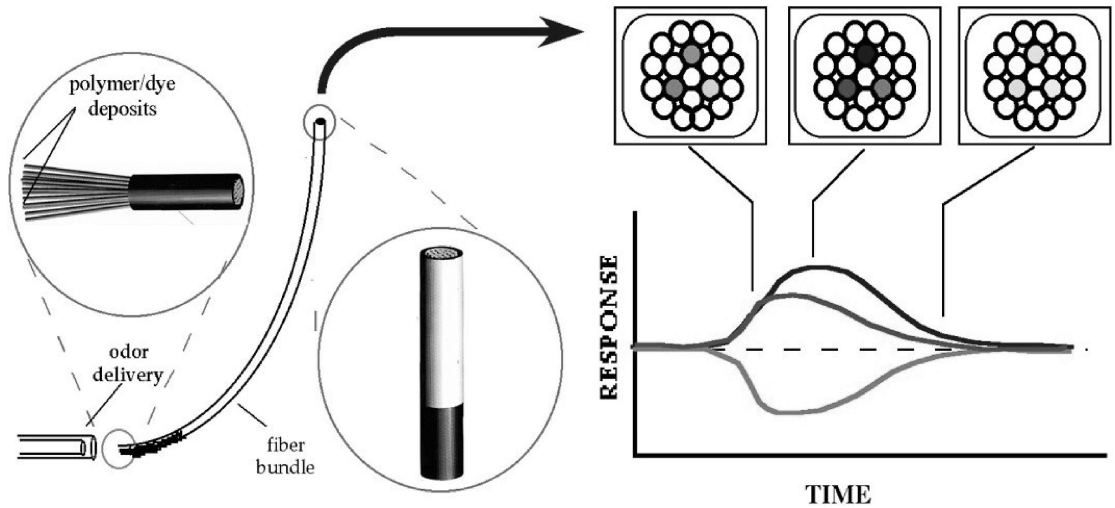


Fig. 1. The plot represents sensor/analyte signatures for three sensors within the bundled 19-sensor array. The data being analyzed for this project are signature patterns of fluorescence change versus time for various analyte mixtures at various concentrations. (This figure was published in Nature, vol. 382: pp. 697-700 (1996); http//:www.nature.com, reprinted by permission.)

observations $x_i^{\phi,\lambda}(t)$, $i = 1, \ldots, n$, for fibers $\phi \in \{1, \ldots, 19\}$ and wavelengths $\lambda \in \{1, 2\}$. While the process is naturally described as functional with t ranging over a 20 second interval, the data as collected are discrete with the 20 seconds recorded at 60 equally spaced time steps for each response. Construction of the database involves taking replicate observations for the various analytes.

The sensor responses are inherently aligned due to the "sniff" signifying the beginning of each observation. The response for each sensor for each observation is normalized by manipulating the individual sensor baselines. This preprocessing consists of subtracting the background sensor fluorescence (the intensity prior to exposure to the analyte) from each response to obtain the desired observation: the change in fluorescence intensity for each fiber at each wavelength. Functional data analysis smoothing techniques are utilized to smooth each sensor response [27].

2.2 Database for a "Needle in the Haystack" Detection Task

The task we address is the identification of an odorant observation z. Specifically, we consider the detection of trichloroethylene (TCE) in complex backgrounds. (TCE, a carcinogenic industrial solvent, is of interest as the target due to its environmental importance as a groundwater contaminant.) The objective is to classify observations as TCE-present ($g(z) = 1$) or TCE-absent ($g(z) = 0$). In addition to TCE in air, eight diluting odorants are considered: BTEX (a mixture of benzene, toluene, ethylbenzene, and xylene), benzene, carbon tetrachloride, chlorobenzene, chloroform, kerosene, octane, and Coleman fuel. Dilution concentrations of 1:10, 1:7, 1:2, 1:1, and saturated vapor are considered. Fig. 2 presents example (unsmoothed) sensor response signals indicating the importance of analyte mixture type, analyte mixture presentation, and fiber band.

Class 0, the TCE-absent class, consists of $n_0 = 352$ observations; the database D_n contains 32 observations of pure air and

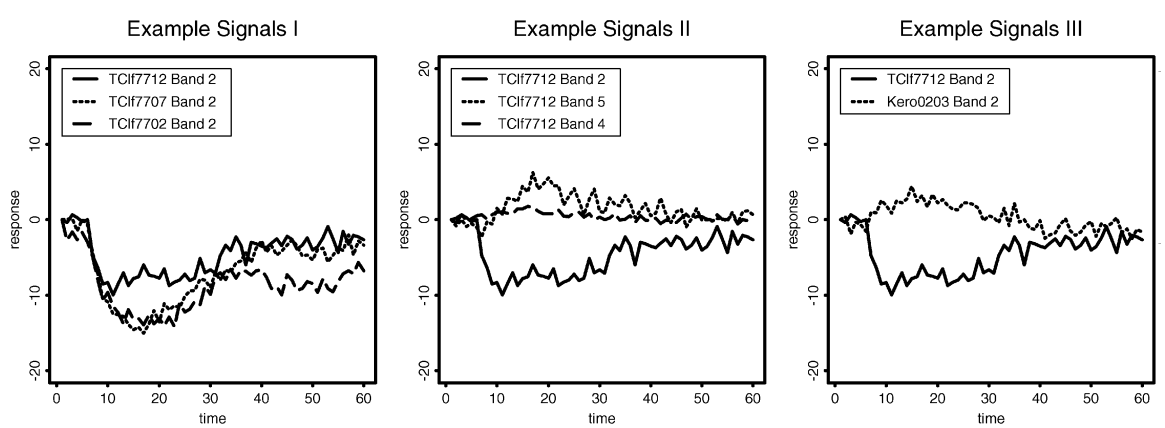


Fig. 2. Depicted are three (unsmoothed) sensor response signal examples: a comparison of a single fiber band for three different presentations of the same analyte mixture (left panel), a comparison of three different fiber bands for a single analyte mixture presentation (middle panel), and a comparison of the same fiber band for two different analyte mixture presentations (right panel).

40 observations of each of the eight diluting odorants at various concentrations in air. There are likewise $n_1 = 760$ class 1 (TCE-present) observations; 40 observations of pure TCE, 80 observations of TCE diluted to various concentrations in air, and 80 observations of TCE diluted to various concentrations in each of the eight diluting odorants in air are available. Thus, there are $n = n_0 + n_1 = 1,112$ observations in the training database D_n. This database is well-designed to allow for an investigation of the ability of the sensor array to identify the presence of one target analyte (TCE) when its presence is obscured by a complex background; this is referred to as the "needle in the haystack" problem.

3 METHODOLOGY

Consider a training database $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ of $d = 38$-dimensional functional random variables X_i representing odorant observations and their associated class labels $Y_i \in \{0, 1\}$ for $i = 1, \cdots, n$. The $X_i^{\phi,\lambda}$ represent signals which are 20 seconds in duration. For each fiber, ϕ and wavelength λ, $X_i^{\phi,\lambda}$ is defined and assumed continuous for time $t \in \mathcal{T} = [0, 20]$, and the X_i take their values in $\mathcal{C}(\mathcal{T})^d$. (Here, $\mathcal{C}(\mathcal{T})$ represents continuous functions on $\mathcal{T}$.) The database D_n consists of n_0 observations from class 0 and n_1 observations from class 1, for $j = 0, 1$, we have $C_j = \{X_i : Y_i = j\}$, $|C_j| = n_j$ and $n = n_0 + n_1$. (The class-conditional training sample sizes n_j are taken to be *design variables* rather than random variables.) For $X_i \in C_j$, the X_i are assumed independent and identically distributed F_j.

The statistical classification methodology proposed here is an ensemble approach to classifier construction [11], [8], [20]. Subsamples of the training data set are used to define the subclassifiers which make up the ensemble. Given a random unidentified odorant observation Z, with class label Y_Z unobserved, each (nonparametric) subclassifier is based on the ranks of the class-conditional interpoint distances from Z to the elements of the subsample. The ensemble classifier is then given by a vote of the subclassifiers.

3.1 A Generalized Wilcoxon-Mann-Whitney Classifier

Our task is to design a classifier $g(Z|D_n)$ with the property that the probability of classification error $L_n = P[g(Z|D_n) \neq Y_Z]$ is as near as possible to the minimal *Bayes error* L^* [6], [21]. Chemistry provides no parametric functional model for the response curves or for the discriminant region boundaries. Exploratory investigation of the response signals (smoothed versions of the examples presented in Fig. 2, for example) suggests no simple parametric functional model is appropriate. Thus, nonparametric functional data discriminant analysis is called for. A common approach to the distribution-free analysis of two or more high-dimensional samples involves the consideration of the interpoint distances [22]. Let $\rho : \mathcal{C}(\mathcal{T})^d \times \mathcal{C}(\mathcal{T})^d \to [0, \infty)$ be an arbitrary (pseudo-) distance applicable to the functional data in question, we consider the class-conditional interpoint distances $\{\rho(Z, X_i) : X_i \in C_j\}$ for $j = 0, 1$. For instance, the L_2 distance is given by

$$\rho(X_1, X_2) = \left(\sum_{\phi=1}^{19} \sum_{\lambda=1}^{2} \int_{\mathcal{T}} (X_1^{\phi,\lambda}(t) - X_2^{\phi,\lambda}(t))^2 dt \right)^{1/2}. \quad (1)$$

Other examples of particular interest considered in the sequel include L_1 (sum of integrated absolute differences) and pseudodistances which ignore or weight individual functional dimensions representing particular fibers ϕ and/or wavelengths λ. These latter pseudodistances are useful for exploratory data analysis and sensor design purposes in addition to classification. Since the fibers have different physical characteristics, different weighting for each fiber may be appropriate. Different pseudodistances yield different dimensionality reduction or "feature extraction" (see Section 4.5).

In practice, some form of functional smoothing is appropriate due to sensor noise. We discuss the specifics of our choice of smoother in Section 4.4 below. Letting $s : \mathcal{C}(\mathcal{T})^d \to \mathcal{C}(\mathcal{T})^d$ represent the particular smoothing procedure employed, we consider the smoothing-derived (pseudo-) distance ρ' defined as $\rho'(X_1, X_2) = \rho(s(X_1), s(X_2))$. The class-conditional interpoint distances under consideration are then $\{\rho'(Z, X_i) : X_i \in C_j\}$ for $j = 0, 1$. (Note that even when ρ is a legitimate distance, ρ' may be a pseudodistance since the smoother may take $X_1 \neq X_2$ to $s(X_1) = s(X_2)$.) For more general applications, a more sophisticated distance employing some form of time—warping may be appropriate. The experimental design and data collection used here yields observations which are temporally aligned and makes the additional sophistication and complexity of such a method unnecessary.

The choice of (pseudo-) distance ρ and the functional smoothing s employed are interdependent aspects of the classifier design which require exploratory analysis and experimental investigation. This topic will be revisited in Section 4.5.

Fig. 3 presents example empirical class-conditional interpoint distance distributions using the L_2 metric and an experimentally determined level of functional smoothing. For the two (arbitrarily selected) exemplars considered in Fig. 3, analysis of class-conditional location parameter estimates for the interpoint distance distributions will yield correct classification. Indeed, the proposed classifier presented here will be interpreted in precisely this framework.

The individual subsample classifiers which make up our ensemble classifier are constructed based on the distances from the unidentified odorant observation Z in question to subsets of the class-conditional training data. To begin building the subclassifier, subsets $S_j \subset C_j$ of size r_j are chosen from the class j training data, $j = 0, 1$. Then, for some $k_j \leq r_j$, the distance from Z to the k_1th nearest element of S_1 is compared to the distance from Z to the k_0th nearest element of S_0. Define $\rho_k(Z, S) = \rho(Z, X)_{(k)}$, where the order statistic $\rho(Z, X)_{(k)}$ represents the kth smallest of the distances $\{\rho(Z, X)_i = \rho(Z, X_i) : X_i \in S\}$. Then, the subsample classifier

$$g_{S_0,S_1}(Z; r_0, r_1, k_0, k_1) = I_{\{\rho_{k_1}(Z,S_1) \leq \rho_{k_0}(Z,S_0)\}} \quad (2)$$

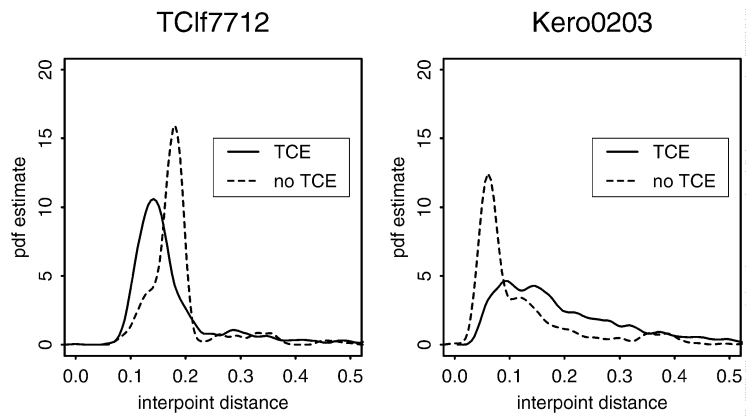


Fig. 3. Interpoint distance distributions for olfactory classification. Depicted are class-conditional probability density estimates for the interpoint distances from a given analyte to the library of training samples. The task is the detection of the presence of TCE (trichloroethylene) in complex backgrounds. The left panel compares the distances from an analyte containing TCE in chloroform to the exemplars in the library which contain TCE with the distances to the exemplars which are TCE-free. The right panel presents the analogous results for a test analyte (kerosene) which does not contain TCE. We see that the class-conditional interpoint distance distributions reflect the presence or absence of TCE.

based on class-conditional subsamples S_0 and S_1 identifies to which class-conditional training subset the unidentified observation Z is "closest." ($I_{\{\}}$ denotes the indicator function.) The subsample classifiers (2) are parameterized by the class-conditional subset sizes r_0, r_1 and the ranks k_0, k_1 used to define closeness.

The proposed classifier considers a vote of all such subsample classifiers. For $r_j \leq n_j$, the set of all appropriate subsample pairs is

$$\Delta(n_0, n_1, r_0, r_1) = \{(S_0, S_1) : S_j \subset C_j \text{ and } |S_j| = r_j\}, \quad (3)$$

where $|S|$ denotes the cardinality of the set S. An ensemble vote of subsample classifiers of the form (2), given by

$$\tau(Z; n_0, n_1, r_0, r_1, k_0, k_1) = (1/|\Delta|) \sum_{S_0 \subset C_0} \sum_{S_1 \subset C_1} g_{S_0,S_1}(Z), \quad (4)$$

yields the test statistic of interest. The statistic τ defined in (4) represents a tally of the number of pairs $(S_0, S_1) \in \Delta$ such that the distance from Z to the k_1th nearest element of S_1 is less than or equal to the distance from Z to the k_0th nearest element of S_0, normalized by the total number of pairs $|\Delta|$. Thus, the proposed classifier is related to previous methods which consider combination of nearest-neighbor classifiers; see, for instance, [30] and references contained therein.

Here, the three decision problem of labeling Z as class 0, class 1, or "no decision" is approached from the viewpoint of deciding, for some τ_0, $\tau(Z) < \tau_0$, $\tau(Z) > \tau_0$, or $\tau(Z) = \tau_0$, as in, for example, [1, p. 183]. Large values of τ indicate that class 1 wins the vote, while small values of τ favor class 0. More precisely, given the null distribution ($H_0 : F_0 = F_1$) F_τ of the statistic τ (which depends on $k_0 \leq r_0 \leq n_0$ and $k_1 \leq r_1 \leq n_1$) and confidence parameters $\alpha, \beta \geq 0$ such that $\alpha + \beta \in [0, 1]$, our proposed ensemble classifier takes the form

$$\begin{aligned} g(Z; n_0, n_1, r_0, r_1, k_0, k_1, \alpha, \beta) = \\ I_{\{F_\tau(\tau(Z)) > 1-\alpha\}} - I_{\{\beta \leq F_\tau(\tau(Z)) \leq 1-\alpha\}}. \end{aligned} \quad (5)$$

(A classification of -1 corresponds to "*no decision*.") Choosing $\alpha = \beta$ treats both types of classification error equally; $\alpha = \beta = 1/2$ yields

$$\begin{aligned} g(Z) &= I_{\{F_\tau(\tau(Z)) > 1/2\}} - I_{\{F_\tau(\tau(Z)) = 1/2\}} \\ &= I_{\{\tau(Z) > F_\tau^{-1}(1/2)\}} - I_{\{\tau(Z) = F_\tau^{-1}(1/2)\}} \end{aligned} \quad (6)$$

and demands that a decision always be made (unless the observed value of τ is precisely $F_\tau^{-1}(1/2)$, the median of F_τ), while $\alpha = \beta \in (0, 1/2)$ allows for no decision unless τ provides sufficient evidence in favor of one class or the other. Choosing $\alpha > 1/2$ forces the classification of Z as class 1 unless there is overwhelming evidence in favor of class 0.

Two special cases in (6) are particularly noteworthy. With $r_0 = n_0$, $r_1 = n_1$, and $k_0 = k_1 = 1$, classification via (6) is the well-known 1-nearest-neighbor method [4], [6], [21]. At the other extreme, with $r_0 = r_1 = k_0 = k_1 = 1$, (4) becomes

$$\tau(Z) = (1/(n_0 n_1)) \sum_{X_i \in C_1} \sum_{X_j \in C_0} I_{\{\rho(Z, X_i) \leq \rho(Z, X_j)\}} \quad (7)$$

and the statistic τ is identified as the classical Wilcoxon-Mann-Whitney statistic [34], [23] applied to the class-conditional interpoint distance samples. In this case, the classifier (6) uses that most popular distribution-free statistic for testing equality of location; under the null hypothesis H_0, the sampling distribution F_τ is symmetric about $\tau_0 = F_\tau^{-1}(1/2) = 1/2$ and Z is labeled as belonging to the class whose interpoint distance distribution median is determined to be smallest (nearest to zero).

In general, the statistic τ is a location parameter estimate. The probability parameter of interest is

$$T(r_0, r_1, k_0, k_1) = \int_0^\infty F_{\rho_{k_1}(Z,S_1)} dF_{\rho_{k_0}(Z,S_0)}, \quad (8)$$

where S_0 and S_1 are random class-conditional subsamples of size r_0 and r_1 and the $F_{\rho_{k_j}(Z,S_j)}$ represent the distribution function for the (univariate) class-conditional random

variables $\rho_{k_j}(Z, S_j)$. Thus, τ represents a generalization based on subsample order statistics [35] of the Wilcoxon-Mann-Whitney statistic. The statistic τ is a U-statistic and can be shown to have an asymptotically normal distribution [35]. However, for small sample sizes—and especially for unbalanced designs or unequal subset sizes k_j, in which case F_τ is skewed—inference based on the asymptotic distribution is inappropriate; the exact distribution (or a small-sample approximation thereof—see Section 3.4) is necessary. A recurrence for the exact distribution F_τ under the null hypothesis H_0 is given in [25] for $k_0 = k_1 = 1$. Classification based on the generalized Wilcoxon-Mann-Whitney statistic τ given in (4) is particularly relevant to interpoint distance-based nonparametric discriminant analysis [25]. Different choices for the parameters yield desirable power characteristics against different alternatives [35]. The issue of *adaptively* selecting the parameters r_0, r_1, k_0, k_1 will be addressed in Section 3.6.

3.2 Extension to $K > 2$ Classes

The proposed classifier has been developed for the simple two-class problem. The extension to $K > 2$ classes can be addressed in two ways. The statistic τ can be generalized to the K sample case and a recurrence for the joint distribution $F_{\tau_1,\ldots,\tau_K}$ is available [25]. Another approach is that of addressing the K class problem through consideration of a collection of two class subproblems [9], [14].

3.3 Relationship to Machine Learning

From a machine learning perspective, the classifier g based on τ is a classic example of "classification by ensemble" [11], [8], [20]. The statistic τ represents the most fundamental approach to constructing ensembles of classifiers. In a manner similar to bagging [2], subsamples S_0 and S_1 are taken (without replacement) from the training database D_n, and $I_{\{\rho_{k_1}(Z,S_1)\le\rho_{k_0}(Z,S_0)\}}$ represents a classifier for Z based on these subsamples. Thus, all possible subsample classifiers obtained are then combined in τ via the simplest possible method for combining individual classification decisions from an ensemble of classifiers: an unweighted vote.

Observing a value of $\tau > 1/2$ means that a majority of the subclassifiers $I_{\{\rho_{k_1}(Z,S_1)\le\rho_{k_0}(Z,S_0)\}}$ in the ensemble favors class 1. A more appropriate classification criterion is the event $\{I_{\{F_\tau(\tau(Z)>1/2)\}} = 1\}$. This event represents evidence in favor of class 1 vs. class 0 in that the vote count is *probabilistically large* (under H_0). Thus, the classifier proposed in (6), based on the unweighted ensemble vote τ, accounts for the character of the distribution F_τ. As noted above, this distribution is strongly influenced by unequal sample sizes ($n_0 \neq n_1$), unequal subset sizes ($r_0 \neq r_1$), and/or unequal rank choices ($k_0 \neq k_1$). In effect, (6) implicitly weights the ensemble votes in a probabilistically appropriate way.

The combination methodology employed here is straightforward and allows for analysis via mathematical statistics. However, more elaborate combination methods such as those presented in [20] may be beneficial in terms of classification performance. In particular, using fewer carefully selected subsets so that the ensemble does not employ so many classifiers is worthy of consideration, but makes the analysis of the statistic significantly more difficult.

3.4 Computational Considerations

For large sample sizes such as those encountered in the olfactory classification task, the calculation of the observed value of τ via (4) and of the distribution F_τ via the available recurrence, are computationally intensive exercises. For the example, results presented in Sections 4.2 and 4.3, the following estimators are used.

Let S_u be a uniform random sample of size u from the collection of subset pairs Δ. The estimator for τ is given by

$$\hat{\tau} = (1/u) \sum_{(S_0,S_1)\in S_u} I_{\{\rho_{k_1}(Z,S_1)\le\rho_{k_0}(Z,S_0)\}}. \tag{9}$$

The estimator standard deviation $\sigma_{\hat{\tau}} \le 1/(2\sqrt{u})$, indicating how large u must be taken (and, consequently, the required computational demand) in order to have an estimator with some prescribed accuracy. Equation (9) can be employed using either observed data or sequences generated under the null hypothesis. To obtain the quantile estimator for F_τ, we consider a collection $\{\hat{\tau}_1, \ldots, \hat{\tau}_v\}$ of such estimators taken under H_0. Then,

$$\hat{F}_\tau(t) = (1/v) \sum_{i=1}^{v} I_{\{\hat{\tau}_i \le t\}} \tag{10}$$

with an accuracy dependent on u, v, and t.

3.5 Classifier Consistency

Our hypothesis testing approach to the two-class decision problem ([1], p. 183) allows us to address the issue of "classifier consistency" from the standpoint of consistent tests of hypotheses. Note that the null hypothesis $H_0 : F_0 = F_1$ implies $F_{\rho(z,X_i|Y_i=0)} = F_{\rho(z,X_i|Y_i=1)}$ for any fixed observation z. For simplicity, consider as alternative hypotheses *stochastic ordering*; the random variable $\rho(z, X_i|Y_i = 0)$ is defined to be stochastically smaller than $\rho(z, X_i|Y_i = 1)$, denoted as

$$\rho(z, X_i|Y_i = 0) <^{st} \rho(z, X_i|Y_i = 1),$$

if

$$F_{\rho(z,X_i|Y_i=0)}(x) \ge F_{\rho(z,X_i|Y_i=1)}(x)$$

for every x, with strict inequality for at least one x. From Fig. 3, we see that, for the left panel in which the test observation (TClf7712) is TCE-present (class 1), we have $\rho(z, X_i|Y_i = 0) >^{st} \rho(z, X_i|Y_i = 1)$. For the TCE-absent observation (Kero0203) depicted in the right panel of Fig. 3, $\rho(z, X_i|Y_i = 0) <^{st} \rho(z, X_i|Y_i = 1)$ as desired for a class 0 observation. For situations such as these, we have the following result.

Theorem. *For a fixed observation z, the classifier g given in (6) based on the statistic τ given in (4) is consistent against alternatives of stochastic ordering of the class-conditional interpoint distance distributions.*

Proof. For fixed values of r_0, r_1, k_0, k_1, as $n_0, n_1 \to \infty$ with $n_0/(n_0 + n_1) \to \zeta \in (0, 1)$, $\tau(z)$ is asymptotically normal under H_0 and $lim\ F_\tau^{-1}(1/2) = T\ a.s.$, where T is given by (8). Under $H_A : \rho(z, X_i|Y_i = 0) >^{st} \rho(z, X_i|Y_i = 1)$ (see, for example, Fig. 3, left panel) $lim\ \tau(z) > T\ a.s.$ and $lim\ I_{\{\tau(z)>F_\tau^{-1}(1/2)\}} = 1\ a.s.$, while under $H_A : \rho(z, X_i|Y_i = 0) <^{st} \rho(z, X_i|Y_i = 1)$ (see, for example, Fig. 3, right panel)

$lim\ \tau(z) < T\ a.s.$ and $lim\ I_{\{\tau(z)>F_\tau^{-1}(1/2)\}} = 0\ a.s.$ We conclude that under either stochastic ordering alternative, the observation z is correctly classified almost surely as the class-conditional training sample sizes grow to infinity. □

See [35] for details regarding the asymptotics of $\tau(z)$.

3.6 Adaptive Parameter Selection

Different observations z yield different class-conditional interpoint distance distributions and, therefore, different choices for the parameters r_0, r_1, k_0, k_1 will be appropriate. Adaptive selection of the parameters, based on the class-conditional interpoint distances $\rho(z, X_i|Y_i = j)$, involves choosing r_0, r_1, k_0, k_1 for the classifier via

$$(r_0^*, k_0^*, r_1^*, k_1^*) = \arg \min_{\substack{1 \le k_0 \le r_0 << n_0 \\ 1 \le k_1 \le r_1 << n_1}} \Psi(\{\rho(z, X_i|Y_i=0)\}_{i=1}^{n_0}, \{\rho(z, X_i|Y_i=1)\}_{i=1}^{n_1}; r_0, k_0, r_1, k_1) \tag{11}$$

for some appropriate criterion function Ψ. Optimal choices against stochastic ordering alternatives can be made in terms of the Pitman Asymptotic Efficacy (PAE) notion of asymptotic power [35]. This asymptotic optimality result hinges on the asymptotic normality of the statistic τ. In practice, the issue of adaptively selecting the parameters given class-conditional interpoint distance samples is more problematic. The adaptive approach proposed here and used in the experiments in Section 4 considers a reasonable surrogate criterion Ψ: minimizing order statistic variance V. Intuitively, the sample interpoint distance order statistic with the smallest variance should provide a reliable cue upon which to base the test for stochastic ordering. (Further justification for this surrogate criterion for minimization is based on the form of the rigorously derived PAE criterion for maximization; the sample interpoint distance order statistic variance is the denominator of the PAE [35].) For $j = 0, 1$, we choose (r_j^*, k_j^*) to satisfy

$$(r_j^*, k_j^*) = \arg \min_{1 \le k_j \le r_j \le r_j^{max} << n_j} \hat{V}(\rho(z, X_i|Y_i = j)_{(k_j:r_j:n_j)}), \tag{12}$$

where $\rho(z, X_i|Y_i = j)_{(k_j:r_j:n_j)}$ denotes the k_jth smallest of a random subset of size r_j taken from the n_j sample class-conditional interpoint distances. This variance is simple to calculate (see, e.g., [5]). (For implementation purposes, we must choose $r_j^{max} << n_j$.)

Note that the character of the interpoint distance distributions (as indicated by the shape of the corresponding probability density estimates) depends on the unidentified observation z under investigation. Thus, different observations z will yield different parameter choices as selected via the adaptive procedure.

To illustrate, consider applying the adaptive parameter selection presented in (12) to the two examples depicted in Fig. 3. Each test observation gives rise to its own unique class-conditional interpoint distance distributions and it is this pair of distributions that determine the parameters.

In Case 1, a TCE-present observation (TClf7712) is held out and the relevant training set contains $n_0 = 352$ TCE-absent exemplars and $n_1 = 759$ TCE-present exemplars. The left panel of Fig. 3 depicts the associated class-conditional interpoint distance probability density estimates with $\rho(z, X_i|Y_i = 0) >^{st} \rho(z, X_i|Y_i = 1)$. Equation (12) yields parameter values $r_0^* = 35, k_0^* = 4, r_1^* = 66, k_1^* = 5$. Using these choices, (9) yields $\hat{\tau} = 0.9825$ and (10) yields $\hat{F}_\tau^{-1}(1/2) = 0.6568$. Thus, (6) yields $I_{\{\hat{\tau}(z)>\hat{F}_\tau^{-1}(1/2)\}} = 1$ and the test observation is correctly classified as TCE-present (class 1).

Similarly, in Case 2, a TCE-absent observation (Kero0203) is held out, and the relevant training set contains $n_0 = 351$ TCE-absent exemplars and $n_1 = 760$ TCE-present exemplars. The right panel of Fig. 3 depicts the associated class-conditional interpoint distance probability density estimates for this case with $\rho(z, X_i|Y_i = 0) <^{st} \rho(z, X_i|Y_i = 1)$. The adaptively selected parameter values are

$$r_0^* = 19, k_0^* = 1, r_1^* = 65, k_1^* = 1.$$

Here, $\hat{\tau} = 0.2896$, $\hat{F}_\tau^{-1}(1/2) = 0.7044$, and $I_{\{\hat{\tau}(z)>\hat{F}_\tau^{-1}(1/2)\}} = 0$; the test observation is correctly classified as TCE-absent (class 0).

4 Results

The classification methodology developed in Section 3 is designed for the functional artificial nose chemical sensor data described in Section 2. We present example results for the aforementioned "needle in the haystack" detection task comparing the performance of our proposed generalized Wilcoxon-Mann-Whitney classifier with that of the conventional k-nearest-neighbors classifier. The L_2 (sum of integrated squared differences) distance (1), and an experimentally determined level of (polynomial smoothing spline) smoothing (see Section 4.4) is used throughout. The comparison criterion of interest is therefore straightforward: classifier performance as measured by the probability of classification error $L_n(g) = P[g(Z|D_n) \neq Y_Z]$.

As noted in Section 3.4, for large sample size problems computational considerations preclude the use of the exact ensemble classifier (6) which sums over all subset pairs. The experimental results presented in Section 4.1, using a reduced database, use the exact classifier. Comparitive results with well–known classifiers are presented for this reduced database problem. The full "needle in the haystack" database is considered in Sections 4.2 and 4.3; here, the approximation method of Section 3.4 is used.

4.1 TCE Detection: Low Concentration

We first consider a simplified problem which, despite its simplicity, is of interest in its own right: the detection of a *low concentration* of trichloroethylene (TCE) in complex backgrounds. The problem is simplified in the sense that the sample sizes are relatively small and consideration of a single concentration should yield a simpler discriminant surface. For the results presented in this section, the classifier given by (6) is employed and τ is computed exactly. For additional simplicity, we consider the case $k_0 = k_1 = 1$ so that the available recurrence [25] provides the exact distribution F_τ.

The database D'_n considered in this example consists of $n = 160$ observations. Class 0, the TCE-absent class, is represented by $n_0 = 80$ observations; ten observations of pure air and ten observations each of seven of the diluting odorants at a concentration of 1:2 in air. There are likewise $n_1 = 80$ class 1 (TCE-present) observations; ten observations of TCE diluted to a 1:10 concentration in air and 10 observations each of TCE diluted to a 1:10 concentration in a 1:2 concentration of seven of the diluting odorants in air. (Due to a missing observation at this lowest concentration, Benzene is omitted from the analysis.) This reduced database D'_n includes 80 of the most difficult to detect TCE-present observations in the overall database D_n.

Let K represent the class of k-nearest-neighbors classifiers with $k \in \{1, 3, ..., 37, 39\}$ [6], [4] and G represent the class of generalized Wilcoxon-Mann-Whitney classifiers (6) parameterized by $k_0 = k_1 = 1$ and $r_0 = r_1 = r \in \{1, 3, ..., 37, 39\}$. Each class contains 20 classifiers. Let (Z, Y) be a random pair representing an unidentified observation and its class label, and define $\theta(g|D_n) = 1 - L_n(g) = P[g(Z|D_n) = Y]$ to be the probability that classifier g, trained on database D_n, correctly classifies the unidentified observation Z.

Cross-validation yields

$$max_{g \in G} \hat{\theta}(g|D'_n) = 120/160 = 0.75$$
$$> max_{g \in K} \hat{\theta}(g|D'_n) = 117/160 = 0.73125,$$

where

$$\hat{\theta}(g|D_n) = (1/n) \sum_{i=1}^{n} I_{\{g(X_i; D_{(i)}) = Y_i | D_n\}} \tag{13}$$

is the leave-one-out, or deleted, estimate of the probability of correct classification [6], [21]. Here, $D_{(i)}$ represents the training database D_n with observation (X_i, Y_i) deleted. Employing the adaptive procedure for parameter selection ((12) with the additional constraint that $k_0 = k_1 = 1$) yields $\hat{\theta}(g|D'_n) = 126/160 = 0.7875$.

For comparison, the results on this data set of several commonly used classifiers are considered. As noted, the best k-nearest-neighbors classifier (with $k = 3$) yields $\hat{\theta}(g|D'_n) = 0.73125$. The adaptive nearest-neighbor procedure developed in [13] yields $\hat{\theta}(g|D'_n) = 0.75$, the tree classifier CART [28], [3] yields $\hat{\theta}(g|D'_n) = 0.725$, and a support vector machine (SVM) implementation [31], [18] yields $\hat{\theta}(g|D'_n) = 0.75$. The performance improvement of the adaptive Wilcoxon-Mann-Whitney classifier over the best of these competitors is statistically significant at the $p = 0.05$ level, by McNemar's test [28].

Further investigation indicates moderately near neighbors are the most useful for classification in this problem. For k-nearest-neighbors, performance is optimized with $k = 3$ and degrades significantly and rapidly as k increases. The classification performance of our proposed classifier (6) on this problem is best for larger values of r (optimal performance for the class of generalized Wilcoxon-Mann-Whitney classifiers G is obtained with $r = 33$) which again correspond to heavier weighting on the near neighbors than on the farther neighbors. (Recall that the extreme cases of the generalized Wilcoxon-Mann-Whitney classifier with $k_0 = k_1 = 1$ are the classical Wilcoxon-Mann-Whitney median test for $r_0 = r_1 = 1$ and the classical 1-nearest-neighbors classifier for $r_j = n_j$.)

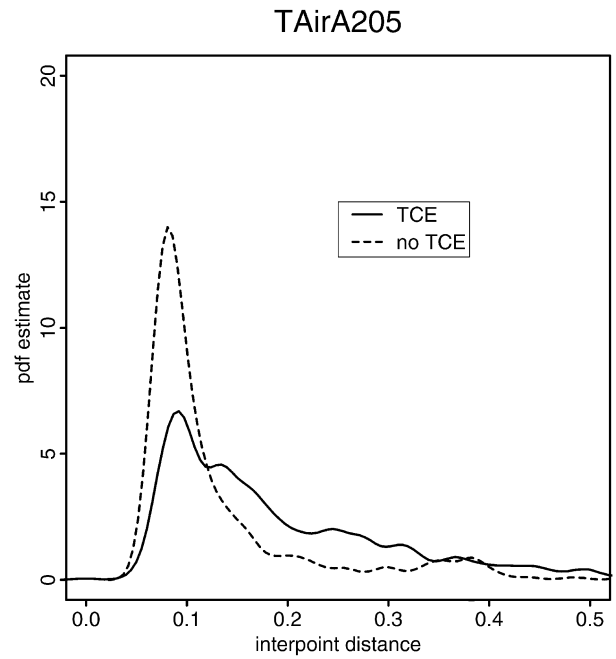


Fig. 4. Interpoint distance distributions for a misclassified TCE-present observation.

Analysis of the parameters selected via the adaptive generalized Wilcoxon-Mann-Whitney procedure indicates that different diluting analytes yield parameter choices with significantly different character. For example, the value of the adaptively selected parameter r_0^* is between 18 and 26 for all ten of the cases in which the test observation is air, while $r_0^* \geq 30$ for the cases in which the test observation is a mixture of Coleman fuel and air. This investigation illustrates the potential of the adaptive procedure and is the subject of continuing investigation as an avenue for impacting sensor design.

4.2 TCE Detection: Entire Database

Returning to the overall database D_n of $n = 1,112$ observations at various concentration levels, we now present results based on (6) using (9) and (10) to estimate the observed value of the statistic τ and the distribution F_τ, respectively. The deleted estimate of classification performance yields $max_{g \in K} \hat{\theta}(g|D_n) = 973/1,112 = 0.875$; the best k-nearest-neighbors classifier makes 139 errors. (Comparing this performance on the entire database with the analogous performance on the low-concentration observations reported in Section 4.1 supports the claim that the low concentration observations are among the most difficult in D_n.)

For the adaptive generalized Wilcoxon-Mann-Whitney classifier $\hat{\theta}(g|D_n) = 1,062/1,112 \approx 0.955$. (The parameters u and v (see Section 3.4) are chosen based on computational consideration; $u = v = 2,500$ so that leave-one-out cross-validation can be performed on this large database in a reasonable amount of time.) Our ensemble approach results in the elimination of more than 64 percent of the 139 k-nearest-neighbors errors. This performance improvement is statistically significant at the 0.05 level.

TABLE 1
Confusion Matrix for Case I

Estimated Class \ True Class	0	1
0	$\frac{327}{352}$	$FN = \frac{25}{760}$
1	$FP = \frac{25}{352}$	$\frac{735}{760}$

Annotation: false positives (*FP*); false negatives (*FN*).

TABLE 2
Confusion Matrix for Case II

Estimated Class \ True Class	0	1
0	$\frac{302}{352}$	$FN = \frac{7}{760}$
"No Decision"	$\frac{25}{352}$	$\frac{18}{760}$
1	$FP = \frac{25}{352}$	$\frac{735}{760}$

Annotation: false positives (*FP*); false negatives (*FN*).

As an illustrative example of an arbitrarily selected misclassified observation, consider Fig. 4. Here, we plot the class-conditional interpoint distance densities for one of the lowest concentration TCE-present observations (TAirA205). Note that $\rho(z, X_i|Y_i = 0) <^{st} \rho(z, X_i|Y_i = 1)$. It is not surprising that the adaptive generalized Wilcoxon-Mann-Whitney classifier, based on interpoint distances, misclassifies this observation. Indeed, it is unclear that any (reasonable) classifier will correctly classify this observation as TCE-present.

4.3 Utility of Classification Significance

A consequence of our approach is the availability of a confidence measure (significance) for the classification decision. This is analogous to the desired "*no decision*" decision which drove the development of the (k, l) nearest-neighbor classifier [15], [6], in which no decision is made unless at least $l > k/2$ observations are from the same class.

For illustrative purposes, consider an operational scenario in which the cost of incorrectly classifying a TCE-present observation as TCE-absent is greater than the cost of a false positive error. Furthermore, assume that a "no decision" classification can be used to indicate the need for the collection of an additional observation. ("No decision" carries with it an implicit cost in terms of time and/or treasure, precluding the overuse of this option.)

A selection of $\alpha = 1/2$ and $\beta = 1/3$ for the classifier (5) is qualitatively appropriate for this scenario. The choice of $\alpha = 1/2$ results in a classification of TCE-present if the evidence supports class 1 over class 0, no matter how weak this support. Setting $\beta = 1/3$ implies that the unidentified observation z will be considered TCE-absent only if the evidence strongly supports this assertion. Weak evidence favoring class 0 over class 1 ($1/3 \leq F_\tau(\tau) \leq 1/2$) will yield a "no decision" classification, reducing false negative errors and setting in motion the machinery of additional investigation.

We investigate the performance on database D_n of classifier (5) for two cases. For Case I, we consider $\alpha = \beta = 1/2$, and, thus, (5) reduces to (6). As presented in Section 4.2, (6) results in 50 classification errors for this case. As depicted in the confusion matrix (Table 1), these 50 errors are accounted for as $25/352 = 0.071$ false positives (classifying TCE-absent as TCE-present) and $25/760 = 0.033$ false negatives. With the settings $\alpha = 1/2$ and $\beta = 1/3$ for Case II, chosen to reduce the false negative errors (at an acknowledged cost of some reinvestigation), the overall performance becomes 32 classification errors and 43 "no decisions." More importantly, for the illustrative application, the false negative error rate is reduced to $7/760 = 0.009$, at the cost of introducing a "no decision" rate of $43/1,112 = 0.039$. For Case II, the 760 TCE-present observations yield 735 correct classifications (the same as for Case I, since $\alpha = 1/2$ in both cases), seven incorrect classifications, and 18 "no decisions" (see Table 2). These remaining seven false negatives are a subset of the 80 lowest concentration TCE-present observations investigated in Section 4.1.

4.4 Choice of Smoother

For our purposes, the sensor responses are smoothed separately for each fiber ϕ and wavelength λ; thus $s : \mathcal{C}(\mathcal{T}) \to \mathcal{C}(\mathcal{T})$. We utilize polynomial smoothing splines with the level of smoothing determined by cross-validation [12], [27]. As a competing approach, consider using kernel smoothing [27], [32] to smooth each sensor response. A bandwidth of $h = 1.1$ for a Gaussian kernel was determined by visual inspection of smoothed curves and classification performance experiments. (Note that the level of smoothing suggested by classification performance experiments is, in terms of visual inspection of smoothed curves, significantly less than anticipated. That is, optimizing classification performance over the smoothing bandwidth h yields response curves more similar to the signals depicted in Fig. 2 than to the cartoon curves of Fig. 1.) Experiments utilizing kernel smoothing indicate no improvement over polynomial smoothing splines in terms of classification performance.

4.5 Choice of Distance Function

Investigation of

$$\rho(X_1, X_2) = \sum_{\phi=1}^{19} \sum_{\lambda=1}^{2} \int_{\mathcal{T}} |s(X_1^{\phi,\lambda})(t) - s(X_2^{\phi,\lambda})(t)| dt \qquad (14)$$

suggests that L_1 provides no significant improvement over L_2. In fact, L_1 yields an empirical performance degradation.

Given an experimentally determined smoother s and nonnegative weights $w_{\phi,\lambda}$, the weighted L_p pseudodistance

$$\rho(X_1, X_2) = (\sum_{\phi=1}^{19} \sum_{\lambda=1}^{2} w_{\phi,\lambda}^p \int_{\mathcal{T}} |s(X_1^{\phi,\lambda})(t) - s(X_2^{\phi,\lambda})(t)|^p dt)^{1/p} \tag{15}$$

for $1 \leq p < \infty$ permits exploratory analysis of the various sensor (fiber, wavelength) bands. For instance, with $p = 2$ comparing the experimental performance presented above (which uses (15) with $w_{\phi,\lambda} = 1$ for all ϕ, λ) against the performance obtained using weights which eliminate from consideration one of the wavelengths ($w_{\phi,\lambda} = 0$ for one λ and all ϕ while $w_{\phi,\lambda} = 1$ for the other λ and all ϕ) indicates that the signals for both wavelengths λ contain valuable information for the classification problem at hand. Similarly, investigations using the 19 different distances given by eliminating from consideration exactly one fiber (setting $w_{\phi,\lambda} = 0$ for one fiber ϕ) and the 19 distances given by considering exactly one fiber ($w_{\phi,\lambda} = 0$ for all but one ϕ) indicate that each fiber is important and no fiber suffices. Finally, experimental evidence suggests that choosing p much greater than 2 (say, 30) yields significantly improved classification performance; see [26].

The investigation of distances of the form (15) is highly relevant to continuing efforts in classifier design. Furthermore, these analyses represent a primary way in which our classification results can impact sensor design.

5 Discussion

A careful analysis of a well-designed data set yields evidence that the interpoint distance-based adaptive generalized Wilcoxon-Mann-Whitney ensemble classifier is appropriate for difficult olfactory classification tasks. In addition, the availability of classification significance levels increases the operational utility of the approach for many applications.

A discussion of the implications of our results to chemical sensor data analysis must begin with the fact that applications require a high level of classification performance at much lower concentrations than those considered here. Both the sensor design and the methods for olfactory signal processing and pattern recognition must continue to improve.

Finally, while the needle in the haystack detection problem is of significant intrinsic interest, the olfactory classification problem in general and the Tufts database D_n in particular call for a multiclass investigation.

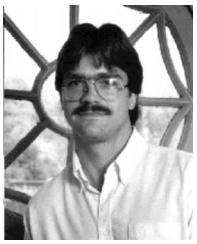
Acknowledgments

Funding for this effort was provided in part by Defense Advanced Research Projects Agency as administered by the Air Force Office of Scientific Research under contract DOD F49620-99-1-0213 and by the US Office of Naval Research grant N00014-95-1-0777. The database used in this article was provided by David Walt's laboratory at Tufts University. Additional domain expertise was provided by Peter Jurs' laboratory at Penn State University and by John Kauer's laboratory at Tufts University. The author would like to thank Jeff Solka, Dave Marchette, Adam Cannon, Jingdong Xie, Lenore Cowen, and the three anonymous referees for their helpful comments.

References

[1] P.J. Bickel and K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics.* Oakland, Calif.: Holden-Day, 1977.
[2] L. Breiman, "Bagging Predictors," *Machine Learning,* vol. 24, no. 2, pp. 123-140, 1996.
[3] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees.* London: Chapman & Hall, 1984
[4] T.M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Information Theory,* vol. 13, pp. 21-27, 1967.
[5] H.A. David, *Order Statistics,* New York: Wiley, 1970.
[6] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition,* New York: Springer, 1996.
[7] T.A. Dickinson, J. White, J.S. Kauer, D.R. Walt, "A Chemical-Detecting System Based on a Cross-Reactive Optical Sensor Array," *Nature,* vol. 382, pp. 697-700, 1996.
[8] T.G. Dietterich, "Machine Learning Research: Four Current Directions," *AI Magazine,* vol. 18, no. 4, pp. 97-136, 1997.
[9] J.H. Friedman, "Another Approach to Polychotomous Classification," technical report, Stanford Univ., 1996. (unpublished).
[10] R. Guitierrez-Osuna, "Olfactory Signal Processing and Pattern Recognition," *IEEE Spectrum,* vol. 35, no. 9, p. 28, Sept. 1998.
[11] L. Hansen and P. Salamon, "Neural Network Ensembles," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, pp. 993-1001, 1990.
[12] T.J. Hastie and R.J. Tibshirani, *Generalized Additive Models,* London: Chapman and Hall, 1990.
[13] T. Hastie and R. Tibshirani, "Discriminant Adaptive Nearest Neighbor Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, pp. 607-616, 1996.
[14] T. Hastie and R. Tibshirani, "Classification by Pairwise Coupling," *Ann. Statistics,* vol. 26, no. 2, pp. 451-471, 1998.
[15] M. Hellman, "The Nearest Neighbor Classification Rule with a Reject Option," *IEEE Trans. Systems Science and Cybernetics,* vol. 6, pp. 179-185, 1970.
[16] T.K. Ho, J.J. Hull, and S.N. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 1, pp. 66-75, Jan. 1994.
[17] *IEEE Spectrum,* special issue on Electronic Noses, vol. 35, no. 9, pp. 22-38 Sept. 1998.
[18] T. Joachims, "Making Large-Scale SVM Learning Practical," *Advances in Kernel Methods—Support Vector Learning,* B. Scholkopf, C.J.C. Burges, and A.J. Smola, eds., pp. 169-184, MIT Press, 1999.
[19] G. Kaplan and R. Braham, "Special Report: A Nose is a Nose is a Nose?" *IEEE Spectrum,* vol. 35, no. 9, p. 22, Sept. 1998.
[20] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 3, pp. 226-239, Mar. 1998.
[21] S.R. Kulkarni, G. Lugosi, and S. Venkatesh, "Learning Pattern Classification—A Survey," *IEEE Trans. Information Theory,* vol. 44, no. 6, pp. 2178-2206, 1998.
[22] J.-F. Maa, D.K. Pearl, and R. Bartoszynski, "Reducing Multidimensional Two-Sample Data to One-Dimensional Interpoint Comparisons," *Ann. Statistics,* vol. 24, pp. 1069-1074, 1996.
[23] H.B. Mann and D.R. Whitney, "On a Test Whether One of Two Random Variables is Stochastically Larger than the Other," *Ann. Math. Statistics,* vol. 18, pp. 50-60, 1947.
[24] H.T. Nagle, S.S. Schiffman, and R. Guitierrez-Osuna, "The How and Why of Electronic Noses," *IEEE Spectrum,* vol. 35, no. 9, pp. 22-34, Sept. 1998.
[25] C.E. Priebe and L.J. Cowen, "A Generalized Wilcoxon Mann-Whitney Statistic," *Comm. Statistics: Theory and Methods,* vol. 28, no. 12, pp. 2871-2878, 1999.
[26] C.E. Priebe, D.J. Marchette, and J.L. Solka, "On the Selection of Distance for a High-Dimensional Classification Problem," *Proc. Statistical Computing Section of the Am. Statistical Assoc.,* 2000.
[27] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis.* New York: Springer, 1997.
[28] B.D. Ripley, *Pattern Recognition and Neural Networks,* Cambridge Mass.: Cambridge Univ. Press, 1996.
[29] *Science,* special issue on Olfaction, vol. 286, no. 5540, pp. 703-728, Oct. 1999.
[30] D.B. Skalak, "Prototype Selection for Composite Nearest Neighbor Classifiers," PhD Thesis, Dept. of Computer Science, Univ. of Massachusetts, Amherst, 1997.
[31] V.N. Vapnik, *The Nature of Statistical Learning Theory.* New York: Springer, 1995.

[32] M.P. Wand and M.C. Jones, *Kernel Smoothing*. London: Chapman & Hall, 1995.
[33] J. White, J.S. Kauer, T.A. Dickinson, and D.R. Walt, "Rapid Analyte Recognition in a Device Based on Optical Sensors and the Olfactory System," *Analytical Chemisty*, pp. 2191-2202, vol. 68, 1996.
[34] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics*, vol. 1, pp. 80-83, 1945.
[35] J. Xie and C.E. Priebe, "Generalizing the Mann-Whitney-Wilcoxon Statistic," *J. Nonparametric Statistics*, vol. 12, pp. 661-682, 2000.



Carey E. Priebe received the BS degree in mathematics from Purdue University in 1984, the MS degree in computer science from San Diego State University in 1988, and the PhD degree in information technology (computational statistics) from George Mason University in 1993. From 1985 to 1994, he worked as a mathematician and scientist in the US Navy research and development laboratory system. Since 1994, he has been a professor in the Department of Mathematical Sciences, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland. His research interests are in computational statistics, kernel and mixture estimates, statistical pattern recognition, and statistical image analysis.