PERGAMON

Pattern Recognition 36 (2003) 45–60

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Characterizing the scale dimension of a high-dimensional classification problem ☆

David J. Marchette[a,*], Carey E. Priebe[b]

[a] *Code B10, Naval Surface Warfare Center, Dahlgren, VA 22448–5100, USA*
[b] *Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218–2682, USA*

Received 18 January 2001; received in revised form 16 October 2001; accepted 17 December 2001

Abstract

Classification of high-dimensional data is inherently difficult. We present an exploratory data analysis methodology for characterizing the scale dimension of a classification problem. The idea is to characterize the support of one distinguished target class as a collection of balls covering the class, with each ball centered at an observation in that class such that the radius is maximal without containing observations from the other classes. The scale dimension is defined to be the number of distinct radii (ball sizes) required to cover the class without covering observations from the other class. A greedy algorithm is used to fit the balls. The balls then provide a description of the support of the target class, with information about the complexity of the classification problem implicit in the number, radii, adjacency and position of the balls. Clustering the balls by radius and pruning the cluster tree yields an estimate of the scale dimension for the problem. We illustrate the methodology with pedagogical simulations and a chemical sensor data analysis application. Published by Elsevier Science Ltd on behalf of Pattern Recognition Society.

Keywords: Scale dimension; Classification; Exploratory data analysis; Interpoint distance; Reduced kernel estimator; Random graph; Class cover; High-dimensional data; Artificial nose

1. Introduction

This article presents an exploratory data analysis methodology for the investigation and characterization of the *scale dimension* of high-dimensional classification problems. Scale dimension is defined in terms of the number of equal radius balls needed to characterize the problem. The balls are centered on exemplars from one class, referred to as the target class, such that they cover no observations of the other class. The ball centers are then the only target class observations retained by the classifier. Classification is performed by projecting observations based on the distance to the ball centers. This results in a form of reduced kernel estimator classifier. Section 2 develops the methodology, Section 3 presents pedagogical simulation examples designed to give an intuitive understanding of the algorithm, and Section 4 presents a case study for a chemical sensor data analysis application.

1.1. Statistical pattern recognition

We consider statistical pattern recognition in the supervised case (classification as opposed to clustering). An available training database D_n consists of (X, Y) pairs; $D_n = [(X_1, Y_1), \ldots, (X_n, Y_n)]$, where the multi-variate or function-valued (Ξ-valued, say) random variables X_i represent training observations (data collected for exemplar i) and their associated class labels Y_i take their values

☆ Support for this effort is 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, Office of Naval Research grant N00014-95-1-0777, Office of Naval Research via the NSWC ILIR program, and the Acheson J. Duncan Fund for the Advancement of Research in Statistics.

* Corresponding author. Tel.: +1-540-653-2736; fax: +1-540-653-2641.

E-mail address: marchettedj@nswc.navy.mil (D.J. Marchette).

0031-3203/02/$22.00 Published by Elsevier Science Ltd on behalf of Pattern Recognition Society.
PII: S0031-3203(02)00042-0

in $\{0,1\}$. We shall consider the two-class problem throughout, although the methodology can be extended to the multi-class case. The class-conditional samples are represented by $\mathscr{X}_j = \{X_i : Y_i = j\}$ for $j = 0, 1$, with $\mathscr{X}_0 \cup \mathscr{X}_1 = \mathscr{X}$ and $\mathscr{X}_0 \cap \mathscr{X}_1 = \emptyset$. The cardinality of the set $\mathscr{X}_j$ is $|\mathscr{X}_j| = n_j$, and therefore the database D_n consists of n_0 observations from class 0 and n_1 observations from class 1, for a total of $n = n_0 + n_1$ observations. (The class-conditional training sample sizes n_j may be taken to be *design variables* rather than random variables.) For $X_i \in \mathscr{X}_j$, the X_i are assumed to be independent and identically distributed F_j. For simplicity, we will further assume that the random variables X_i are continuous; that is, the probability density functions f_j exist. Hence, ties occur with probability zero, and the observations are distinct almost surely. This will be implicitly assumed throughout.

A common approach to the distribution-free analysis of two or more high-dimensional samples involves the consideration of the interpoint distances [1]. The choice of distance is of fundamental importance; in general, we let $\rho(\cdot,\cdot) : \Xi \times \Xi \to \mathbb{R}_+ = [0,\infty)$ be an appropriate distance function. Then, for $X_i \in \mathscr{X}$ and $W \subset \mathscr{X}$, we define

$$\rho(X_i, W) = \min_{X' \in W} \rho(X_i, X').$$

Thus, $X_i \in W \Rightarrow \rho(X_i, W) = 0$ while $X_i \notin W \Rightarrow \rho(X_i, W) > 0$ a.s.

1.2. Motivation

When the dimensionality of the original data (the dimensionality of Ξ) is large, it is difficult at best to perform statistical pattern recognition in the domain space—the "curse of dimensionality" suggests that the sample size n is likely too small. See for example Refs. [2] or [3]. We are interested in understanding the geometry of the classification problem, and to this end we would like to be able to determine the scale of the data in a given region of its support. Thus, we are looking for witness sets W_l and their associated balls. In particular, we want to know the ball centers and radii that cover the support of the $\mathscr{X}$ observations.

How shall we choose the witness sets W_l; that is, how shall we choose the number d, the size of the witness sets, and the witness elements therein? Designate a distinguished *target class* $J \in \{0,1\}$. Then the nontarget class is $1 - J$. (The proposed procedure is nonsymmetric in the two classes, meaning that the results depend upon the choice of J.) Consideration of multiple witness sets allows consideration of multiple scales. The lth scale applies to the local region near the witness elements of the lth witness set. We will show how this gives a way of exploring the geometry of the target class relative to class $1 - J$.

One reason for performing this mapping is to allow a better understanding of the data through visualization or other low-dimensional exploratory data analysis. A second reason, mentioned above, is the inherent difficulty of performing analysis and estimation in high-dimensional spaces; classification performance may be superior under a reduced complexity model. Thus, the selection of an appropriate cover is of great interest to the pattern recognition practitioner.

2. Methodology

2.1. Class covering

Given an integer $d \geqslant 1$, let $\mathscr{W} = [W_1 \cdots W_d]'$ be a vector of witness sets $W_l \subset \mathscr{X}_J$ for $l = 1, \ldots, d$, and let $\mathscr{R} = [r_1 \cdots r_d]$ be a vector of nonnegative scalars (representing radii); $r_l \in (0,\infty)$ for $l = 1, \ldots, d$. Let $B(x,r) = \{x' : \rho(x,x') < r\}$ be the (open) ball in Ξ of radius r centered at x. If

$$\mathscr{X}_J \subset \bigcup_{l=1}^{d} \bigcup_{X \in W_l} B(X, r_l),$$

then $(\mathscr{W}, \mathscr{R})$ is said to be a *proper cover* of class J. (For specificity, we will use the adjective *proper* to denote covers which contain all of the target class observations in order to distinguish them from generalizations defined in the sequel.) If, furthermore,

$$\mathscr{X}_{1-J} \cap \left(\bigcup_{l=1}^{d} \bigcup_{X \in W_l} B(X, r_l) \right) = \emptyset,$$

then $(\mathscr{W}, \mathscr{R})$ is said to be a *pure* cover of class J. We say that d is the *dimensionality* of the cover. If $\mathscr{R} = [r \cdots r]$—all radii are the same—we say that the cover $(\mathscr{W}, \mathscr{R})$ is *homogeneous*. Similarly, we say the cover is *heterogeneous* if the radii are not equal.

Pure covers exist with dimensionality $d \leqslant n_J$ and witness set cardinality $|W_l| \leqslant n_J$.

Example 1. There exists (a.s.) $\varepsilon > 0$ such that $(\mathscr{W} = [\mathscr{X}_J],\ \mathscr{R} = [\varepsilon])$ is a pure (and trivially homogeneous) cover with dimensionality $d = 1$ and $|W_1| = n_J$.

Example 2. For $d = n_J$ there exists (a.s.) some small $\varepsilon > 0$ such that letting $\mathscr{W}$ be the n_J-vector with elements $W_i = \{X_i\}$ for each $X_i \in \mathscr{X}_J$ and setting $\mathscr{R} = [\varepsilon \cdots \varepsilon]$ yields a pure homogeneous cover. Thus, $d = n_J$ and $|W_i| = 1$.

Example 3. Let $\mathscr{W}$ be the n_J-vector with elements $\{X_i\}$ for each $X_i \in \mathscr{X}_J$ and setting $\mathscr{R} = [d(X_1, \mathscr{X}_{1-J}) \cdots d(X_n, \mathscr{X}_{1-J})]$ yields a pure heterogeneous cover. This cover will be important in the sequel. Once again, $d = n_J$ and $|W_i| = 1$ for this cover.

The above examples represent the phenomenon of overfitting, which is in general to be avoided.

We call a cover $(\mathscr{W}, \mathscr{R})$ for which $|W_l| = 1$ for all l a *unit* cover. Thus (from Example 2), there always (a.s.) exists a pure proper homogeneous unit cover with $d \leqslant n_J$. A construction for approximating a *minimal*—that is, one for which d is as small as possible—pure proper homogeneous unit cover is given in Ref. [4], and computational complexity results for this construction are provided.

The cover defined in Example 3 defines the *class cover catch digraph* as follows. Following Ref. [5], we define a *graph* to be a set of vertices V and unordered pairs of vertices called edges. The set of edges is denoted E. We write the edge u, v, between vertices u and v, as uv. The vertices u and v are called *neighbors*. If the pairs are ordered, we call the graph a *directed* graph, or *digraph*. If uv is a directed edge, then v is a neighbor of u. If there is no edge vu, then u is not a neighbor of v. A sphere digraph [6] is constructed from a set of balls as follows. Each ball corresponds to a vertex of the digraph, and there is a directed edge from one vertex to another if the center of the ball corresponding to the second vertex is contained in the interior of the ball corresponding to the first. The class cover catch digraph (CCCD) is the sphere digraph corresponding to the balls in Example 3. That is, it is the digraph on n_J vertices, corresponding to the n_J balls centered on the $X_i \in \mathscr{X}_J$, with a (directed) edge between u and v if and only if the ball corresponding to u covers the center of the ball corresponding to v. The operations that we will be performing on the covers can be understood in terms of operations on the underlying CCCD. Priebe et al. [7] discuss a particular one-dimensional case of a CCCD. The CCCD is thus the sphere digraph with spheres of maximal radius (maximal in the sense of not covering observations from class $\mathscr{X}_{1-J}$).

The covers described above are typically redundant. We would like to find a cover that uses fewer balls, both because it is a more compact representation, and because it allows faster computations. To this end we utilize the concept of a dominating set.

A set of vertices D is called a *dominating set* of G if every vertex of G not in D is a neighbor of a vertex of D. A dominating set of minimum cardinality is called a *minimum dominating set*. The cardinality of a minimum dominating set of G is denoted $\gamma(G)$, or just γ if the underlying graph is understood. A dominating set that contains no (strict) subsets which are themselves dominating sets is called a *minimal dominating set*. Note that a minimum dominating set is minimal, but the converse is not necessarily true.

The CCCD is a type of *random graph*. The edges are defined in terms of a random sample drawn from the two classes. This defines a class of random graphs as those which can be realized as CCCDs for a particular classification problem.

Given a CCCD, there are a number of ways one can use the cover to construct a classifier. We define one such as follows. Let $\mathscr{W} \subset \mathscr{X}_J$ be a set of points with associated balls $\{B(s,r) | s \in \mathscr{W}\}$. We classify an observation x as class J if $x \in \bigcup_{s\in\mathscr{W}} B(s,r)$. Otherwise, we assign the observation class $1-J$. This can easily be extended by scaling by the radius of the balls, or by constructing the CCCD on $\mathscr{X}_{1-J}$ and using a tie-breaking rule when the observation is contained in balls from each class. We will only be concerned with the simplest of these, as our purpose is to explore the concept of scale dimension rather than the construction of the classifier. Thus, we define the CCCD classifier $g_{\mathscr{W}}$ as

$$g_{\mathscr{W}}(x) = 1 - J + (2J-1) I \left\{ x \in \bigcup_{s\in\mathscr{W}} B(s,r) \right\}.$$

It is easy to see that $g_{\mathscr{W}}(x) = J \Leftrightarrow x$ is interior to one of the balls defined by $\mathscr{W}$, otherwise $g_{\mathscr{W}}(x) = 1 - J$.

The class-conditional resubstitution error rate estimate for class J for the CCCD classifier $g_{\mathscr{W}}$ is given by

$$\hat{L}_J^{(R)}(g_{\mathscr{W}}) = \frac{1}{n_J + n_{1-J}} \left(\sum_{x\in\mathscr{X}_J} I \left\{ x \notin \bigcup_{s\in\mathscr{W}} B(s,r) \right\} + \sum_{y\in\mathscr{X}_{1-J}} I \left\{ y \in \bigcup_{s\in\mathscr{W}} B(s,r) \right\} \right).$$

Note that if $(\mathscr{W}, \mathscr{R})$ is a pure proper cover of class J, then $\hat{L}_J^{(R)}(g_{\mathscr{W}}) = 0$.

While a pure proper cover $(\mathscr{W}, \mathscr{R}) \Rightarrow \hat{L}_J^{(R)}(g_{\mathscr{W}}) = 0$, the converse is not true. We will argue, in the sequel, in favor of the utility of *heterogeneous nonunit nonpure* covers.

2.2. Algorithm

The goal of the algorithm is to find a pure proper unit cover of minimal dimensionality. This will provide us with a compact representation of the classification region, and reduce the computational burden of the classifier. The optimization problem under consideration may be stated as

find $(\mathscr{W}, \mathscr{R})$
so as to
minimize the number of witness sets d
subject to the constraint that $(\mathscr{W}, \mathscr{R})$
is a pure proper unit cover of target class J.

Note that the cover is not constrained to be homogeneous. Let us define d_{min} to be the minimizing number of witness sets for the optimization problem stated above. Conditional on the training set D_n, d_{min} is well defined even though there is not necessarily a unique minimizing cover. In the graph theory terminology defined above, we are concerned with finding a minimum dominating set for the CCCD. We have $1 \leqslant d_{min} \leqslant d_{homogeneous} \leqslant n_J$, where $d_{homogeneous}$ is the minimizing number of witness sets for the analogous optimization problem in which the cover is constrained to be homogeneous [4].

Consider the random $n_J \times n_J$ binary matrix $A = [a_{i,j}]$ with elements given by $a_{i,j} = I\{\rho(X_i, X_j) < \rho(X_i, \mathscr{X}_{1-J})\}$ for $X_i, X_j \in \mathscr{X}_J$ and $i, j = 1, \ldots, n_J$. That is, $a_{i,j} = 1$ if X_j is an element of the pure open ball of maximum possible radius centered at X_i. (Here “pure” means “contains no class

$1-J$ observations".) A is the augmented adjacency matrix corresponding to the (random) CCCD. Then the goal is to find a solution to the optimization problem

Minimize $\vec{1}^{\mathrm{T}}\vec{x}$
subject to $A^{\mathrm{T}}\vec{x} > \vec{0}$ (or, equivalently, $A^{\mathrm{T}}\vec{x} \geqslant \vec{1}$)
N.B. $\vec{x} \in \{0,1\}^{n_J}$, and the inequalities are component-wise .

The elements x_i of the binary vector $\vec{x}$ of length n_J indicate whether there is a ball centered at observation $X_i \in \mathscr{X}_J$; i.e., X_i is a witness element. The unit cover sought is given by the observations indicated by the nonzero elements of a vector $\vec{x}$ which solves the optimization problem, and their associated radii. This can be seen to be an NP-hard optimization problem by transformation to the classical "dominating set" problem of graph theory (see Refs. [4,8]).

Our deterministic "greedy heuristic" algorithm [9] proceeds as follows:

```
**Set** $\mathscr{J} = \mathscr{X}_J,\ d = 0,\ l = 0$
**While** $\mathscr{J} \neq \emptyset$
    $l = l + 1$
    $X = argmax_{X' \in \mathscr{X}_J} |B(X', \rho(X', \mathscr{X}_{1-J})) \cap \mathscr{J}|$ [break ties arbitrarily]
    $W_l = \{X\}$
    $r_l = \rho(X, \mathscr{X}_{1-J})$
    $\mathscr{J} = \mathscr{J} \setminus \{X' : \rho(X, X') < r_l\}$
**EndWhile**
**Set** $d = l$
**Return** $(\mathscr{W} = [W_1 \cdots W_d]', \mathscr{R} = [r_1 \cdots r_d]')$.
```

The algorithm presented herein is fast, and is guaranteed to find a pure proper (heterogeneous a.s.) unit cover of the specified target class J. The algorithm is *not* guaranteed to find a minimal cover; we define $\hat{d}_{min}$ to be the cover dimensionality obtained by the algorithm. Given class-conditional distributions and training sample sizes n_j, d_{min} and $\hat{d}_{min}$ are random variables. Conditional on the training sample D_n, d_{min} and $\hat{d}_{min}$ are scalars and $\hat{d}_{min} \geqslant d_{min}$. An interesting open question concerns the magnitude of $\hat{d}_{min} - d_{min}$.

2.3. Cover generalizations

The drawback, in general, of requiring pure proper covers is clear: overfitting. We define the *cardinality* c of the cover to be the total number of witness elements; $c = \sum_{l=1}^{d} |W_l|$. Requiring purity may result in a cover with an unacceptably large cardinality. For example, a small number of nontarget class observations amongst the mass of target observations will force c to be large, in which case allowing a small amount of impurity might greatly reduce c. Similarly, the requirement for a proper cover may result in a cover with an unacceptably large dimension (and cardinality) due to a small number of outlying target class observations, each of which requires its own witness element. We now consider generalizations of the pure proper covers presented above.

For $0 \leqslant s \leqslant n_J$, define $(\mathscr{W}, \mathscr{R})$ to be an *s-missing* cover of class J if

$$\left| \mathscr{X}_J \cap \left(\bigcup_{l=1}^{d} \bigcup_{X \in W_l} B(X, r_l) \right) \right| \geqslant n_J - s.$$

The proper covers defined earlier are 0-missing covers. Note that for $s \geqslant 1$, s-missing covers are no longer necessarily proper but may or may not be pure. For $0 \leqslant t \leqslant n_{1-J}$, define $(\mathscr{W}, \mathscr{R})$ to be a *t-tainted* cover of class J if

$$\left| \mathscr{X}_{1-J} \cap \left(\bigcup_{l=1}^{d} \bigcup_{X \in W_l} B(X, r_l) \right) \right| \leqslant t.$$

Note that 0-tainted covers are pure. For $t \geqslant 1$, t-tainted covers are no longer necessarily pure but may or may not be proper.

2.4. Cover simplification

The combinatorics of finding directly an s-missing t-tainted cover for specified s and t is daunting. Therefore, we propose beginning with a pure proper unit cover $(\mathscr{W}, \mathscr{R})$ generated by an approximation algorithm such as the one presented above and "simplifying" the cover ex post facto.

We have the notion of *collapsing a cover*. This operation decreases the number of witness sets while preserving purity and properness. The idea is to take an original cover $(\mathscr{W}, \mathscr{R})$ and cluster those witness sets with identical radii into a single, larger witness set. Let $\mathscr{R}' = [r'_1 \cdots r'_{d'}]'$ be a vector whose elements are the $d' \leqslant d$ distinct radius values in the cover $(\mathscr{W}, \mathscr{R})$. For $l \in \{1, \ldots, d'\}$, let $W'_l = \bigcup_{l' : r_{l'} = r'_l} W_{l'}$ and $\mathscr{W}' = [W'_1 \cdots W'_{d'}]'$. Then $(\mathscr{W}', \mathscr{R}')$ is a cover of dimensionality d'. The result of collapsing a pure proper heterogeneous unit cover $(\mathscr{W}, \mathscr{R})$ of dimension d is a pure proper heterogeneous (not necessarily unit) cover $(\mathscr{W}', \mathscr{R}')$ of dimension $d' \leqslant d$ in which the radii r'_l are distinct. (These distinct radii are exact "fundamental scales" for the cover—a collection of ball sizes which can yield a pure proper cover.) Note that any homogeneous cover can be collapsed to a cover of dimensionality $d' = 1$. For example, the cover in Example 1 can be obtained by collapsing the cover in Example 2.

This idea of collapsing the cover carries over into the graph domain as well. Let $\mathscr{S} = \{S_1, \ldots, S_k\}$ be a collection of disjoint sets of vertices of a digraph G. We call the *condensation* of G the graph on the vertices $\{S_1, \ldots, S_k\}$ such that there is a (directed) edge from S_i to S_j if and only if there is an edge from an element of S_i to an element of S_j in G. (This definition of condensation is slightly broader than that given in Ref. [10], where the sets are required to be the strong components.) We denote the condensed graph $G_{\mathscr{S}}$.

While no two radii in a cover produced by the foregoing greedy algorithm will be identical (almost surely), we consider the generalization to *radii ranges*, in which the elements of $\mathscr{R}$ in a cover $(\mathscr{W}, \mathscr{R})$ are intervals rather than scalars. That is, $r_l \mapsto (r_l^{min}, r_l^{max})$, where

$r_l^{min} = \max_{X \in \bigcup_{X' \in W_l} B(X', r_l) \cap \mathscr{X}_J} \rho_{W_l}(X)$ is the minimum radius such that the most remote target point that needs to be covered remains covered and $r_l^{max} = \rho_{W_l}(\mathscr{X}_{1-J})$ is the maximum allowable radius. If we consider radii ranges in the operation of collapsing a cover then perhaps dimensionality can be reduced further, as the requirement for radii equality can be replaced with a notion of overlapping ranges. Thus, $\mathscr{R}'$ is a vector of values such that, for each radius range (r_l^{min}, r_l^{max}) in the original cover $(\mathscr{W}, \mathscr{R})$, there exists at least one element $r_l' \in (r_l^{min}, r_l^{max})$. Thus, any cover for which there exists an $r \in (r_l^{min}, r_l^{max})$ for all l can be collapsed to a cover of dimensionality $d' = 1$.

For a pure proper heterogeneous unit cover $(\mathscr{W}, \mathscr{R})$ of dimensionality $\hat{d}_{min}$ obtained via the greedy algorithm, let $(\mathscr{W}', \mathscr{R}')$ denote the cover obtained by the operation of collapsing $(\mathscr{W}, \mathscr{R})$ (using radii ranges) so that the dimensionality $\hat{d}'_{min}$ of $(\mathscr{W}', \mathscr{R}')$ is minimal. Then $\hat{d}'_{min} \leqslant \hat{d}_{min}$. The cardinality $c = \hat{d}_{min}$ is unchanged by the operation of collapsing the cover $(\mathscr{W}, \mathscr{R})$. Furthermore, $\hat{L}_J^{(R)}(g_{\mathscr{W}'}) = \hat{L}_{1-J}^{(R)}(g_{\mathscr{W}'}) = 0$.

As noted above, Cannon and Cowen [4] give an algorithm for finding a homogeneous unit cover which is approximately minimal among all homogeneous unit covers. This cover can be collapsed to a cover of dimensionality $d' = 1$. The cost of a focus on homogeneity is a potentially large number of witness elements ($d_{min} \leqslant d_{homogeneous}$). Thus, minimal collapsed dimensionality ($d' = 1$) is traded for high cardinality c. Heterogeneity may yield a higher collapsed dimensionality ($d' \geqslant 1$). The gain may be a smaller total number of witness elements. This issue will be considered further in the examples (Sections 3 and 4).

Collapsing a pure proper cover yields a pure proper cover. We now consider the notions of *clustering a cover* and *pruning a cover*.

Given a cover $(\mathscr{W}, \mathscr{R})$, pruning the cover (removing elements from the cover) will decrease cardinality (total number of witness elements). Pruning can also reduce the dimensionality of the cover. Pruning a pure proper unit cover yields a pure s-missing unit cover. Pruning a cover (at $p \geqslant 1$) simply eliminates those witness sets which account for fewer than p of the target class observations. That is, the pruned cover $(\mathscr{W}_p', \mathscr{R}_p')$ has as its vector of witness sets those witness sets W_l in the original cover for which $|\mathscr{X}_J \cap \bigcup_{X \in W_l} B(X, r_l)| \geqslant p$. (Pruning at $p = 1$ is no pruning at all.) For example, pruning singleton witness sets which cover only their own witness element to get an s-missing cover will alleviate overfitting. So, given a pure proper heterogeneous unit cover $(\mathscr{W}, \mathscr{R})$ of dimensionality $\hat{d}_{min}$ and cardinality $c = \hat{d}_{min}$ obtained as the algorithmic output, pruning at level p yields a pure heterogeneous unit cover $(\mathscr{W}'_p, \mathscr{R}'_p)$ of cardinality $c_p \leqslant \hat{d}_{min}$ and dimensionality $d_p \leqslant \hat{d}_{min}$.

In the CCCD, pruning corresponds to removing vertices (and the edges to and from them). This results in an *induced subgraph* of G, which we denote $prune_p(G)$.

Given a cover $(\mathscr{W}, \mathscr{R})$ of dimensionality d', clustering the cover decreases range-space dimensionality without effecting the cardinality c. The resulting cover may be t-tainted and/or s-missing. The clustering operation proceeds as follows. First the cover radii are clustered, yielding a dendogram. For instance, the examples in Sections 2 and 3 below consider complete linkage clustering [11]. Given the dendogram output of this clustering, there is (a.s.) a canonical clustering into a cover of dimensionality d for each $1 \leqslant d \leqslant d'$. For a given target dimensionality d this canonical clustering is given by cutting the dendogram at a level for which there are d branches and defining for each branch the witness set W_l' for the clustered cover $(\mathscr{W}'_d, \mathscr{R}'_d)$ to be the union of the leaf witness sets W_l under that branch. So, given a pure heterogeneous unit cover $(\mathscr{W}'_p, \mathscr{R}'_p)$ of cardinality $c_p \leqslant \hat{d}_{min}$ and dimensionality $d_p \leqslant \hat{d}_{min}$ obtained by pruning at level p the pure proper heterogeneous unit algorithmic output $(\mathscr{W}, \mathscr{R})$, we can for any $d \leqslant d_p$ produce a cover $(\mathscr{W}'_{p,d}, \mathscr{R}'_{p,d})$ of cardinality c_p and dimensionality d.

The radii within each cluster are all set to be equal. Several choices for the radius of the cluster are available, such as minimum, average, median, etc. We choose to use the minimum of the radii within the cluster. Thus, the balls are all reduced to be the size of the smallest ball contained in the cluster. This guarantees that no new $\mathscr{X}_{1-J}$ observations are covered, at the possible expense of the properness of the cover.

At the graph level, clustering produces a collection $\mathscr{S} = \{S_1, \ldots, S_k\}$ of clusters, which in turn produces a collapsed graph $G_{\mathscr{S}}$. Pruning this results in a final graph $prune_c(G_{\mathscr{S}})$, where the c denotes the pruning criterion used.

Pruning decreases the cardinality, while both pruning and clustering decrease the dimensionality. As a rule, the (resubstitution) classification error will increase under both pruning and clustering. For the special case of collapsing the cover, where equal radius balls are combined, the resubstitution error is not effected.

Our proposed methodology produces a cluster tree for the pure proper heterogeneous unit cover produced by the greedy algorithm. Nonunit covers, as well as missing and tainted covers, are generated via pruning, collapsing, and clustering.

2.5. Scale dimension

Given a cover $(\mathscr{W}, \mathscr{R})$, two important parameters characterizing the complexity of the cover, and hence of the classification problem, are the dimensionality d and the cardinality c. (Recall that the results are asymmetric in target class.)

We begin with a pure proper unit heterogeneous cover produced by the greedy algorithm with dimensionality $\hat{d}_{min} \geqslant d_{min}$. Consider first the parameter c as a function of the level of pruning p. At $p = 1$ (no pruning) $c_p = \hat{d}_{min}$ and $d_p = \hat{d}_{min}$; c_p and d_p decrease as p increases, but at a cost of increasing $\hat{L}_J^{(R)}$. The trade-off is to choose p so as to reduce the cardinality c_p (and hence reduce overfitting) without too dramatic an adverse effect on $\hat{L}_J^{(R)}$.

The second curve of interest represents dimensionality, and considers the error $\hat{L}_J^{(R)}$ as a function of the level of clustering d. When d is large, $\hat{L}_J^{(R)}(g_{\mathscr{W}'_{p,d}})$ will be small. In particular, $d = d_p$, the dimensionality of the pruned cover $(\mathscr{W}'_p, \mathscr{R}'_p)$, implies $\hat{L}_J^{(R)}(g_{\mathscr{W}'_{p,d}}) = \hat{L}_J^{(R)}(g_{\mathscr{W}'_p})$ by construction. Furthermore, for $d = d'_p$, the dimensionality of the collapsed pruned cover, implies $\hat{L}_J^{(R)}(g_{\mathscr{W}'_{p,d}}) = \hat{L}_J^{(R)}(g_{\mathscr{W}'_p})$. As $d \searrow 1$, $\hat{L}^{(R)}(g_{\mathscr{W}'_{p,d}})$ will increase. The location of the "elbow" in this curve can be interpreted as the "scale dimension" of the high-dimensional classification problem, and represents the number of "fundamental scales" required for the target class cover.

Given $p \geqslant 1$ and $\alpha \in [0,1]$, the *scale dimension* or the "elbow" in the dimensionality vs. error curve is defined to be $d^\star = d^\star_{\alpha,p} = \min\{\arg\min_d \hat{L}_J^{(R)}(g_{\mathscr{W}'_{p,d}}) + \alpha d\}$. Our investigations involve beginning with the algorithmically generated cover $(\mathscr{W}, \mathscr{R})$ and investigating simultaneously: (i) the error $\hat{L}_J^{(R)}$ as a function of the level of clustering d and (ii) the cardinality c_p of this curve as a function of the level of pruning p.

The reduction in complexity provided by the dominating set can have a large effect on the computational requirements of the classifier. The scale dimension also reduces the complexity, by requiring fewer radii to be specified. The idea is to estimate the fundamental scales of the data. Fig. 1 illustrates the benefits of the reduction in complexity. Two simulations were performed. In each, 2-dimensional, 2 class data were drawn. The classes had disjoint compact supports. The CCCD, its dominating set, and the scale dimension were computed. The error between the ball coverage and the true target distribution was then computed. This error is defined as the area of nontarget support covered by the balls plus the area of target support not covered. This was repeated 100 times, and box plots of the error are depicted for each case. In both simulations the nontarget observations corresponded to uniform observations drawn from a rectangle containing the target support. In the top plot the target support was a thin rectangle, while in the bottom plot the target support consisted of the disjoint union of 4 balls, 2 of radius 1 and 2 of radius 2.

For the error computation using the scale dimension, the balls were clustered into $d^\star$ clusters and for each cluster all the radii were set to the minimum radius within the cluster. This has the effect of reducing the amount of overfitting of the nontarget support. Other values for the radius are possible, such as the median or average of the radii in the cluster. These are not pursued here. The choice was made for purposes of illustration only.

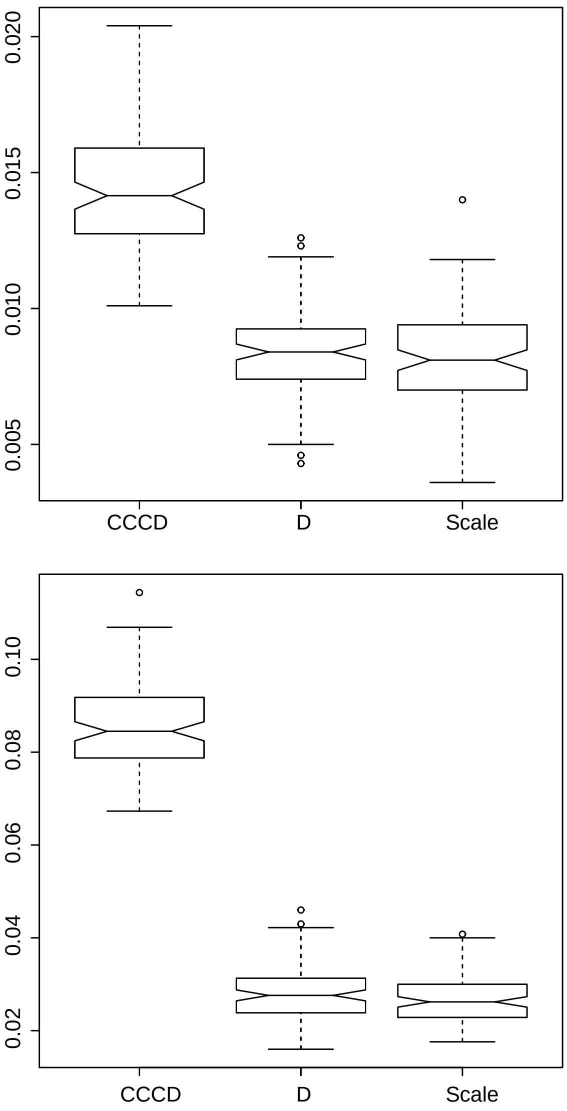

Fig. 1. Box plots of the errors in estimating the support for two different target distributions. In the top plot the target distribution is the rectangle defined by the corners $(1.25,1),(1.75,2)$. There were 100 target observations and 1800 nontarget observations within the region defined by the portion of the rectangle $(0,0),(3,3)$ exterior to the support of the target class. In the bottom plot the support is the union of the balls $B((0,0),1), B((3,0),1), B((5,4),2), B((0,4),2)$. There were 300 target observations and 500 nontarget observations within the region defined by the difference of the rectangle $(-3,-3),(8,8)$ and the target support. In each case 100 replications were performed to obtain the box plots.

As can be seen by the plots, using the dominating set dramatically reduces the error in the estimation of the target class support. This reduction is a result of the reduction in the overfitting of the CCCD. Since the CCCD covers all the area to the nearest nontarget observations it tends to err in the direction of covering too much of the nontarget support. The reduction by the dominating set tends to reduce this overfitting, at the expense of possibly covering less of the

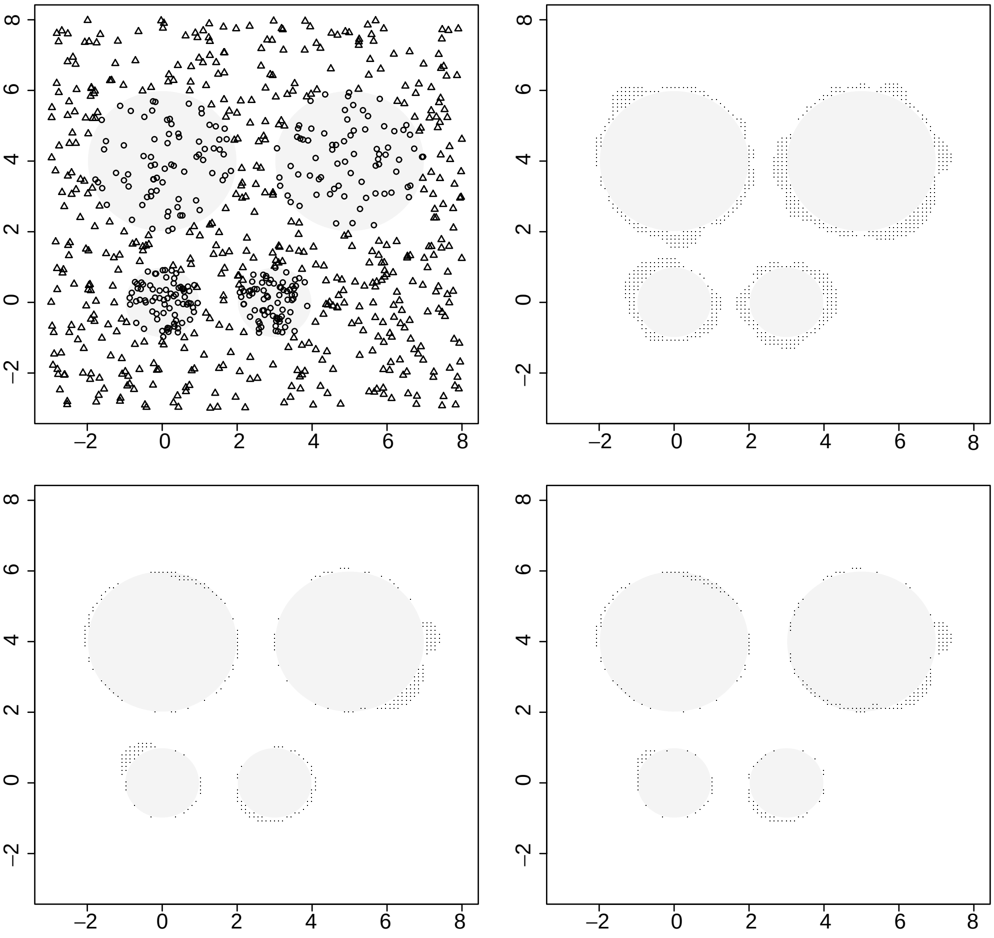

Fig. 2. An example of the effect of reduction of complexity. The support of the target class is indicated as gray circles. The upper left depicts the data overlayed on a depiction of the target support. The upper right, lower left and lower right depict the error regions for the CCCD, dominating set, and scale dimension respectively, where regions of error are indicated by the small black dots.

target support. As these examples illustrate, the reduction in overcoverage of the nontarget tends to be greater than the undercoverage of the target.

This is illustrated for a particular example in Fig. 2. In this case we see a data set plotted over a representation of the support, with the regions of error for the CCCD, the dominating set, and the scale dimension. This shows the trade-off between (nearly) complete coverage of the target support and overcoverage of the nontarget support provided by the reduction to the dominating set.

The scale dimension is best thought of as a method for investigating the structure of the data, rather than as a way to further reduce the complexity of the CCCD. If used as a classifier, the choice for cluster radius is critical for the functioning of the classifier. An alternative use would be as an input to an algorithm to cover the target class under the restriction of using at most $d^\star$ different sized balls.

3. Simulation examples

To demonstrate the concepts developed above, we present four pedagogical simulation examples.

3.1. Case I

For Case I the domain space class-conditional scatter plot ($\Xi = \mathbb{R}^2$) and the algorithmically produced cover $(\mathscr{W}, \mathscr{R})$ for the target class observations $\mathscr{X}_J$ are presented in Fig. 3, The class-conditional probability density functions f_j for this case are both uniform; for the target class (observations represented in Fig. 3 by circles) the support of f_J is the union of 2 disjoint balls with different radii, while for class $1-J$ the observations (represented in Fig. 3 by triangles) the density is uniform on $[-0.5, 1.5] \times [-0.5, 1.5] \backslash support(f_J)$.

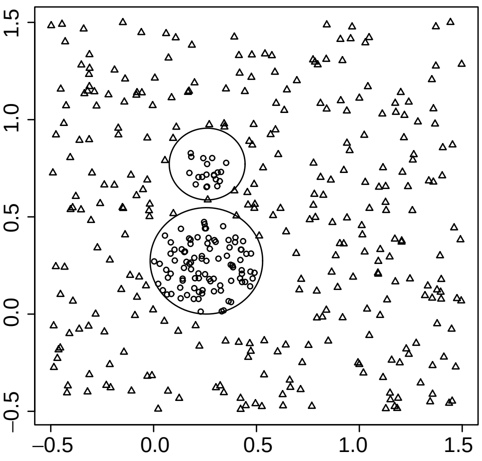

Fig. 3. Case I. The domain space class-conditional scatter plot ($\Xi = \mathbb{R}^2$) and the algorithmically produced cover ($\mathscr{W},\mathscr{R}$) for the target class observations $\mathscr{X}_J$ (represented by "o"s). The scale dimensional $d^\star = 2$, and the cardinality $c = 2$; a cover with two fundamental scales and two witness elements yields $\hat{L}_J^{(R)} = 0$.

For this simplest nontrivial problem there are two fundamental scales. Note, however, that the data, being random, need not cooperate. That is, there is positive probability that no target class observation will fall close enough to the center of one or both target class domain regions to allow for a pure proper unit cover with exactly two witness sets. Furthermore, the algorithm used to determine the cover is not guaranteed to find the best cover. Fig. 3 indicates that, in this case, the probability gods have smiled on us and the algorithm is successful with $\hat{L}_J^{(R)} = 0$ using exactly two unit witness sets. No pruning is necessary, so the cardinality for this example is $c_1 = 2$. Our error $\hat{L}_J^{(R)}(g_{\mathscr{W}'_{1,2}}) = 0$ so the scale dimensional for this example is $d^\star = 2$.

3.2. Case II

For Case II, the domain space class-conditional scatter plot and the algorithmically produced cover ($\mathscr{W},\mathscr{R}$) for the target class observations $\mathscr{X}_J$ are presented in Fig. 4(a) and the dendogram for the complete linkage clustering of the ball radii $\mathscr{R}$ is presented in Fig. 4(b). The algorithmic result is a cover ($\mathscr{W},\mathscr{R}$) with dimensionality $\hat{d}_{min} = 6$. For this example there are two fundamental scales; *support*(f_J) is the union of one big ball and a collection of smaller balls arranged in an inverted "∪". The algorithmic result is 1 big ball and 5 smaller balls with similar radii. None of the small balls are superfluous, so pruning (at $p = 2$, say) has no effect. The domain-space depiction and the dendogram suggest a clustering at $d = 2$; one witness set of size one and one witness set of size five—the latter consisting of the five witness elements from the 5 small balls. This suggests a final cover of ($\mathscr{W}'_{2,2}, \mathscr{R}'_{2,2}$) with dimensionality $d = 2$ and cardinality $c = 6$. Fig. 5 gives the error $\hat{L}_J^{(R)}(g_{\mathscr{W}'_{2,d}})$ as a function of the level of clustering d for this example. Clearly $d^\star = 2$. (Precisely, $d^\star_\alpha = 2$ for $\alpha \in [0.02, 0.16)$.) The cardinality c_p of this curve as a function of the level of pruning p is equally clear; $\hat{L}_J^{(R)}(g_{\mathscr{W}'_{p,2}}) = 0$ and $c_p = 6$ for p about 15 and then, once pruning takes effect, $\hat{L}_J^{(R)}(g_{\mathscr{W}'_{p,2}})$ begins to increase as c_p decreases for p increasing beyond 15. Clearly $c = 6$ total witness elements is the correct number to use for this example.

Fig. 6 shows the intersection graph of the cover from Fig. 4(a). In this, there is an edge between two vertices if their balls intersect. We can see that this provides us with some

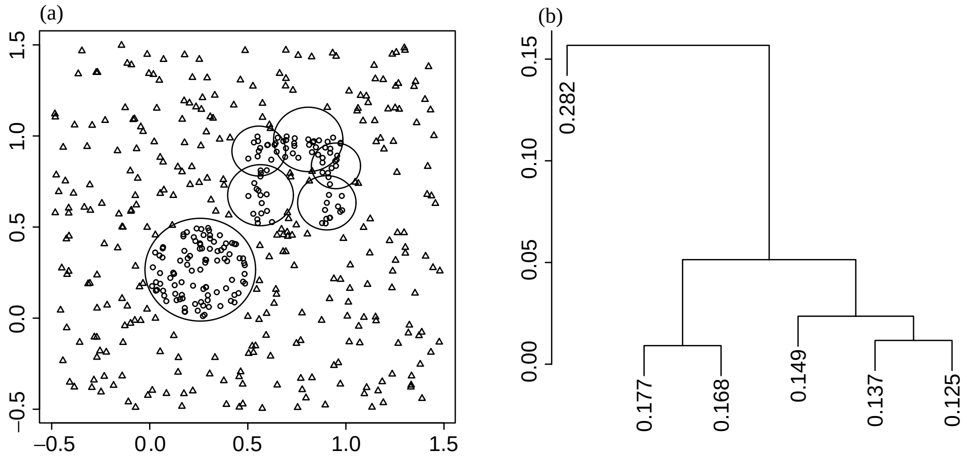

Fig. 4. Case II. (a) The domain space class-conditional scatter plot ($\Xi = \mathbb{R}^2$) and the algorithmically produced cover ($\mathscr{W},\mathscr{R}$) for the target class observations $\mathscr{X}_J$ (represented by "o"s). The cover depicted is of dimensionality $\hat{d}_{min} = 6$. (b) The dendogram for the complete linkage clustering of the 6 ball radii $\mathscr{R}$. The radii label the leaves of the tree. (The canonical two-clustering uses the two fundamental scales, 1 large and 5 small balls. See Fig. 5.)

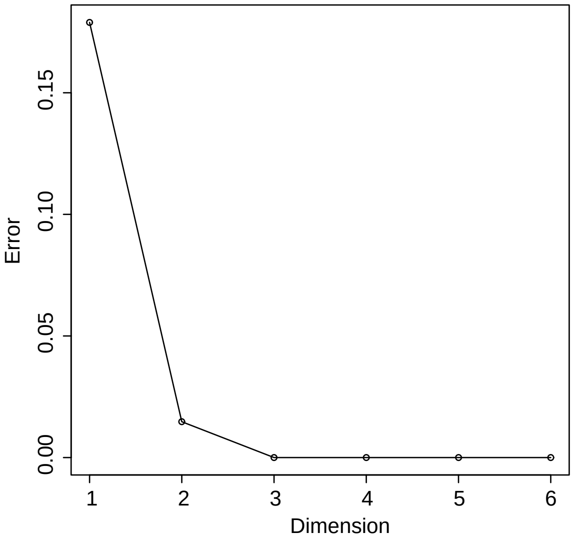

Fig. 5. Case II. The dimensionality vs. error curve. The scale dimension $d^\star = 2$, and the cardinality $c = 6$. (The former result is based on the location of the "elbow" in the curve. The latter stems from inspection of Fig. 4; pruning the smallest of the covering balls increases the error significantly.)

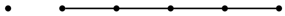

Fig. 6. Case II. The intersection graph of the cover in Fig. 4(a). An edge occurs between two vertices if and only if their corresponding balls intersect. The singleton corresponds to the large ball, while the chain corresponds to the five balls defining the "U".

information about the geometry of the problem. The graph has two components, indicating that the support is disconnected, and the chain corresponds to the "U" indicating that there is a "tube" in the support. Further information can be obtained by careful analysis of the relationships between the balls.

3.3. Case III

We now consider the case where the two distributions overlap. This is illustrated in Fig. 7. This is essentially the same as Case I, where now the distributions overlap in a small annulus about each circle. This results in a large number of very small balls, each covering a small number of observations, as depicted in Fig. 7(a).

Using this cover, C_1, we have the dendogram depicted in Fig. 7(b). Fig. 7(c–d) depicts the cover and dendogram resulting from $prune_5(C_1)$.

Another possible approach to producing a smaller cover presents itself if we consider the cover C_2 resulting from the algorithm applied to $\mathscr{X}_{1-J}$. This is depicted in Fig. 7(e). Note that it too has the problem of too many small balls. The idea is to prune this cover (in this case at $p = 5$), and then remove the observations pruned from the training set. The algorithm is then run for $\mathscr{X}_J$ using this reduced training set. The result is depicted in Fig. 7(f).

This second approach suggests the following extension to the algorithm. First, construct covers for each class. Prune at some level p (possibly different for each class) and remove the points pruned from the training set. Then recompute the cover using the reduced training set. The error curves for this example are depicted in Fig. 8. Both approaches suggest the same value for the scale dimension, $d^\star = 2$.

3.4. Case IV

It is important to remember that the scale dimension is not related to the Euclidean dimension of the data. To illustrate this we drew data from the following distributions:

Class 0 : $N(0, I)$,
Class 1 : $N(\mu_1, I)$,

where $\mu_1 = (\mu, 0, \ldots, 0)$ is (possibly) nonzero only in the first coordinate, and I is the identity matrix. Two-hundred observations were drawn from each class. We plot the error curves for these data as μ runs from 0 to 8 in Fig. 9. As can be seen, the scale dimension is 8 or greater for $\mu = 0$ (complete overlap of the classes), and nearly 3/4 of the observations (143) are needed to cover the class, indicative of the overlap. By $\mu = 3$ we start to see some indication of an elbow, which is quite pronounced at $\mu = 5$. The scale dimension in this case is 2, and it is clear that one cluster covers the bulk of the data while the other covers points within the region of high overlap of the classes, close to the classification boundary.

This example also points out a slightly nonintuitive aspect of the scale dimension. Rather than a single "scale" corresponding to the target distribution, there are different scales, defined by the different sized balls, as the ball centers transition into the region of high overlap. As can be seen from the plots in Fig. 9, for reasonably low overlap the scale dimension is 2, corresponding to the small balls in the overlap region and large balls away from the overlap. Ultimately, as the difference in means increases, the scale dimension goes to 1.

Contrast this example with Cases I–III. In the first two, the scale dimension corresponded closely to the number of different sized components of the support of the target class density. Case III illustrated the effect of overlap, as investigated in this example. This illustrates the fact that scale dimension is fundamentally tied to the classification problem. It, along with the cover itself, provides information about the relative amount of overlap in various regions of the domain.

4. Artificial nose experimental example

This section presents an empirical investigation of the Tufts artificial nose sensor data. These data are taken from

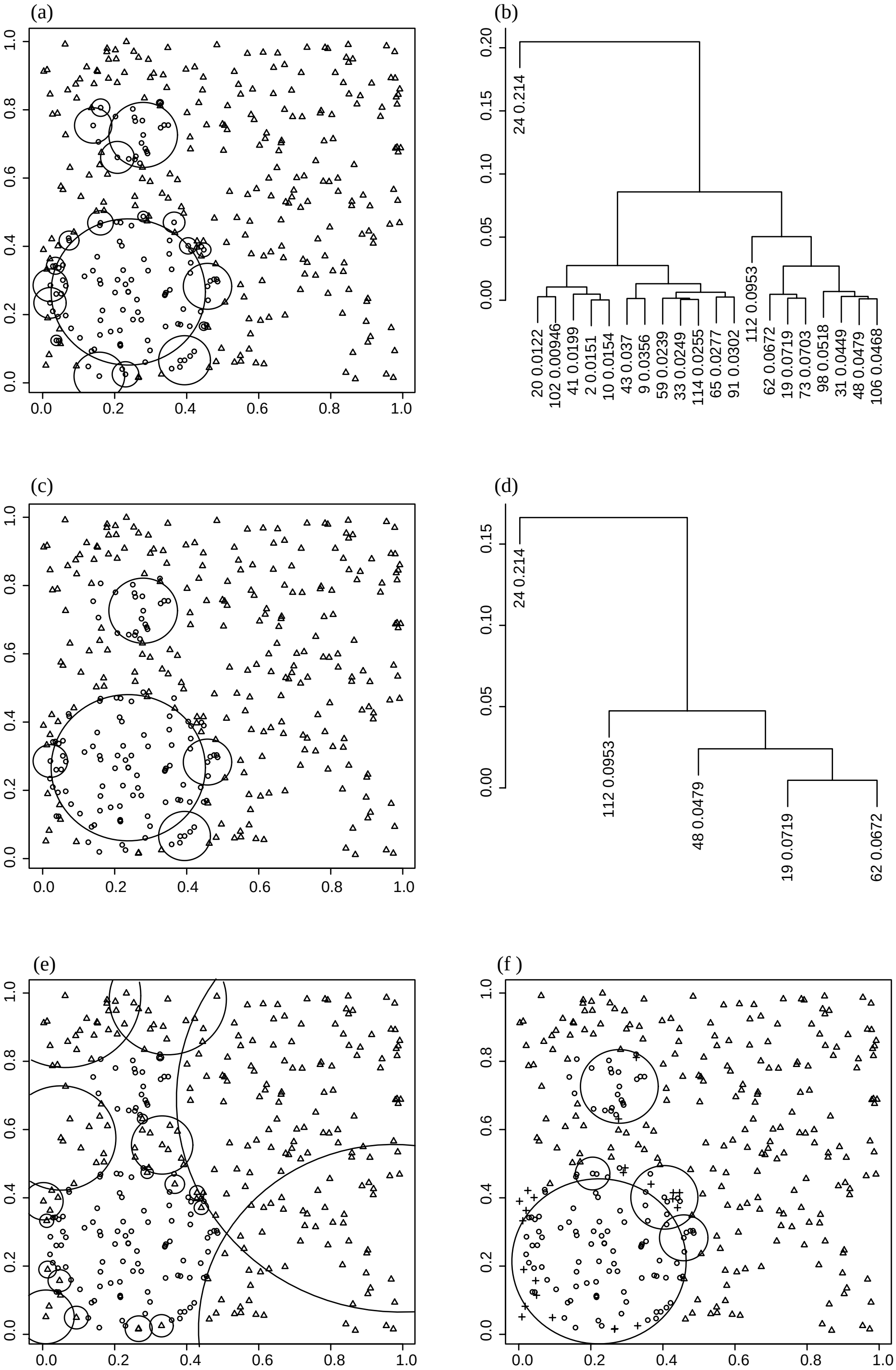

Fig. 7. Case III. (a) The domain space class-conditional scatter plot and the cover $(\mathscr{W}, \mathscr{R})$ for the target class observations $\mathscr{X}_J$ (represented by "o"s). (b) The dendogram for the clustering of the ball radii. (c) The pruned cover, pruned at $p = 5$. (d) The dendogram for the pruned cover. (e) The cover produced using $\mathscr{X}_{1-J}$. (f) The cover for $\mathscr{X}_J$ resulting from eliminating those $\mathscr{X}_{1-J}$ observations pruned at $p = 5$. Observations removed from processing are indicated with a "+".

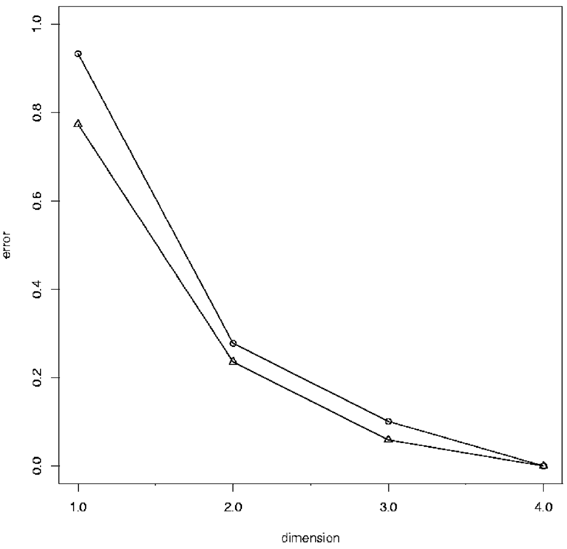

Fig. 8. Case III. The dimensionality vs. error curves. The triangles correspond to the default algorithm (Fig. 4(c)) while the circles correspond to the reduced version (Fig. 4(f)).

a fiber optic system constructed at Tufts University. The Tufts sensor consists of a 19-fiber optic bundle. The fibers are chemically doped with a solvatochromic dye (see Ref. [12]). This doping results in a sensor for which a change in fluorescence intensity is in response to interactions of the dye in each fiber with the chemical environment [13]. 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. 10).

The data set we will consider here consists of recordings of sensor responses to various analytes at various concentrations. Each observation is a measurement of the fluorescence intensity response at each of two wavelengths (620 and 680 nm) for each sensor in the 19-fiber bundle as a function of time. The Tufts sensor produces functional observations $x_i^{\phi,\lambda}$ for fibers $\phi \in \{1,\dots,19\}$ and wavelengths $\lambda \in \{1,2\}$. (The index $i = 1,\dots,n$ represents the observation number.) While the process is naturally described as functional with t ranging over a 20 s interval, the data as collected are discrete with the 20 s recorded at 60 equally spaced time steps for each response. Construction of the database involves taking replicate observations for the various analytes in various concentrations. Thus, each observation consists of 2280 values: 19 fibers at two wavelengths sampled 60 times.

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 subtracting the background sensor fluorescence (the intensity prior to exposure to the analyte) from each response to obtain the change in fluorescence intensity for each fiber at each wavelength.

The task at hand is the identification of an odorant observation. 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 ground water contaminant.)

In addition to TCE in air, eight diluting odorants are considered: BTEX (a mixture of benzene, toluene, ethylbenzene, and xylene, denoted BTE in the figures), benzene (Ben), carbon tetrachloride (CTe), chlorobenzene (ClB), chloroform (Clf), kerosene (Ker), octane (Oct), and Coleman fuel (WGa). A "T" in front of one of the above trigraphs indicates that the observation contains TCE, and hence is a target observation. Dilution concentrations of 1:10, 1:7, 1:2, 1:1, and saturated vapor are considered. In addition, there are 40 observations of TCE alone, with no confusers (denoted TCE_ below). Fig. 11 presents example sensor response signals indicating the importance of analyte mixture type, analyte mixture presentation, and fiber band.

The database D_n contains $n_0 = 352$ observations from class 0, the TCE-absent class. These consist of 32 observations of pure air and 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 $n = n_0 + n_1 = 1112$ observations in the training database D_n. This database is well designed to allow for 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.

Our goal is to investigate this high-dimensional data analysis problem, and perhaps to understand its structure. Fig. 12 presents the clustering dendogram for the case in which TCE-present is the target class, where pruning is employed at $p = 5$. Fig. 13 depicts four views of the largest component of the catch digraph of the pruned dominating set. This component consists of 47 elements, slightly more than half the elements. In this picture the directions of the arcs have been suppressed to avoid clutter. Recall that the observations that define this graph are very high dimensional, and so it is not surprising that the graph is not planar. We provide several projections of the graph onto the plane to aid in understanding the graph structure. One can see several cycles within this graph, indicating a large amount of overlap of the balls. There are also some chains, providing some evidence that the TCE class is not spherical, but rather somewhat spread out.

The other components of the catch digraph are all of order 3 or fewer, with 2 components of order 3, 4 of order 2, and

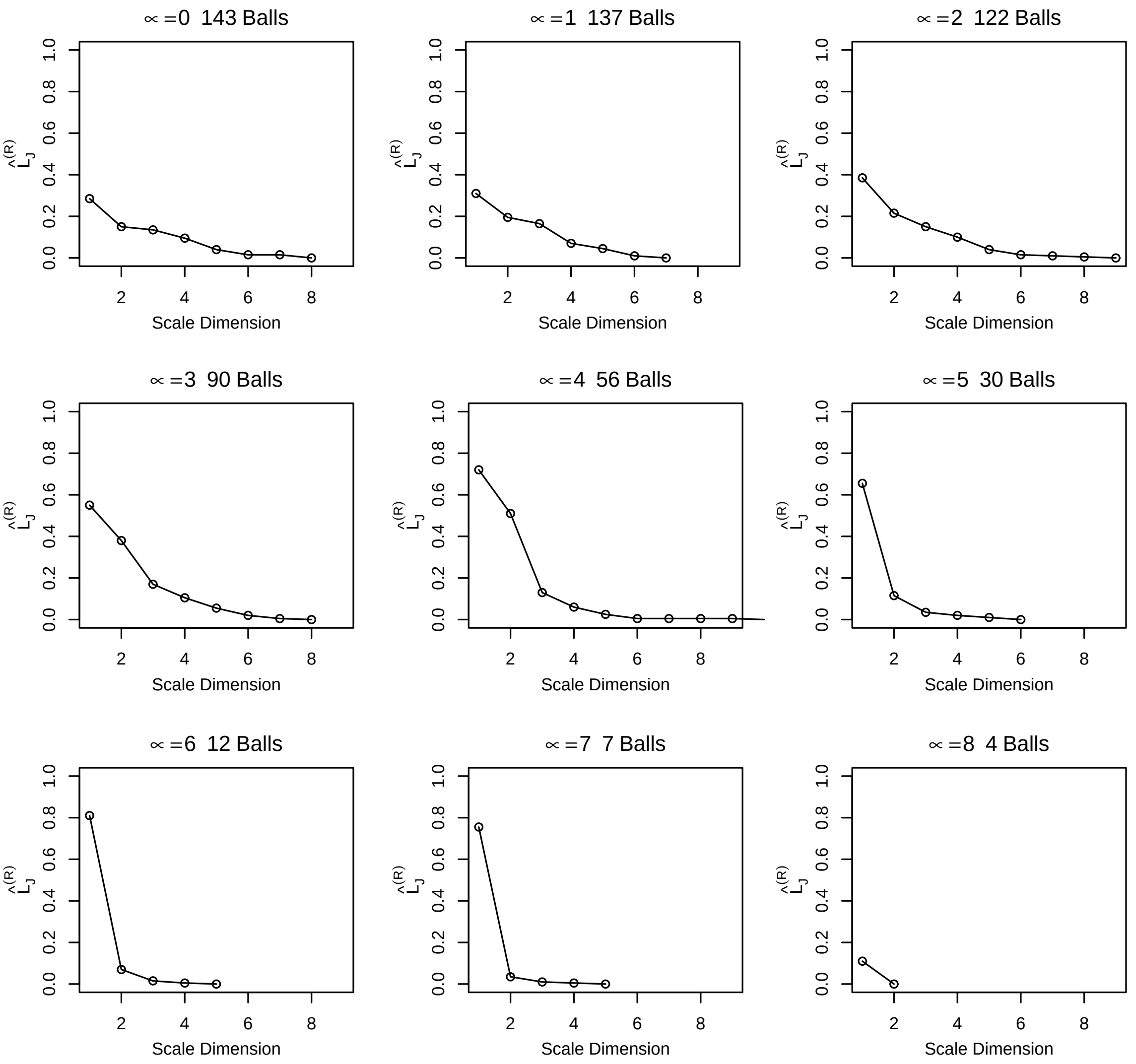

Fig. 9. Case IV. The dimensionality vs. error curves for data from two 50 dimensional normals. The distributions are identical $N(0, I)$ except for one variate, which has a different mean in 8 of the 9 plots (as indicated by the titles). The number of witness elements (vertices) in the graph are also displayed in the titles.

15 of order 1, for a total of 22 components. This indicates some measure of spread of the target class relative to the nontarget class.

The catch digraph does not give a full understanding of the geometry of the class. A related graph is the intersection graph, in which an edge is placed between vertices if their balls intersect. This can provide a better understanding of the geometry of the space. In this case, the intersection graph for the 76 elements of the pruned dominating set is connected. The graph is not the complete graph, in fact the ratio of the number of edges to the total number possible is 0.55, indicating that nearly half the possible edges are missing.

Fig. 14 depicts the histogram of the degrees of the vertices for the intersection graph. The five vertices with smallest degree are all chloroform. In fact, 7 of the smallest 9 are chloroform. The two with the largest degrees (60 and 62) are a CTe and a Clb, respectively. This suggests that the chloroform observations are, in general, outliers. This hypothesis has been verified by investigation of the interpoint distances. This is a potentially valuable insight into the application, as it suggests that distinguishing TClf may be a difficult task.

The scale dimension curves for this experiment are presented in Fig. 15. The error is low for maps with range-space dimensionality down to and including six, and the error increases dramatically when the range-space dimensionality decreases below 6. The analysis suggests that the scale dimension $d^\star \approx 6$. The cardinality for this cover is $c_5 = 76$,

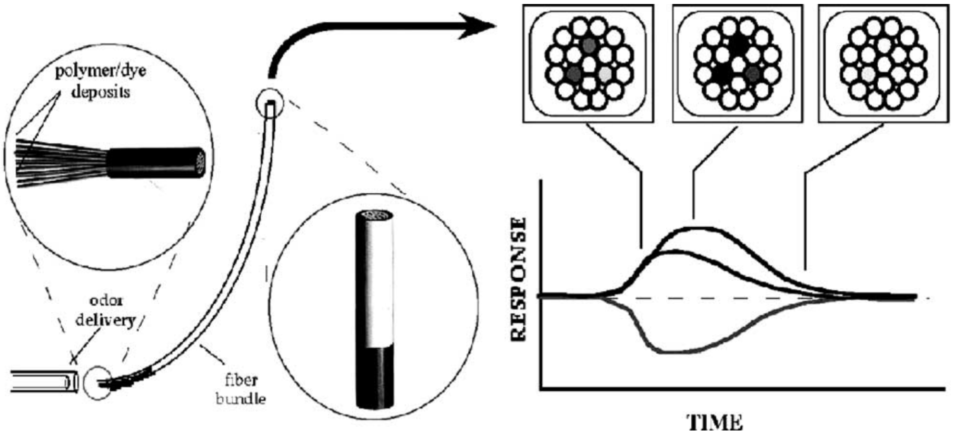

Fig. 10. The plot represents sensor/analyte signatures for three sensors within the bundled 19-sensor array. This figure was published in Nature 382 (1996) 697–700, and is reprinted by permission.

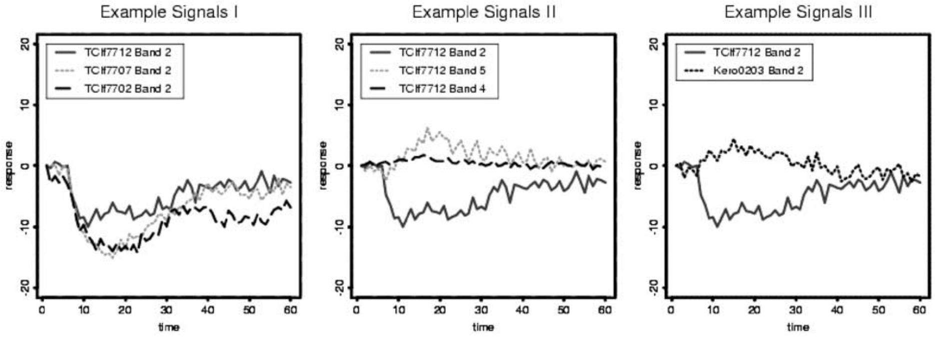

Fig. 11. Depicted are three (unsmoothed) sensor response signal examples: a comparison of a single fiber band for three different presentations of the same analyte mixture (left panel), a comparison of three different fiber bands for a single analyte mixture presentation (middle panel), and a comparison of the same fiber band for two different analyte mixture presentations (right panel). "TClf" is TCE in chloroform and air. (All TClf observations are at the same concentration.) "Kero" is Kerosene, sans TCE. This figure originally appeared in IEEE PAMI 23 (4) (2001) 404–413, C.E. Priebe, Olfactory classification via interpoint distance analysis, copyright IEEE, and is reprinted by permission.

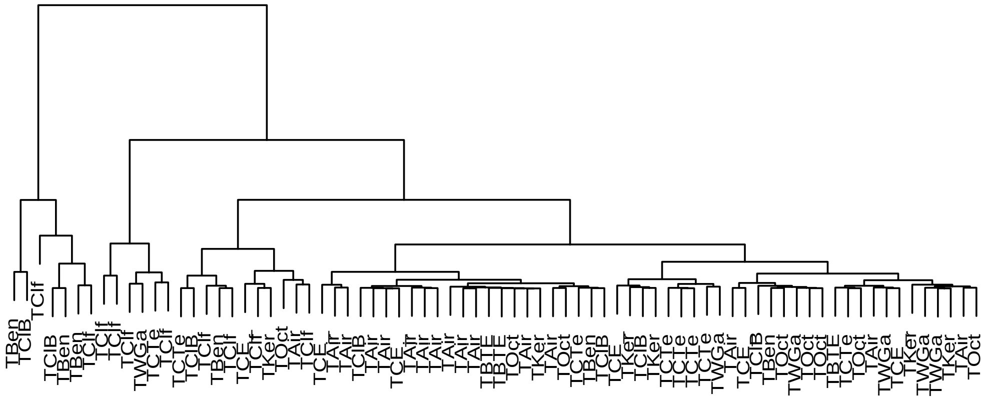

Fig. 12. Cluster tree for the Tufts artificial nose, after pruning at $p = 5$. (The cardinality $c_5 = 76$.) The leaf label indicates the chemical content of the observation.

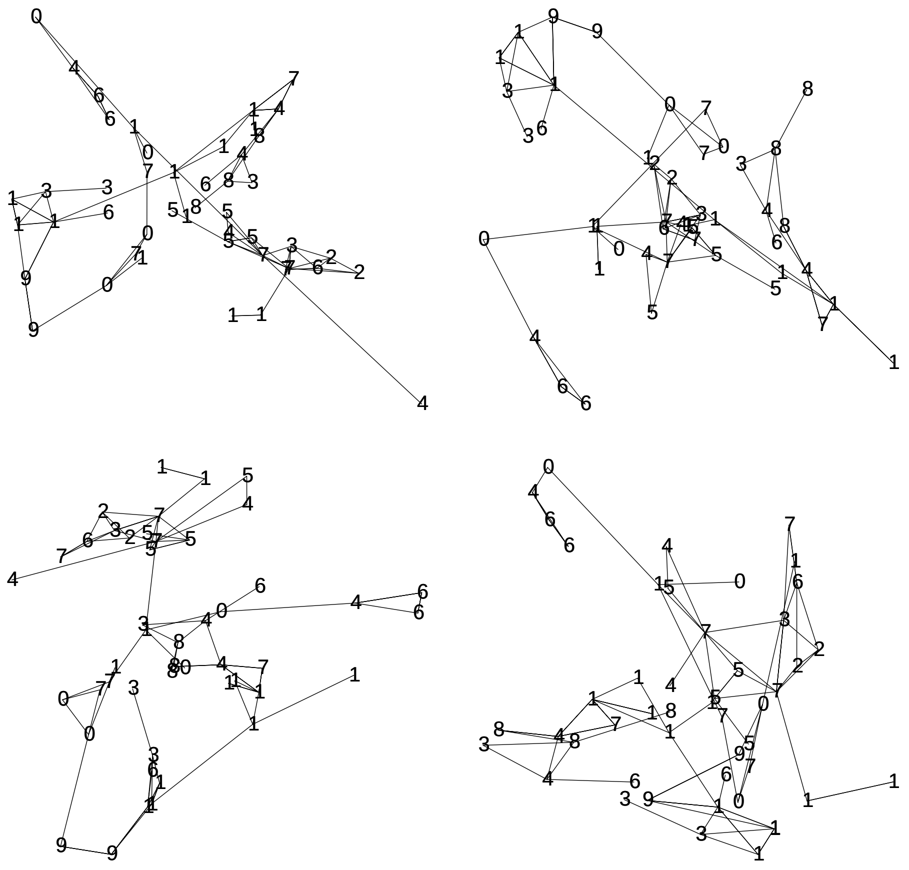

Fig. 13. Four views of the largest component of the catch digraph of the dominating set for the nose data after pruning at $p = 5$. The vertices are numbered according to the chemical confuser: 0 = TWGa, 1 = TAir, 2 = TBTE, 3 = TCE_, 4 = TBen, 5 = TClf, 6 = TClB, 7 = TCTe, 8 = TKer, 9 = TOct.

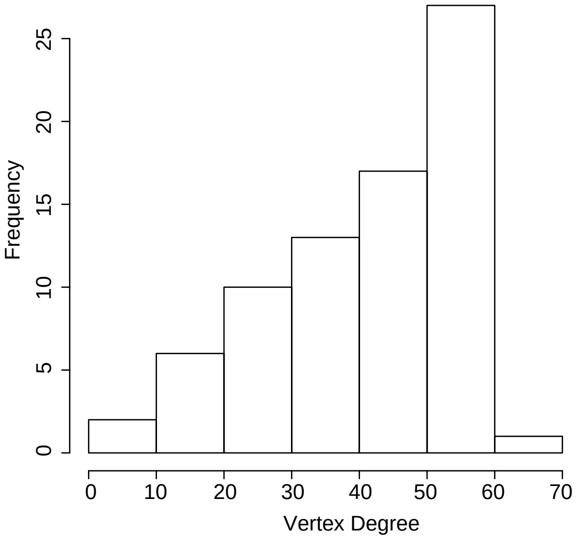

Fig. 14. Histogram of the degrees for the vertices in the intersection graph of the dominating set for the nose data after pruning at $p=5$.

indicating that 10% of the target class training observations are used as exemplars in the model.

Consider the clusters resulting from the scale dimension of 6. Looking at the radii that make up the clusters we obtain the statistics of Table 1. The last (largest radius) cluster, corresponds to a single chloroform observation. The ball associated with this observations contains only 13 chloroform observations. The next to last cluster consists of two benzenes, a chloroform and a chlorobenzene. These cover 101 observations among them. These two clusters have only two observations of overlap, so they are quite distinct.

It is also clear that the vast majority of the balls are relatively small, indicating that much of the data lies close to the nontarget class. The smallest cluster of 52 elements covers 353 observations (nearly half of the total target observations). These are primarily observations consisting of low-concentration TCE and/or high-concentration confusers.

Through analysis of the CCCD, the dominating set, and the related graphs, we obtain quite a bit of information about the structure of the TCE class and its relationship to the

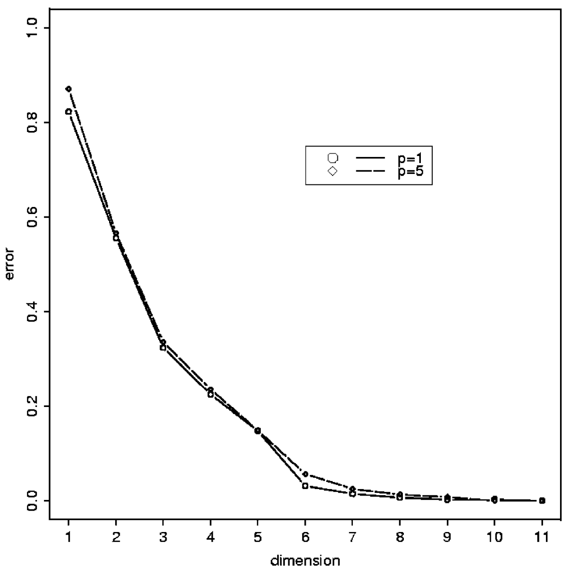

Fig. 15. Complexity characterization for the Tufts artificial nose. (The target class is TCE-present.) The x-axis is the range-space dimensionality of the nonlinear point-to-subset map $\rho_{\mathcal{W}'_p}$, and the y-axis is the estimated error rate $\hat{L}_J^{(R)}(g_{\mathcal{W}'_p})$ for a given level of pruning p. The different curves represent different levels of pruning applied to the cover cluster tree. Pruning at $p = 5$ yields an error nearly as good as that of no pruning ($p = 1$); other levels of pruning yield higher error. This analysis suggests that the scale dimension for this problem is $d^\star \approx 6$. (The cardinality of the cover at $p = 5$ is $c_5 = 76$, or approximately 10% of the target class training observations.)

Table 1
Radii ranges for the 6 clusters of the nose data. The column labeled "Size" indicates the number of elements in the cluster

Min radius	Max radius	Size
281.1803	689.5605	52
737.0896	1104.3983	11
1243.1260	1660.4067	6
2091.1148	2244.6562	2
2422.3908	2653.1308	4
2913.0716	2913.0716	1

confusers. We see that chloroform is somewhat different from the other analytes. It is farthest from the nontarget class, and there is evidence that it clusters. However, it is fairly spread out, as seen by the fact that the observations with smallest degree in the intersection graph are chloroform.

From Table 1 we see that there are few very small or very large balls. Recall that we have pruned at $p = 5$ so we have eliminated balls containing few observations. These balls might reasonably be thought to be ones close to the nontarget observations, and hence of small radius. This is an area for further analysis.

Note that if one were to draw 1112 observations uniformly from the unit cube in R^{2280} one would expect the interpoint distances to be roughly equal ([2, p. 29] shows that all the observations would be in the corners), which would result in a CCCD consisting of very few arcs, and yet the intersection graph would be essentially the complete graph. The analysis shows that there is quite a bit more structure to these data than this.

5. Discussion

In his 1992 book [2, p. 196], Scott writes "Fortunately, it appears that in practical situations, the dimension of the structure seldom exceeds 4 or 5". While our definition of "structure" is quite different than Scott's, our investigation of the Tufts artificial nose data set gives, we think, evidence in favor of Scott's premise; the scale dimension $d^\star$ of this data set, for classification purposes, is determined to be $d^\star \approx 6$.

At this stage, our approach is exploratory. We believe that investigating scale dimension and cover cardinality for high-dimensional classification problems will provide insights into the structure of the data which will, in turn, prove useful in model building. Edward J. Wegman [personal communication] suggests "Data Mining is an extension of exploratory data analysis and has basically the same goals, the discovery of unknown and unanticipated structure in the data. The chief distinction between the two topics resides in the size and dimensionality of the data sets involved. Data mining in general deals with much more massive data sets for which highly interactive analysis is not fully feasible". Thus, we see our methodology as determining the exemplars for *vector quantization* (e.g. *radial basis functions*, *k-means clustering*, *support vector machine*) for data mining applications (see, for example, Refs. [14,15]). Returning to Fig. 12 and Table 1 we find that one of the dimensions is characterized by a single Clf witness element, one by two Ben witness elements, a chloroform and a chlorobenzene, and one by a benzene and a chlorobenzene. Thus, the largest balls, and hence the observations least like nontarget, are chloroform, benzene and chlorobenzene. The implications of this analysis for sensor design and model building are under investigation.

Additional investigations are called for. First, error investigations using the deleted (cross-validation) error estimate $\hat{L}^{(D)}$ as opposed to the resubstitution error estimate $\hat{L}^{(R)}$ will be valuable. Development of an unsupervised methodology, as opposed to the supervised procedure detailed here, is of interest. The Tufts artificial nose chemical sensor data analysis problem is in fact polychotomous; while the *detection* of a distinguished analyte (such as TCE) is often relevant, in many cases the question at hand involves *identification* of an unknown analyte. Popular approaches to the multi-class problem involve addressing multiple two-class subproblems [16,17]; thus, our methodology remains applicable. Further investigation of the multi-class case in ongoing. Finally, inferring topological structure from the cover is difficult, and is an area for future research.

6. Summary

We present an exploratory data analysis methodology for obtaining information about the high-dimensional decision boundary characterizing the dimensionality of a classification problem and provide a nonlinear map under which classification can be performed. We characterize the support of one distinguished target class as a collection of balls covering the class, with each ball centered at an observation in that class such that the radius is maximal without containing observations from the other classes. A greedy algorithm for fitting the balls is proposed. The balls then provide a description of the support of the target class, with information about the complexity of the classification problem implicit in the number, radii, adjacency and position of the balls. Clustering the balls by radius, pruning the cluster tree, and mapping the data based on distances to the clusters yields a nonlinear map to a (usually lower-dimensional) space in which classification can be performed. The range-space dimensionality of this map is defined to be the scale dimension of the classification problem. We illustrate the methodology with pedagogical simulations and an “artificial nose” chemical sensor data analysis application.

Acknowledgements

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 authors would also like to thank Ed Wegman, Dennis Healy, and Jason DeVinney for many helpful discussions.

References

[1] J.F. Maa, D.K. Pearl, R. Bartoszynski, Reducing multidimensional two-sample data to one-dimensional interpoint comparisons, Ann. Stat. 24 (1996) 1069–1074.

[2] D.W. Scott, Multivariate Density Estimation, Wiley, New York, 1992.

[3] A.K. Jain, R.P.W. Duin, Jianchang Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.

[4] A. Cannon, L.C. Cowen, Approximation algorithms for the class cover problem, Ann. Math. Artif. Intell., to appear.

[5] G. Chartrand, L. Lesniak, Graphs & Digraphs, Chapman & Hall/CRC, Boca Raton, 1996.

[6] Hiroshi Maehara, A digraph represented by a family of boxes or spheres, J. Graph Theory 8 (1984) 431–439.

[7] C.E. Priebe, J.G. DeVinney, D.J. Marchette, On the distribution of the domination number for random class cover catch digraphs, Stat. Probability Lett. 55 (2001) 239–246.

[8] S. Arora, C. Lund, Hardness of approximations, in: D.S. Hochbaum (Ed.), Approximation Algorithms for NP-Hard Problems, PWS Publishing Company, Boston, 1997, pp. 399–446.

[9] A.K. Parekh, Analysis of a greedy heuristic for finding small dominating sets in graphs, Inf. Process. Lett. 39 (1991) 237–240.

[10] L.R. Foulds, Graph Theory Applications, Springer, New York, 1992.

[11] J.A. Hartigan, Clustering Algorithms, Wiley, New York, 1975.

[12] J. White, J.S. Kauer, T.A. Dickinson, D.R. Walt, Rapid analyte recognition in a device based on optical sensors and the olfactory system, Anal. Chem. 68 (1996) 2191–2202.

[13] T.A. Dickinson, J. White, J.S. Kauer, D.R. Walt, A chemical-detecting system based on a cross-reactive optical sensor array, Nature 382 (1996) 697–700.

[14] J. Schurmann, Pattern Classification, Wiley, New York, 1996.

[15] N. Cristianini, J. Shawe-Taylor, Support Vector Machines, Cambridge University Press, Cambridge, 2000.

[16] J.H. Friedman, Another approach to polychotomous classification, Technical Report, Stanford University, 1996.

[17] T. Hastie, R. Tibshirani, Classification by pairwise coupling, Ann. Stat. 26 (2) (1998) 451–471.

About the Author—DAVID MARCHETTE received a B.A. in 1980, and an M.A. in Mathematics in 1982, from the University of California at San Diego. He received a Ph.D. in Computational Sciences and Informatics in 1996 from George Mason University under the direction of Ed Wegman. From 1985 to 1994, he worked at the Naval Ocean Systems Center in San Diego doing research on pattern recognition and computational statistics. In 1994 he moved to the Naval Surface Warfare Center in Dahlgren, VA where he does research in computational statistics and pattern recognition, primarily applied to image processing and automatic target recognition.

About the Author—CAREY E. PRIEBE received the B.S. degree in Mathematics from Purdue University in 1984, the M.S. degree in Computer Science from San Diego State University in 1988, and the Ph.D. 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, MD. His research interests are in computational statistics, kernel and mixture estimates, statistical pattern recognition, and statistical image analysis.