International Journal of Image and Graphics, Vol. 2, No. 1 (2002) 1–17

A VISUALIZATION FRAMEWORK FOR THE ANALYSIS OF HYPERDIMENSIONAL DATA

JEFFREY L. SOLKA*,† and BARTON T. CLARK‡

Code B10, NSWCDD, 17320 Dahlgren Rd., Dahlgren, Virginia 22448-5000, USA
†*jsolka@nswc.navy.mil*
‡*clarkbt@nswc.navy.mil*

CAREY E. PRIEBE

Department of Mathematical Sciences, Johns Hopkins University, Baltimore, Maryland 21218-2682, USA
cep@jhu.edu

Received 25 September 2001
Accepted 4 October 2001

The purpose of this article is to describe a new visualization framework for the analysis of hyperdimensional data. This framework was developed in order to facilitate the study of a new class of classifiers designated class cover catch digraphs. The class cover catch digraph is an original random graph technique for the construction of classifiers on high dimensional data. This framework allows the user to study the geometric structure of hyperdimensional data sets via the reduction of the original hyperdimensional space to a cover with a small number of balls. The framework allows for the elicitation of geometric and other structures through the visualization of the relationships between the balls and each other and the observations they cover.

Keywords: Graph; Classifier; Visualization; Data Mining.

1. Introduction

Discriminant and cluster analysis on a set of n observations in d-dimensional space is complicated by the curse of dimensionality.[1] As d grows the number of parameters required to model the data, for example using kernel and mixture-based approaches grows simililarly.[2] Kernel-based methods also have the disadvantage of needing to store all of the observations associated with the training set.

Nevertheless, one is often faced with the need to analyze data sets that suffer from large d, large n or a combination of these two factors. In the two-class case the discriminant analysis problem is characterized by the interpoint distance structure between the class 0 and class 1 observations. These, along with the within class

*Corresponding Author

interpoint distances, determine the geometric structure of the observations in the high-dimensional space.

We (C. E. Priebe) have developed a new graph theoretic classification procedure based on the use of class cover catch digraphs (CCCD). This classifier was specifically designed in order to analyze moderately sized, $n < 10000$ data sets in hyperdimensional spaces. By hyperdimensional we mean those data sets with a d value in the thousands. The reader is referred to Priebe and Marchette[3] and Priebe *et al.*[4] for a more thorough discussion of the details of this procedure.

The focus of our discussions is a visualization framework, the Interactive Hyperspectral Exploratory Data Analysis Tool (IHEDAT), that was developed in order to study the interaction between the CCCD based classifier and hyperdimensional data sets. This tool has proven beneficial in the fine tuning of the CCCD algorithms, in the study of geometric structure within the data sets, and in the identification of latent classes within the data sets.[5]

IHEDAT has been successfully applied to numerous data sets. These include an artificial nose data set, a gene expression data set, and a hyperspectral data set. We will provide an in-depth discussion on the use of IHEDAT to analyse an artificial nose data set.

We will begin with a brief overview of the CCCD methodology followed by an informative two-dimensional pedagogical example. Next, we will provide some brief background material on the visualization components that make up IHEDAT. Finally, the application of IHEDAT to the artificial nose data set will be discussed.

2. Methodology

We begin by considering two disjoint sets of d-dimensional observations, $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \Re^d$ and $\mathcal{Y} = \{y_1, \ldots, y_m\} \subset \Re^d$. We initialize the procedure by designating $\mathcal{X}$ as the "target class." Although our procedure as stated is asymmetric in target class, it can be made symmetric by repeating the procedure choosing the other class as "target" in turn. The reader is referred to Priebe and Marchette, *et al.*[4] for an exposition of this idea as applied to classification.

Proceeding as in Priebe, DeVinney and Marchette,[6] we define the class cover catch digraph (CCCD) $D = (V, A)$ for $\mathcal{X}$ against $\mathcal{Y}$ as follows. Our set of vertices $V = \mathcal{X}$ (the set of target class observations). Now for each $v \in V$, we define $B_v := B(v, \min_{y \in \mathcal{Y}} \rho(v, y)) := \{z \in \Re^d : \rho(v, z) < \min_{y \in \mathcal{Y}} \rho(v, y)\}$ where $\rho : \Re^d \times \Re^d \to \Re_+ := [0, \infty)$ is a distance or pseudo-distance function. We may think of each B_v as an open ball of maximum radius around v such that the ball contains only target class observations. In our discussions within ρ will be L_2 the standard Euclidean metric. There is an arc or directed edge $vw \in A \iff w \in B_v$. So an arc exists between each observation and the observations covered by its ball.

Following standard graph theoretic terminology we define a *dominating set* S for D as a set $S \subset V$ such that, for all $w \in V$, either $w \in S$ or $vw \in A$ for some $v \in S$. Under this definition each vertex is either a member of the dominating set

or is connected to a member of the dominating set via an arc. The domination number $\gamma(D)$ is defined to be the cardinality of the smallest dominating set(s) of D. In general $\gamma(D)$ is invariant and $1 \leq \gamma(D) \leq \text{cardinality}\ (V) = n$. A minimum dominating set for D is defined as a dominating set with cardinality $\gamma(D)$. The production of a minimum dominating set in a general digraph is an NP-Hard problem. In practice we find an approximate minimum dominating set $\hat{S}$ using a greedy algorithm. The reader is referred to Priebe, DeVinney and Marchette,[6] for a full discussion of these methods. We will use $\hat{\gamma} = \text{cardinality}(\hat{S})$ as an estimate for the domination number of the digraph D.

Associated with each $v \in V$ is a radius $r_v := \min_{y \in \mathcal{Y}} \rho(v, y)$. The nature of the discriminant boundary is determined by the radii of these balls. Balls with the same sized radii are in a similar relationship to the nontarget observations. Hence, we apply agglomerative clustering on the radii $\{r_v : v \in \hat{S}\}$, producing a dendogram, or cluster tree.[7] The leaves of this dendogram correspond to the $\hat{\gamma}$ elements of $\hat{S}$.

We can use this dendogram to provide a sequence of "cluster maps" $m_k : R^d \to R^k_+$ for each $k = 1, \ldots, \hat{\gamma}$. Each cluster map produces a disjoint partition of $\hat{S}$. We can produce the cluster map into a given k-dimensional range-space by visually "cutting" the dendogram horizontally at a level which yields k branches, or clusters, $\hat{S}_1, \ldots, \hat{S}_k$. The kth cluster map is then defined as $m_k(x) = [\rho(x, \hat{S}_1), \ldots, \rho(x, \hat{S}_k)]'$, where the distance $\rho(x, S)$ from a point x to a set S is defined as the minimum over $s \in S$ of the distances $\rho(x, s)$. One can think of a particular partition of $\hat{S}$ as a means of capturing the target/nontarget relationship at that particular resolution level.

For each $k = 1, \ldots, \hat{\gamma}$ an empirical risk (resubstitution error rate estimate) $\hat{L}_k$ is calculated as

$$\hat{L}_k := \left(\frac{1}{n+m}\right)\left(\sum_{i=1}^{n} I\left\{x_i \notin \cup_{j=1,\ldots,k} \cup_{v \in \hat{S}_j} B\left(v, \min_{w \in \hat{S}_j} r_w\right)\right\}\right.$$
$$\left. + \sum_{i=1}^{m} I\left\{y_i \in \cup_{j=1,\ldots,k} \cup_{v \in \hat{S}_j} B\left(v, \min_{w \in \hat{S}_j} r_w\right)\right\}\right).$$

By construction, the empirical risk $\hat{L}_{\hat{\gamma}} = 0$. $\hat{L}_k$ may, however, be nonzero for $k < \hat{\gamma}$. We may use the empirical risk as a function of k to select a reasonable cluster map dimensionality. The "model complexity selection" problem is in general a very difficult problem. In our case it is necessary in order to characterize the geometric relationship between the target and nontarget observations.

We now define the "scale dimension" $\hat{d}^\star$ to be the cluster map dimension which minimizes a dimensionality-penalized empirical risk; $\hat{d}^\star_\delta := \min\{\arg\min_k \hat{L}_k + \delta \cdot k\}$ for some penalty coefficient $\delta \in [0, 1]$. By construction, we have $\hat{d}^\star_\delta = \min\{k : \hat{L}_k = 0\}$ for $\delta = 0$ and $\hat{d}^\star_\delta = 1$ for $\delta = 1$. The value of δ ultimately determines the sharpness required to locate the "elbow" or bend in the plot of empirical risk versus cluster map dimension. Hence, the scale dimension is the abscissa of the elbow in the curve. $\hat{d}^\star_\delta$ is of course an estimate of $d^\star_\delta$, the cluster map dimension

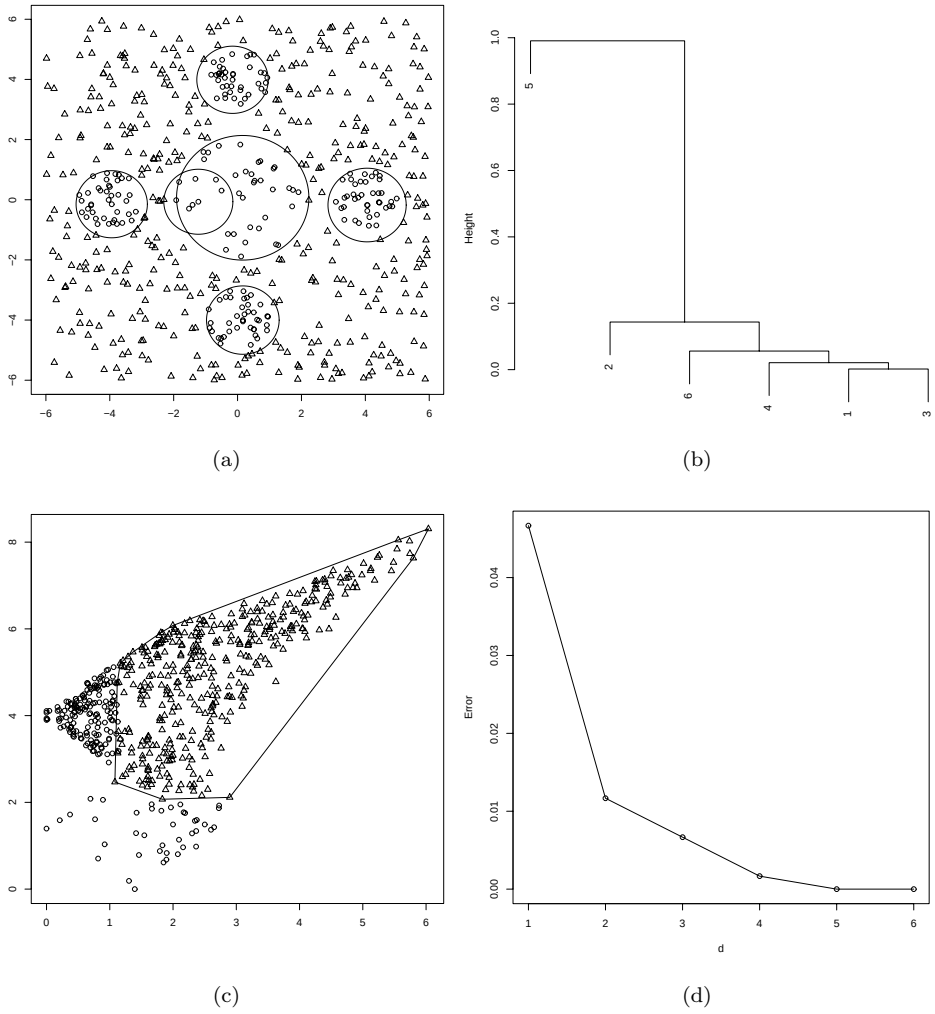


Fig. 1. Depiction of the simulation example. (a) The domain space class-conditional scatter plot, with dominating set $\hat{S}(\hat{\gamma} = 6)$ for the target class observations (represented by "∘"s). (b) The dendogram for the six radii. (c) The class-conditional scatter plot resulting from cluster map m_2 (with the convex hull of the projected non-target class observations). (d) The scale dimension curve.

which minimizes the penalized probability of misclassification. One way to think of the scale dimension is as the number of different sized balls needed to cover the class. In Fig. 1(a) we depict data of two scales, a large scale region where the ball is and a smaller scale region corresponding to the "+." The scale dimension $\hat{d}^{\star}$ is designed to estimate the number of scales appropriate to the data. Figure 1(d) presents the scale dimension plot for this data set.

$m_{\hat{d}^{\star}}$ is the cluster map of interest, which will be used for our exploratory data analysis. There are numerous relationships that one would like to be able to study

within this visualization framework. First and foremost, we are interested in ascertaining the interpoint distance relationship between the target and the nontarget observations. Second, we are interested in ascertaining the relationship between the target observations and the open balls. In particular, we would like to know which ball contains a particular observation. Finally, we are interested in studying the spatial relationship of the open balls to one another. Each of these needs is addressed within the framework of IHEDAT. Each of these will be discussed in turn in the subsequent sections.

3. Two-Dimensional Example

Let us consider, for the purpose of illustration, a simple two-dimensional simulation example. For this case the domain space class-conditional scatter plot and the algorithmically produced dominating set $\hat{S}$ (with $\hat{\gamma} = 6$) and the associated radii (one large and a collection of five smaller) for the target class observations are presented in Fig. 1(a), the dendogram for the complete linkage clustering of these six radii is presented in Fig. 1(b), and the two-dimensional range space class-conditional scatter plot (the result of the application of the cluster map m_2 to the observations of Fig. 1(a)) is presented in Fig. 1(c). In Fig. 1(d) we present a plot of the estimated classification error as a fuction of the number of clusters. The elbow in the curve corresponding to a scale dimension of 2 is clearly visible.

In this pedagogical example, clearly the scale dimension is 2. This is evidenced by the fact that there are two obvious clusters of ball radii.

4. Visualization Components

We will now discuss the visualization components that are used within the IHEDAT. Throughout this section we will refer to Fig. 2.

We need a convenient method to display the spatial relationship of the target and nontarget observations. This information is contained within the interpoint distance matrix. An examination of a tabulation of these values is somewhat daunting and hence, we have chosen to encode each of the values as a grayscale level which we then render as a grayscale image, see Fig. 2(a). This type of approach has been used previously to study data structure under clustering based permutations.[8] Greyscale renderings of the interpoint distance matrix has been used more recently by Marchette and Solka,[9] as a method for outlier detection.

The interpoint distance matrix can also be used to study the spatial structure of the dominating set elements. In this case the interpoint distances are measured from the center of one dominating set element to the center of another dominating set element. Once computed these can be rendered as a grayscale image as we do with the full data set interpoint distances.

The agglomerative clustering methodology is usually displayed using the dendrogram. We follow this practice and also include color information that displays the various clusters. Figure 2(b) illustrates an example dendrogram.

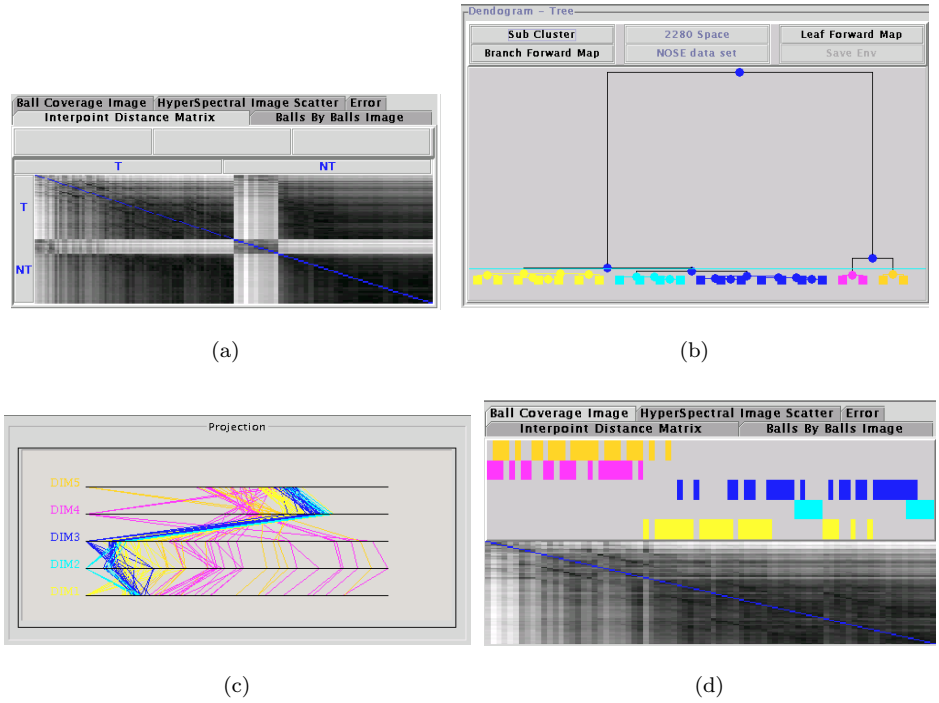


Fig. 2. Visualization components. (a) The interpoint distance matrix. (b) Radii based dendrogram. (c) Parallel coordinates plot. (d) Ball coverage plot.

Given a cluster map $m_k : R^d \to R^k_+$ we are interested in rendering the projected observations in the k-dimensional space. Although $k \ll d$, k is still typically larger than three and hence, one is faced with the task of plotting the observations in the high dimensional space. One relatively simple solution to this probably was posed by Wegman,[10] and Inselberg.[11] This method solves the problem of the impossibility of placing coordinate axes perpendicular to one another in the dimensions $d > 2$ by placing them parallel to one another. A point in Euclidean space becomes a broken line in parallel coordinates. The reader is refered to Wegman[10] for a full discussion of the rendering of various structures in parallel coordinate space. We use parallel coordinates to render the target observations in the k-dimensional space obtained via an application of the cluster mapping in Fig. 2(c).

We have discussed ways to visualize the target/nontarget relationships, the dominating set element relationships, and the cluster mapping. We are still interested in being able to display the relationship between the dominating set elements/balls and the target observations. By construction, each target is covered by the ball corresponding to at least one dominating set element. In fact it is often the case that each target point resides in more than one dominating set element. We have chosen to display this information by superimposing a color representation of dominating set containment above a grayscale rendering of the target versus target interpoint

distance matrix. Each column in the interpoint distance matrix corresponds to one of the target observations. Above this column we place a series of colored pixels. These colored pixels correspond to the same color scheme used in the dendrogram for a particular cluster map. In this manner the coloring of the dendrogram and the coloring of the dominating set containment map is consistent. Figure 2(d) presents an example of this visualization component. From this representation one can see, for example, that the orange and magenta clusters cover all the observations in the top one half (lexicographically) of the data. We will investigate these insights more thoroughly in the next section.

We have discussed the various visualization components used in the IHEDAT system. As is always the case in the construction of a visualization system, there are numerous other ways to display the information inherent in the CCCD system. The exact inner workings of the IHEDAT system is best illustrated via the analysis of an example data set.

5. Exploration of an Artificial Nose Data Set

IHEDAT was originally developed to study the relationship between an artificial nose data set and the CCCD classifier. The artificial nose data set consists of the response of a nonlinear artificial nose to a sequence of chemicals. These chemicals consist of a target compound (trichloroethylene (TCE)) at various concentrations along with various confusers (chloroform, hexane, Coleman fuel, etc.). A particular analyte consists of the target compound at a particular concentration as diluted with air and possible confuser. The analyte is drawn across the fibers of the system in a 20 second pulse. Each of the nose's 19 fiber responses is sampled 60 times and at two wavelengths during the 20 second interval. Hence, each system response consists of a point in a 2280 dimensional space. The reader is referred to Priebe[12] for a full discussion of the artificial nose data collection process.

The data set for our discussions consists of 80 target observations (consisting of TCE along with air and benzene at various concentrations) along with 40 nontarget observations (consisting of benezene in air). This data set was chosen in order to illustrate the IHEDAT capabilities while not overloading the reader.

Figure 3 presents a screen snapshot of the IHEDAT system. The upper left window presents a grayscale rendering of the target(T), nontarget(NT) versus T,NT interpoint distance matrix. The upper right window contains a parallel coordinates plot of the target observations rendered in the cluster map space. We point out that the observations have been colored to match the cluster map representated by the colored dendrogram residing in the lower left-hand corner of the frame. The lower right-hand window contains a file selection widget that allows one to choose the observations that reside in the target and nontarget classes. Now let's take a closer look at each of the windows in turn.

The lower left hand window contains a colored dendrogram that represents the cluster mapping. Each of the terminal leaves of the dendrogram corresponds to a

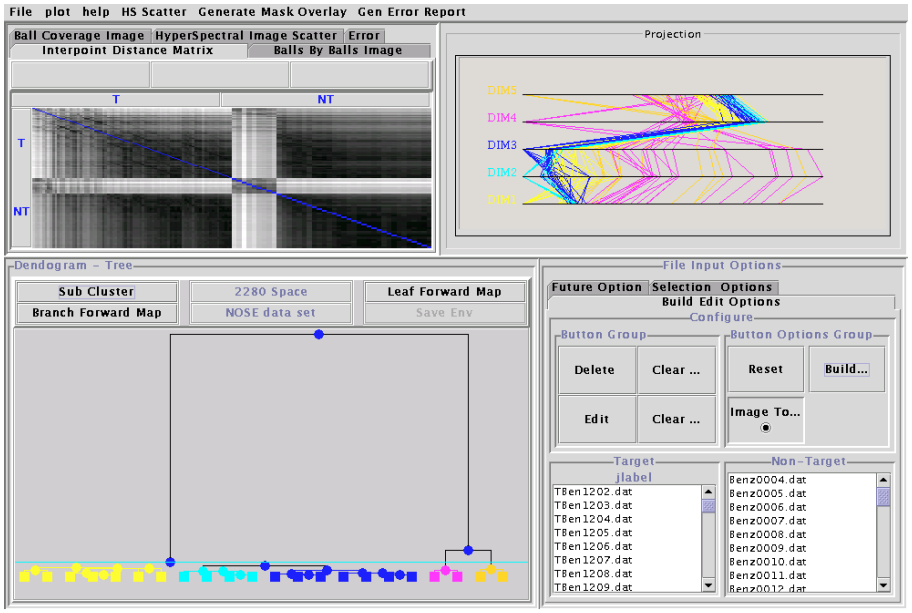


Fig. 3. IHEDAT interpoint distance matrix plot for the TCE and benzene versus benezene artificial nose data set.

dominating set element. We have cut the dendrogram at five clusters as indicated by the scale dimension procedure. The first seven dominating set elements colored yellow reside in one cluster. The next four elements, colored cyan, reside in the next cluster. Altogether the reader can discern five different clusters. We have provided the capability to interact with the dendrogram to ascertain the center, radius, and target containment for each of the dominating set elements. This capability will be discussed shortly.

Next, we turn our attention to the interpoint distance matrix displayed in the upper left-hand corner. The observations are ordered based on the selection made by the user with the data selection widget. The pattern of selected observations is as follows. We have ten replicates with target (trichloroethylene (TCE)) at a concentration of 1:1 and nontarget (benezene (BEZ)) at a concentration of 1:2, ten replicates with TCE at a concentration of 1:1 and BEZ at a concentration of 1:7, ten observations with TCE at a concentration of 1:2 and BEZ at a concentration of 1:2, ten replicates with TCE at a concentration of 1:2 and BEZ at a concentration of 1:7, ten replicates with TCE at a concentration of 1:7 and BEZ at a concentration of 1:2, ten replicates with TCE at a concentration of 1:7 and BEZ at a concentration of 1:7, ten replicates with TCE at a concentration of 1:10 and BEZ at a concentration of 1:2, and ten replicates with TCE at a concentration of 1:10 and BEZ at a concentration of 1:7. So the concentration of TCE in the target observations is nonincreasing as we proceed down the list. The nontarget observations consist of nine observations of satrated BEZ, two observations with BEZ

diluted in air at a concentration of 1:1, nine observations with BEZ diluted in air at a concentration of 1:2, ten observations with BEZ diluted in air at a concentration of 1:7, and ten observations with BEZ diluted in air at a concentration of 1:10. Examining the interpoint distance matrix for target versus target, we see that the high concentration target observations seem to be at the greatest distance from the low concentration target observations. This is indicated by the white block in the lower left-hand and upper-right hand portion of the target versus target section of the diagram. Continuing on with our analysis of the target versus target portion of the diagram, we notice that the low concentration target observations are close to the other low concentration target observations. This is indicated by the dark block in the lower right-hand portion of the target versus target portion of the interpoint distance image.

We next turn our attention to the target versus nontarget portion of the interpoint distance image. We first note that the high concentration target observations are at a large distance from the low concentration nontarget observations. This is evidenced by the white rectangle in the upper right-hand and lower left-hand corners of the interpoint distance image. We next note that the low concentration target observations are at a close distance to the low concentration nontarget observations. This is indicated by the dark rectangle midway down the interpoint distance matrix image on the right-hand side along with it symmetrical representation that appears in the lower left middle nontarget section.

Perhaps one of the most striking features of the interpoint distance image is the white plus sign running vertically in the nontarget portion of the image and horizontally slightly below the midpoint of the image. This visual feature is a clear indication that these sets of observations are outliers as compared to the rest of the target and nontarget observations. The nine observations resident in this plus sign are the nontarget obsrvations consisting of saturated benzene vapor. This type of insight is representative of some of the possibilities that can be obtained using IHEDAT.

Now consider the parallel coordinates plot in the upper right-hand section. Each target observation has been rendered in the parallel coordinates plot produced using the cluster map transformation. The purpose of this plot is to evaluate the observations for structure in the cluster map space. Each observation has been colored according to which dominating element contains the observation. There is a problem of overplotting here in that each target observation may be covered by more than one dominating set element.

Next, we wish to touch upon some of the other dendrogram-based operations that we have provided. We have previously mentioned the capability to cut the dendrogram to create a resultant cluster map. We also can click on a particular dominating element/terminal node and have the target observations that are covered by this node highlighted in the target selection widget in the lower right-hand corner and also in the cluster map parallel coordinates plot in the upper right-hand corner. The center and radii of the dominating set element are also indicated during this operation. Sometimes we are interested in knowing which target

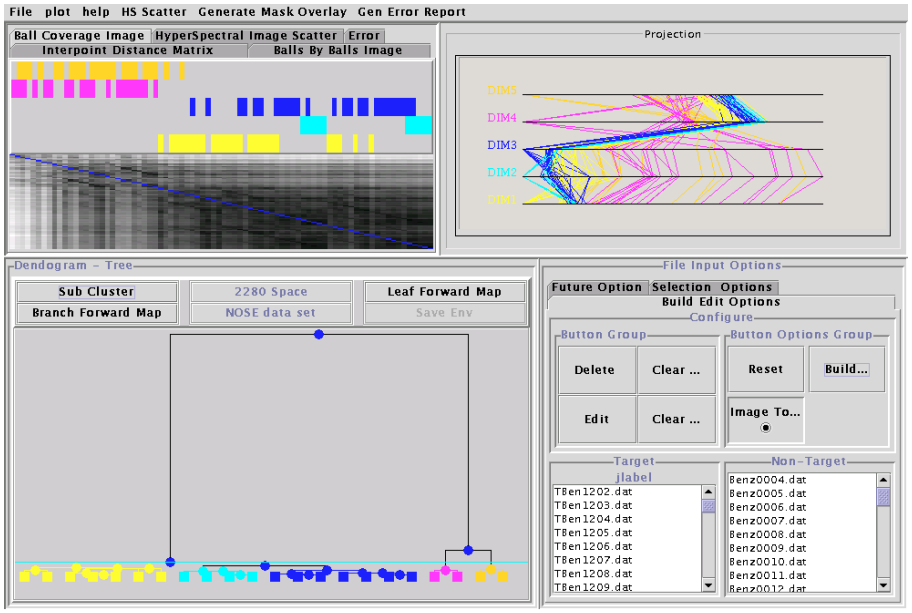


Fig. 4. IHEDAT ball coverage plot for the TCE and benzene versus benzene artifical nose data set.

observations are covered by a particular cluster of dominating set elements. This is accomplished by selecting the node in the dendrogram that sits right above a particular cluster of dominating set elements. This "branch forward" operation also highlights the appropriate observations in the target point selection widget and parallel coordinates plot.

Next, we draw the readers attention to Fig. 4. The only difference between Fig. 3 and Fig. 4 is that in Fig. 4 we have moved the "ball coverage" window in the upper left-hand portion of the frame to the front. The numerous windows in the upper left-hand corner reside in a set of overlaid "tab pane" windows. The bottom portion of this window contains the target versus target interpoint distance grayscale image. Above each column (observation) we have provided a series of colored pixels indicating the cluster containment relationships. For example, the left most set of target observations are only covered by elements of the orange and magenta dominating set clusters.

Once one has ascertained the representative radii associated with each of the dominating set elements, this information can be incorporated into the analysis of the ball coverage diagram. In the case being considered the orange and magenta clusters are associated with larger dominating set radii. Given our previous discussions on the order of data by concentration, this indicates that these elements cover the high concentration target observations. It is in keeping with our intuition that these high concentration target observations are farther to the nontarget observations than the low concentration target observations are to the nontarget

observations. This fact is also supported by our previous analysis of the {T,NT} versus {T,NT} interpoint distance matrix.

The use of the ball coverage window to aid in the analysis of the cluster map produced parallel coordinates plot is a bit more tenuous but we still feel that such an analysis can prove beneficial. The magenta and orange clusters cover the high concentration observations. These observations are far from one another in the original high-dimensional space. This is evidenced by the lightness of the interpoint distance matrix in this area. The target observations colored magenta and orange in the parallel coordinates plot seem to be more spread out than those observations that are colored yellow, green and blue. The reader is urged to compare their parallel coordinates structure with that of the observations colored yellow, cyan and blue that correspond to the lower concentration observations. The later travel much more tightly through the parallel coordinates plot.

The above arguments can be made more clear using a sequence of branch forward maps on the low and high concentration observations. In Fig. 5 we have performed branch forward operations on the orange and magenta balls. Now the observatiosn covered by the orange and magenta balls have been highlighted in red in the parallel coordinates plot. Similarly, one can chose to perform the branch forward operation at the branch point directly above the yellow, cyan and blue balls. Figure 6 presents the result of this operation. The closer spatial proximity of the lower concentration observations are clearly revealed in the parallel coordinates plot.

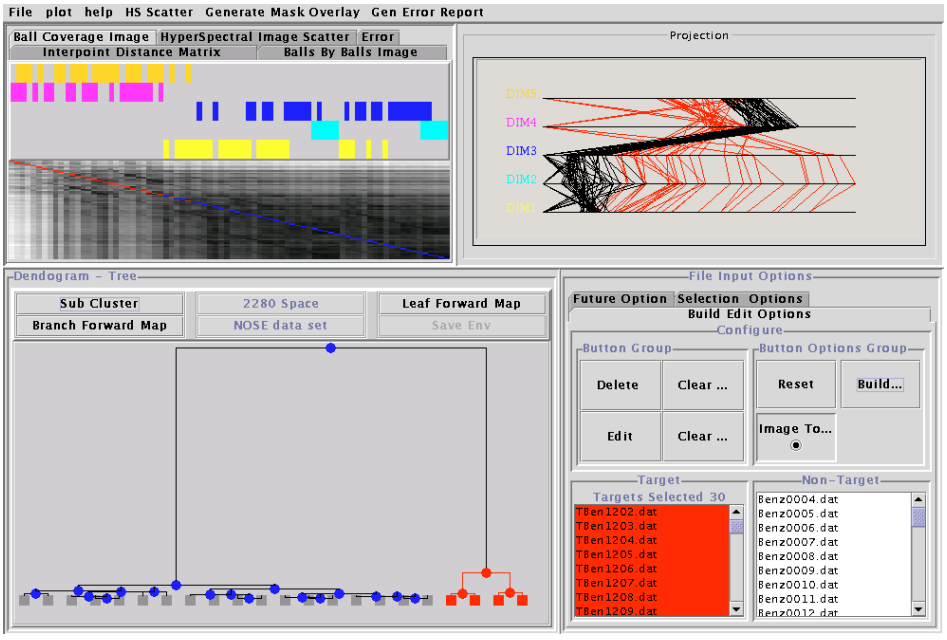


Fig. 5. IHEDAT forward mapping operation on the high concentration observations in the TCE and benzene versus benezene case.

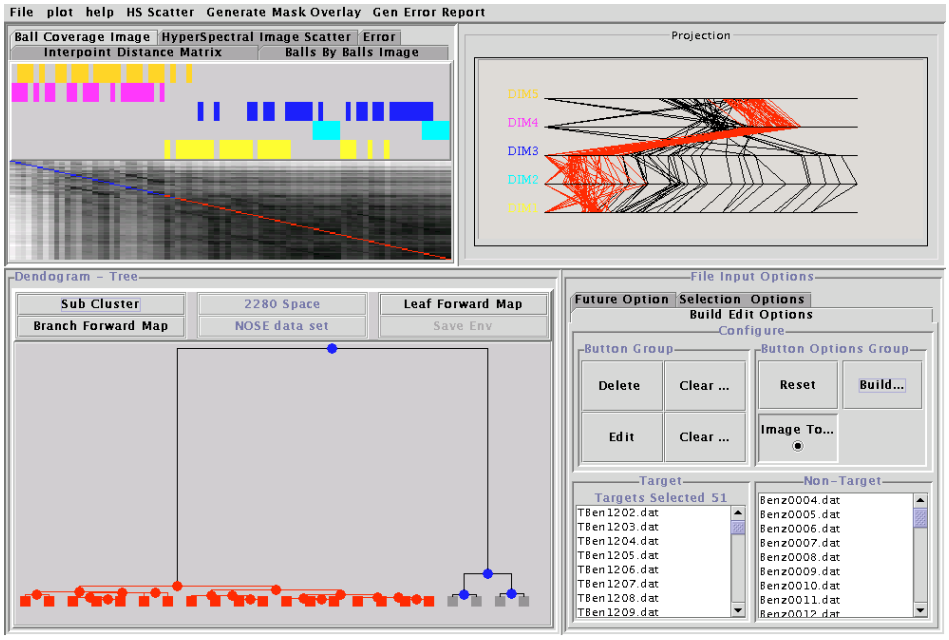


Fig. 6. IHEDAT forward mapping operation on the low concentration observations in the TCE and benzene versus benezene case.

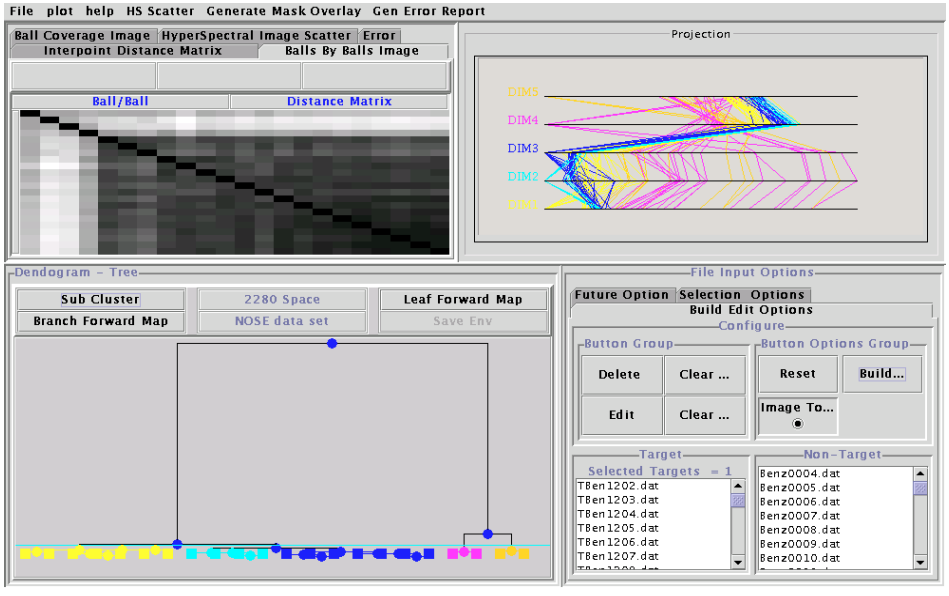


Fig. 7. IHEDAT ball versus ball interpoint distance matrix for the TCE and benzene versus benzene artificial nose data set.

We next turn our attention to an analysis of the interpoint distance image based on the dominating set elements, (see Fig. 7). We first note that the yellow dominating set elements that cover some of the lower concentration target observations seem to be the farthest to one another and the other balls. This is indicated by the light band in the left-hand corner of the the balls versus balls interpoint distance matrix. We next note that the spread of the remaining balls is fairly low.

We conclude our discussions in this section with a brief examination of two more nose test cases. In the first of these we have selected BTEX as the confuser. BTEX is a mixture of benzene, toluene, ethylbenzene and xylene. The concentrations for the TCE and BTEX in the 80 target observations are as follows, ten (1:1,1:2), ten (1:1,1:7), ten (1:2, 1:2), ten (1:2, 1:7), ten (1:7, 1:2), ten (1:7, 1:7), ten (1:10, 1:2), and ten (1:10, 1:7). The concentrations for the BTEX in the 40 nontarget observations are as follows, ten saturated, ten 1:2, ten 1:7, and ten 1:10. Figure 8 presents a ball coverage plot for this test case. In this the low concentration observations are covered by the yellow balls. These balls have consistently lower radii as compared to the cyan and magenta balls that cover the higher concentrations.

In our final example the confuser compound is carbon tetrachloride (CTET). The concentrations of TCE and CTET in the 80 target observations are as in the previous example. The concentrations of CTET in the 40 nontarget observations are also as in the previous example. Figure 9 presents a ball coverage plot for this test case. In this case the scale dimension of two indicates that the target observations

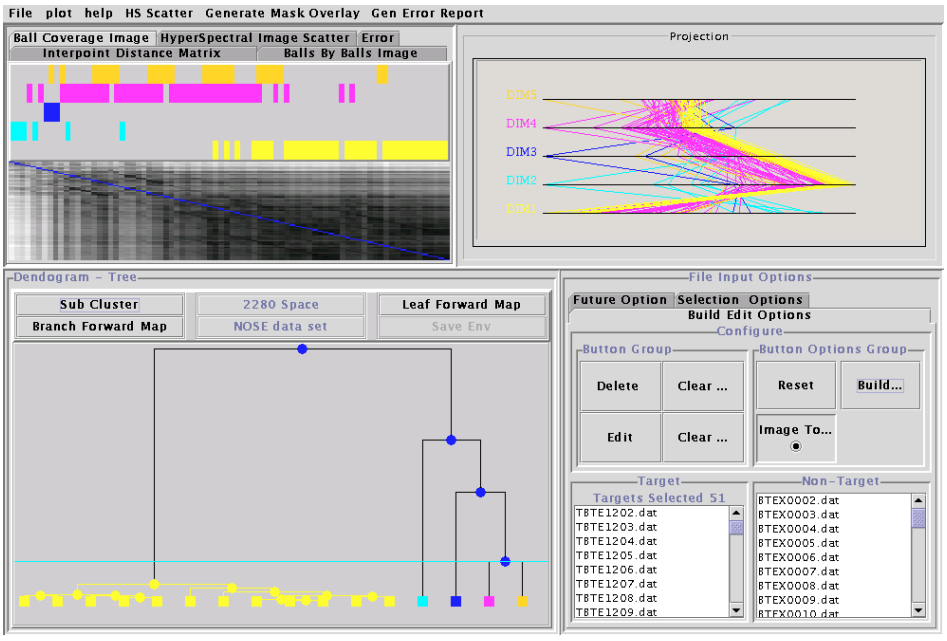


Fig. 8. IHEDAT ball coverage plot for the TCE and BTEX versus BTEX artifical nose data set.

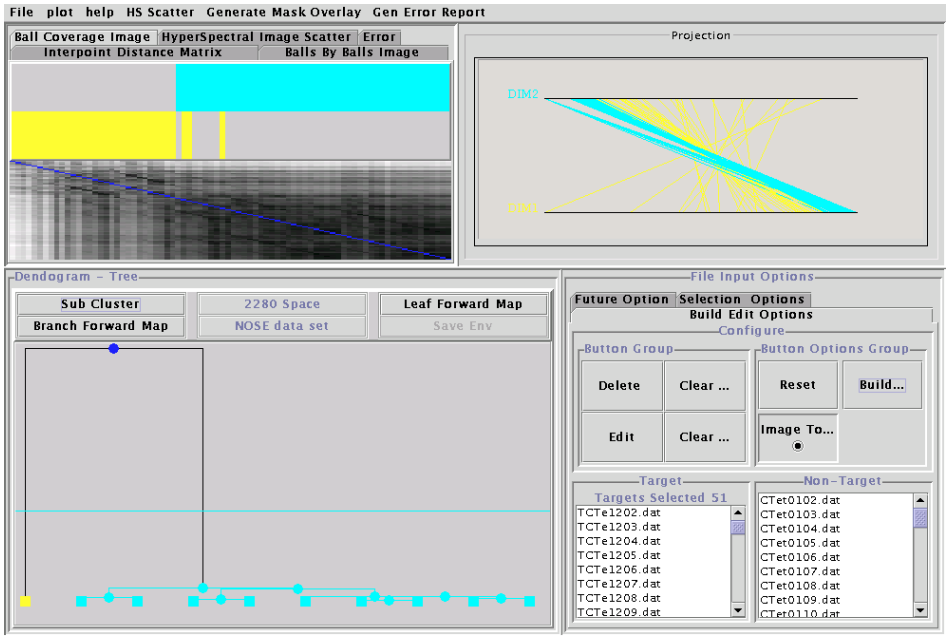


Fig. 9. IHEDAT ball coverage plot for the TCE and CTET versus CTET artificial nose data set.

can be covered by balls with two types of radii. The cyan balls over in the right-hand side of the dendrogram cover the lower concentration observations and have smaller radii than the yellow ball covering the higher concentration observations. In fact, the single yellow ball that covers all of the target observations with TCE at a concentration of 1:1 along with several of the observations with TCE at a concentration of 1:2 has a radius of around 4034. This is as compared to the radii of the cyan balls covering the lower concentrations all of which are at a radii less than 1000.

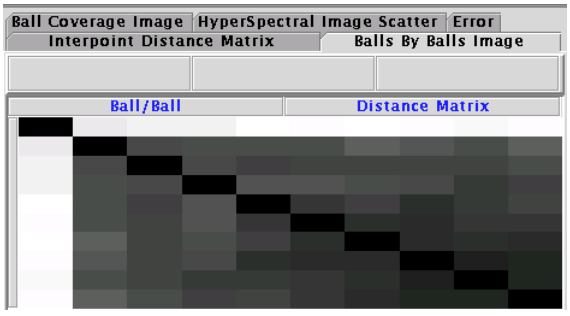


Fig. 10. IHEDAT ball versus ball interpoint distance matrix for the ball versus ball data.

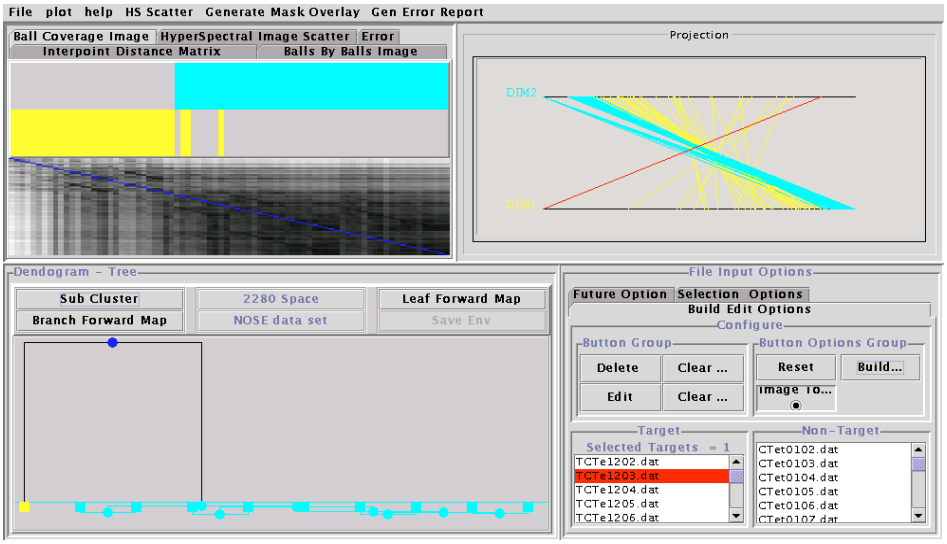


Fig. 11. IHEDAT parallel coordiantes plot for the TCE versus CTET case. High concentration ball center identified.

We may evaluate the location of the single ball that covers these high concentration observations through an examination of the ball versus ball interpoint distance matrix. In Fig. 10 we present the ball versus ball interpoint distance matrix for this data set. We notice that the first ball, the one covering the high concentration observations, is at a large distance from the other balls. This is indicated by the white bar on the left and top of the image. Using a left forward operation one can ascertain that this ball is centered at TCTE1203.dat. This is an observation with a TCE concentration of 1:1 and a CTET concentration of 1:2. A little exploratory data analysis allows one to identify this observation in the parallel coordinates plot, see Fig. 11.

6. Conclusions

We have presented a new framework for the analysis of hyperdimensional data using class cover catch diagraphs. This method, IHEDAT, has been discussed via the analysis of an artificial olfactory data set. IHEDAT is based on several different visualization components, some of which are well known to the statistical community and some of which are novel creations. IHEDAT allows the user to formulate a hypothesis about the geometric relationship between a group of target and nontarget observations.

We have numerous ideas for continuing our work on the IHEDAT. Some of these include provisions for the display of the underlying class cover catch digraph. Some of these focus on the effect of metric space choice on the classification process.

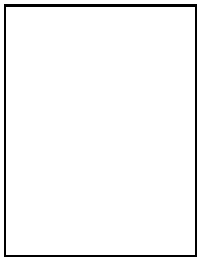
We look forward to new research paths that might be provided by the readers of this article.

Acknowledgments

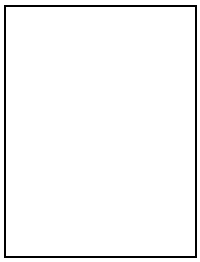
The work of C. E. Priebe was partially supported by Office of Naval Research Grant N00014-01-1-0011 and DARPA Grant F49620-01-1-0395. The work of J. L. Solka and B. T. Clark was partially supported by the Office of Naval Research through the NSWCDD In-house Laboratory Independent Research Program. This work was performed while C. E. Priebe was ASEE/ONR Sabbatical Leave Fellow 2000–2001 (N00014-97-C-0171 and N00014-97-1-1055) at NSWCDD. The authors would like to thank David Marchette for his insightful comments and for his help generating Fig. 1.

References

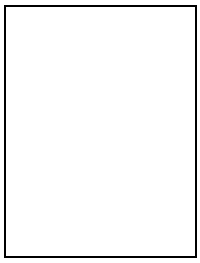
1. R. E. Bellman, *Adaptive Control Processes* (Princeton University Press, Princeton, 1961).
2. D. Scott, *Multivariate Density Estimation* (John Wiley and Sons, New York, 1992).
3. C. E. Priebe and D. J. Marchette, "Characterizing the complexity of a high-dimensional classification problem," *Computing Science and Statistics* **32** (2000).
4. C. E. Priebe, D. J. Marchette, *et al.*, "Classification via class cover catch digraphs," *JHUDMS TR*, in preparation (2002).
5. C. E. Priebe, J. L. Solka, D. J. Marchette, and B. T. Clark, "Class cover catch digraphs for latent class discovery in gene expression monitoring by DNA microarrays," submitted to *Computational Statistics and Data Analysis (Special Issue on Data Visualization)*, 2001.
6. C. E. Priebe, J. G. DeVinney, and D. J. Marchette, "On the distribution of the domination number of random class cover catch digraphs," *Statistics and Probability Letters*, to appear.
7. B. Everitt, *Cluster Analysis* (2nd Edition, Halsted, New York, 1990).
8. M. Minnotte and W. West, "The data image a tool for exploring high-dimensional data sets," in *1998 Proc. ASA Section on Statistical Graphics* (1999).
9. D. Marchette and J. Solka, "Using data images for outlier detection," submitted to *Computational Statistics and Data Analysis (Special Issue on Data Visualization)* 2001.
10. E. Wegman, "Hyperdimensional data analysis using parallel coordinates," *JASA* **85**, 664–675 (1990).
11. A. Inselburg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multi-dimensional geometry," in *Proc. First IEEE Conf. Visualization* (1990), pp. 361.
12. C. Priebe "Olfactory classification via interpoint distance analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence* **23**(4), 404–413 (2001).



Jeff Solka was born in Harrisonburg, Virginia, on January 31, 1955. He earned his B.S. degree in Mathematics and Chemistry from James Madison University in 1978, his M.S. in Mathematics from James Madison University in 1981, his M.S. in Physics from Virginia Polytechnic Institute and State University in 1989 and his Ph.D. in Computational Sciences and Informatics (Computational Statistics) at George Mason University, working under the direction of Prof. Edward J. Wegman, in May of 1995. Since 1984, Dr. Solka has been working in nonparametric estimation and statistical pattern recognition for the Naval Surface Warfare Center, Dahlgren, VA.



Carey Priebe received his B.S. degree in mathematics from Purdue University in 1984, his M.S. degree in computer science from San Diego State University in 1988, and his 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, Maryland. His research interests are in computational statistics, kernel and mixture estimates, statistical pattern recognition, and statistical image analysis.



Ted Clark completed his B.S. in Computer Science at the University of South Carolina in 1988. He has been employed at the Naval Surface Warfare Center since 1988. His research interests include visualization and simulation. He enjoys developing visualization and simulation systems in OpenGL and JAVA.