4

Data Mining Strategies for the Detection of Chemical Warfare Agents

Jeffrey. L. Solka[1,2], Edward J. Wegman[1], and David J. Marchette[2]

[2]Naval Surface Warfare Center (NSWCDD), Dahlgren, VA, [1]George Mason University, Fairfax, VA, USA

CONTENTS

This paper discusses a classification system for the detection of various chemical warfare agents. The data were collected as part of the Shipboard Automatic Liquid (Chemical) Agent Detection (SALAD) system. This system is designed to detect chemical agents onboard naval vessels. We explore the intricacies associated with the construction of various classification systems. Along the way we take time to explore some applications of recently developed statistical procedures in visualization and density estimation to this discriminant analysis problem. We focus our discussion on all phases of the discriminant analysis problem. In the exploratory data analysis phase we provide results that detail the use of histograms, scatter plots and parallel coordinate plots for the selection of feature subsets that are fortuitous to the discriminant analysis problem and the discernment of high dimensional data structure. In the discriminant analysis phase we discuss several semiparametric density estimation procedures along with classical kernel, classification and regression trees, and k-nearest-neighbors based approaches. These discussions include some illustrations of the use of a new parallel coordinates framework for the visualization of high dimensional mixture models. We close our discussions with a comparison of the performance of the various techniques through a study of the associated confusion matrices.

4.1 Introduction

The Shipboard Automatic Liquid (Chemical) Agent Detector (SALAD) is a system designed to detect chemical agents onboard naval vessels. The device takes the form

of an instrument that is fed with tractor feed reagent paper. This paper reacts with chemical droplets to produce a characteristic color change. The device is designed to sit exposed on the ship waiting for chemical agents to rain down on the reactive paper as it travels through the system. At certain periodic intervals a camera captures images of the paper using 13 spectral filters. Intensity measurements at each of these wavelengths are collected and passed to the classification section of the system for additional processing. In this phase of the process the agents are to be classified according to particular chemical type. It is the classification portion of the system that our work has focused on.

Initially data was collected at the Dahlgren Division of the Naval Surface Warfare Center on simulant chemicals, which are designed to produce paper signatures similar to the actual chemical warfare agents. Although these data were provided to us and we did perform some preliminary analysis on the data, this is not the focus of this paper. In addition, data was collected at the GEOMET Center on several of the live agents at various drop sizes. Creation of a classification system for the signatures of the 1 ml drops was the goal of our analysis. Thirteen band signatures were collected on the chemicals GA, GB, GD, GF, VX, HD, L, GDT, HDT, and the paper without any chemical stimulant.

The collection data was presented to us initially as a set of images. Using the Advanced Distributed Region of Interest Tool (ADROIT) [3] the images were diagnosed. The diagnosis procedure consisted of labeling those pixels from each of the various chemical classes along with a subset of the pixels from the background. In this manner a training set and a test set were created. The training set was used to build the classifier and the test set was used to test it. More sophisticated testing procedures, such as the jackknife, were considered but were deemed unnecessary. The training data consisted of 14,236 observations and the test data consisted of 1,868,070 observations.

For the purposes of our analysis we grouped GA, GB, GD, GF, and GDT into the class G; VX into the class V; and HD, L, and HDT into the class H. Each class was then assigned a numerical class label according to the following scheme. G was labeled as class 0, V was labeled as class 1, H was labeled as class 2, and the background was labeled as class 3. The training set consisted of 2,106 observations from the G class, 569 from the V class, 1,088 from the H class, and 10,473 background observations. The test set was comprised of 13,889 observations from the G class, 2,318 from the V class, 6,662 from the H class, and 1,845,201 from the background class.

The design of the classification system was broken into the usual constituent steps of exploratory data analysis, probability density estimation, classifier design, and classifier testing. The purpose of the exploratory analysis phase is to ascertain any underlying structures that exist in the training data. This is in part done in order to choose particularly fortuitous features sets and also to discover any additional structure that would be important in the density estimation portion of the procedure. Standard statistical visualization procedures such as boxplots and histograms are typically applied to univariate projections of the features. In addition one sometimes examines pairs plots, which represent scatterplots of the various features chosen 2 at a time.

Ultimately one would like to be able to examine the distribution of the features in the full higher-dimensional space. The parallel coordinates method (cf. Wegman [11]) is one technique to do this. In this technique one places the coordinate axes parallel to one another in order to plot the points in the higher-dimensional space. This is necessary since one, in dimension higher than 3, ultimately does not have the ability to place the coordinate axes orthogonal to one another. It turns out that there is a natural correspondence between this coordinate system and projective geometry space. In this manner, we can understand how certain geometric structures in Euclidean space are mapped into the parallel coordinates framework.

Once one has performed a preliminary analysis based on these techniques one performs model-based exploratory analysis on the feature set. There are several different techniques for performing model-based density estimation. These range from the fully nonparametric procedures such as kernel density estimation [8], to semiparamteric density estimation procedures such as the adaptive mixtures density estimator (AMDE) [6] [5], and finally fully parametric models such as finite mixture models [4].

In the kernel estimation approach, one models the underlying distribution as a mixture of component densities. Each component density resides at one of the points of the data set and often takes the simple form of a Gaussian. In equation 1 we present the form of the univariate kernel estimator

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} g\left(\frac{x - x_i}{h}\right). \tag{4.1}$$

In the case of multivariate data it is simplest to use a the well-known product kernel.

Mixture models are an alternative to the fully nonparametric kernel estimator. In the simplest case one can model each of the class densities as a single term mixture that has a mean based on each class mean and a common covariance structure. In this case the common covariance matrix is given by

$$\Sigma_c = \sum_{allc} \frac{n_c}{n} \left(\frac{W_c}{n_c - 1}\right) \tag{4.2}$$

where

$$W_c = \sum_{i=1}^{n_c} (x_i^c - \mu^c)^2. \tag{4.3}$$

This classifier is denoted as a linear classifier. One obtains a quadratic classifier by allowing each of the classes to have a different covariance structure.

In the adaptive mixtures density estimation (AMDE) [5] [6] approach one loosens several of the requirements of the above procedures. Namely one allows the number of terms in the model to be driven by the complexity of the data, the location of the terms to be anywhere, the covariances to be nonuniform, and the mixing coefficients to not all be equal. In this case the form of the estimator is

$$\hat{f}(x) = \sum_{i=1}^{m} \pi_i N(x, \hat{\theta}). \tag{4.4}$$

In this equation it is the number of terms that is determined by the complexity of the data, the π's are the mixing coefficients, and N is the normal density determined by parameter set. In this case, the estimator is built in a recursive manner. As each data point is presented, the estimator either updates the existing parameters in the model using a recursive form of the expectation-maximization (EM) algorithm or else adds an additional term to the model as dictated by the complexity of the data.

The last mixture-based approach to be discussed is known as the Shifted Hats Iterated Procedure (SHIP). This method is a hybrid method that employs both kernel estimation techniques and mixture models. The technique switches attention between a mixture model of the data set and a kernel estimate (or more sophisticated semiparametric estimator). The name shifted hats came about since an estimator is typically denoted by the hat symbol and we are shifting our view on which function the estimator represents. A full description of this technique is provided in [7].

The next approach that was used to classify the data is based on the k-nearest-neighbors procedure. In this procedure one, assigns a class label for each of the observations in the test set based on the k closest elements of the training set under an appropriate distance metric such as the standard Euclidean metric. In our case the large cardinality of the training set makes a straight-forward application of the procedure problematic. We have chosen to use a reduced k-nearest-neighbors method as described by Hand 1997 [1]. This approach selects a subset of the full exemplar set for use in a nearest neighbor classifier. The reader is referred to Hand 1997 [1] for a full treatment of the methodology.

The final approach that was used to classify the data was classification and regression trees (CART). In this approach the algorithm attempts to form a sequence of decision planes perpendicular to the coordinate axes that partition the data into class homogeneous regions. The classifier then takes the form of a sequence of simple if tests. The reader is referred to Venables and Ripley 1994 [10] for a full treatment of this approach.

4.2 Results

Initially we chose to evaluate the classification utility of the features individually. In Figure 4.1 we present histogram plots for each of the 13 bands and each of the classes. We notice a fair degree of separability exhibited in features 7 and 11. This separability led us to utilize these features as one possible choice in subsequent analysis.

Next we turn our attention to an examination of the full thirteen-dimensional features. We have chosen to visualize the full higher dimensional feature set through the use of parallel coordinates, which provides us with a convenient venue for the display of higher dimensional data. In Figure 4.2 we present a parallel coordinate plot of the training data. The plot was produced using the ExplorN package that

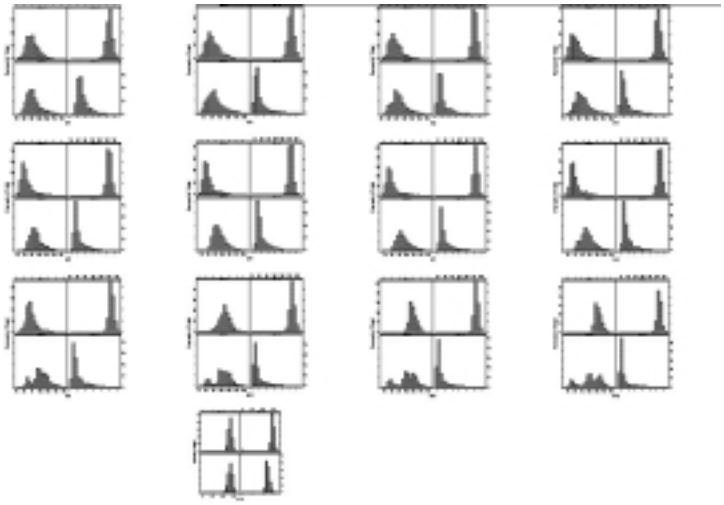

FIGURE 4.1
Histogram plots for each of the 13 bands for each of the 4 classes. The plots are arranged by band from top to bottom. In each case class 0 appears in the lower left corner, class 1 appears in the lower right corner, class 2 appears in the upper left corner, and class 3 appears in the upper right corner.

utilizes saturation brushing, a technique to deal with the overplotting problem associated with large data sets [12] [2].In the plot the color saturation is computed as a function of the number of lines that traverse a given area. In addition when colored lines overlap one another the package mixes the colors together in the usual additive manner. Class 0 is displayed as red, class 1 as green, class 2 as blue, and class 3 (background) as white. The lowest axis in the figure designates the class label.

There are a few things worthy of note in the plot. First we notice the characteristic scalloped appearance of the class 3 data. This visual feature is associated with a multivariate elliptical density as might be found in a normal data set. It is not surprising to find that the background intensities were normally distributed. We also notice that a small subset of the class 3 observations are outliers as indicated by their far left appearance in band 13. This apparent anomaly is the subject of continued investigation. We also note the amount of class separation in bands 7 and 11. This observation is in keeping with our univariate analysis. Once again we notice that the class 3 observations are well separated from the other classes in most of the bands. We finally note the trimodality of class 0 particularly in band 11. This multimodal type behavior is not surprising since we originally collapsed multiple chemical classes into each of the subsequent classes.

The collectors of the SALAD data also proposed an alternate trivariate feature set. This feature set was chosen to mimic a three-band red, blue, green combination. Adding bands 1 and 2 together formed the first feature, the second feature by adding

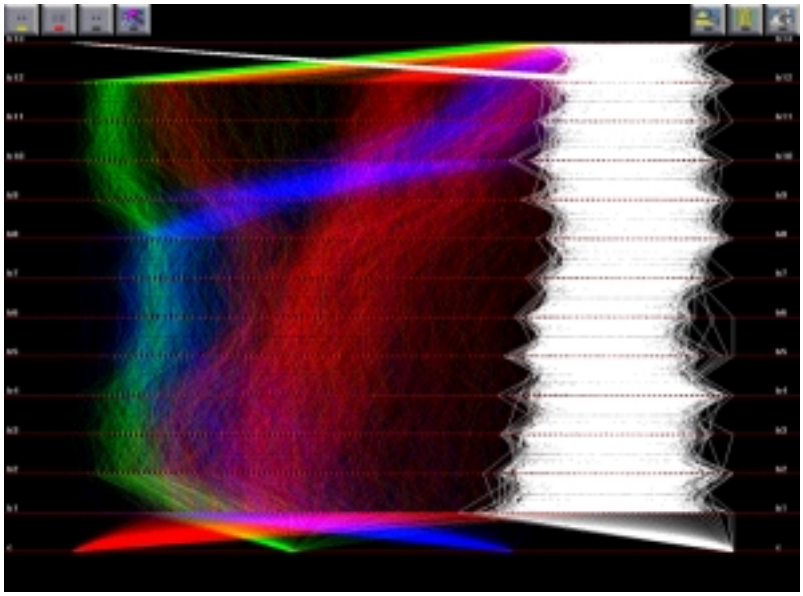

FIGURE 4.2
Parallel coordinates plot of all the training data. Class 0 is rendered in red, class 1 in green, class 2 in blue, and class 3 in white.

bands 6 and 7 and the third feature by adding bands 11 and 12. Besides reducing the dimensionality of the problem to three-space this step also reduced the requirements on the system so that instead of collecting 13 spectral bands they only need to collect 6 bands, or with a modification of the filter set, 3 bands. In Figure 4.3 we present a scatterplot of these features for the 4 classes. The color scheme is identical to the previous plots with the exception that the background observations have been rendered in black. This plot provides additional evidence for the normality of the class three observations. In addition we see clear separation between the background observations and the other classes. Finally we notice that classes 0 through 2 are also moderately well separated in this feature space.

We now turn our attention to some model-based exploratory data analysis. This provides an alternate means to evaluate the utility of the various features. In Figure 4.4 we present plots of SHIP-based probability density estimates for the various classes based on bands 7 and 11. These models help us evaluate the overlap between the various classes. We point out the trimodality of class 0 in both of the features. We note the separation between class 1 and class 2 in the band 11 feature and the separation of class 0 from both class 1 and 2 in the band 7 feature. Finally we note that the background is fairly well separated from the other classes in both bands.

Alternatively one may build bivariate densities for bands 7 and 11 together. In Figure 4.5 we present bivariate kernel density estimates of the training data using a spherical product kernel. In the right-most figure we color each pixel in the band 7

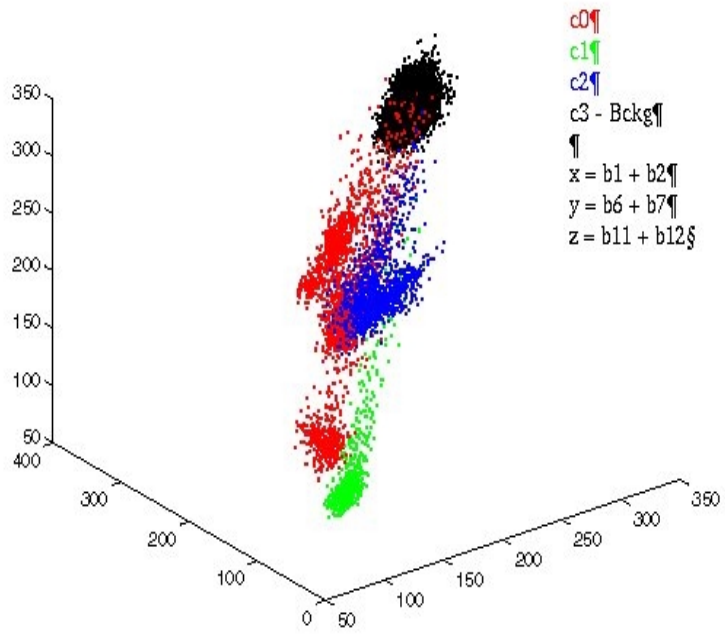

FIGURE 4.3
Scatterplot of the pseudo-RGB features. Class 0 is red, class 1 is green, class 2 is blue, and class 3 is black. Feature 1 (*x*-axis) is band 1 plus band 2, Feature 2 (*y*-axis) is band 6 plus band 7, and Feature 3 (*z*-axis) is band 11 plus band 12.

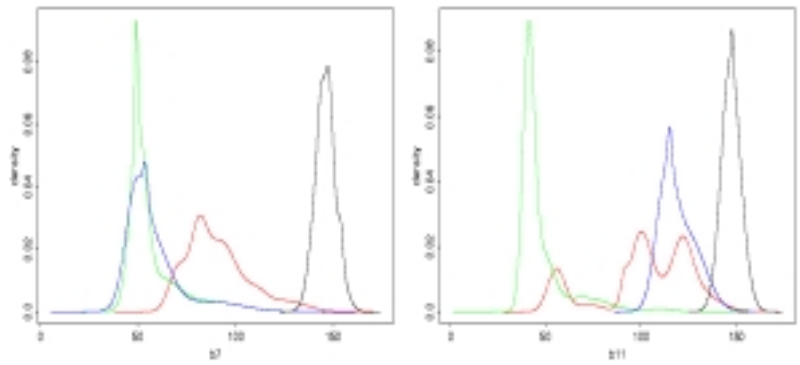

FIGURE 4.4
Univariate SHIP probability density functions for bands 7 and 11.

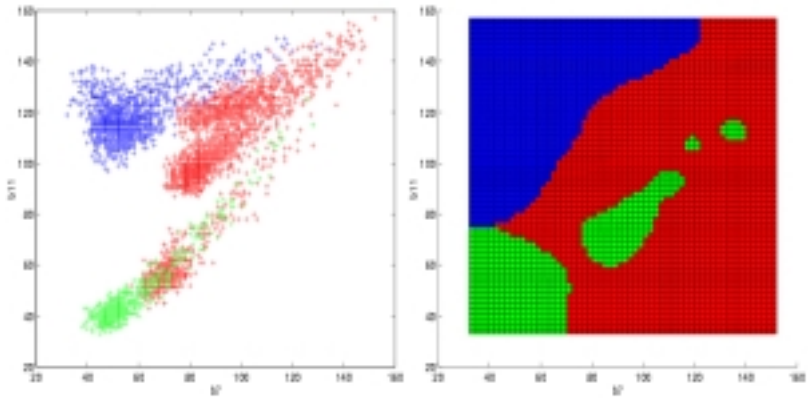

FIGURE 4.5
Multi-class discriminant regions based on a bivariate product kernel. Discriminant regions are plotted on the right and a scatterplot is rendered on the left.

cross band 11 space according to which of the class conditional bivariate probability density functions is higher. In the upper plot we present a scatterplot of the data. As before class 0 is colored in red, class 1 in green, and class 2 in blue. The background class has been omitted in this particular illustration. Pixels where the class conditional probabilities fell below a threshold have been colored white.

In the next set of images we consider the visualization of AMDE models based on the thirteen-dimensional training data. Each model consists of a mixture of thirteen-dimensional Gaussian terms. It is difficult in general to ascertain the match between the training data and the model given the high-dimensional nature of the data. In Figure 4.6 we plot the AMDE model for class 0. The data are rendered in yellow in the plot. In addition we have plotted the means of the terms that constitute the mixture model in red. The first axis has been used to plot a value equal to the scaled mixing coefficients in the case of the mixture terms and a dummy variable in the case of the data. The rendered grayscale images at the bottom of the plot represent the covariance structure of the terms in the mixture with the mixing coefficients explicitly spelled out below the images. White represents a large value in the covariance matrix.

There are a few relationships between the data and the model that are made clear by this plot. We notice that the term with the largest mixing coefficient tracks right through the center of the data set. The covariance image tracks the variability of the data fairly well. The next term, descending by mixing proportion, has a much tighter covariance structure as is indicated by the darkness of the rendering. This term tracks the left-most mode of the data set. The last two terms have very small proportions .

Next we turn our attention to some general discussions on the CART models that were built on the training data. In Figure 4.7 we present the plot of a classification tree based on the training data. The reader will notice that the CART procedure did not utilize all of the 13 bands but merely a subset of them. Specifically bands 1, 5, 7,

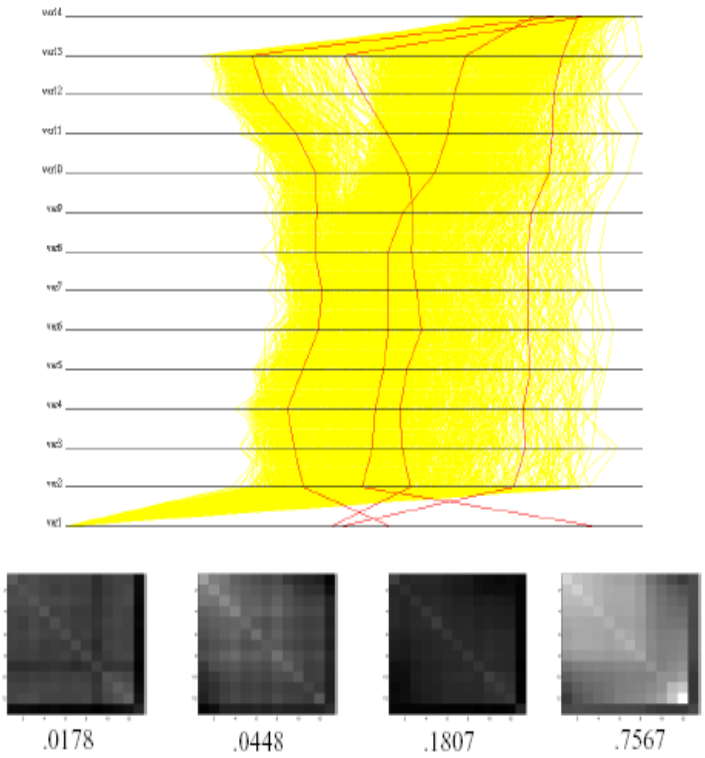

FIGURE 4.6
Parallel coordinates plot of the data, rendered in yellow, and the means of the mixture terms, rendered in red for class 0. The first axis represents a value proportional to the mixing coefficient in the case of the mixture terms. The gray-scale images represent the covariance structure of the given term. The numerical values of the mixing coefficients appear below the images.

8, 10, 12, and 13 were used in the model. These exceed the number of bands "hand picked" by the data collectors by 1, but interestingly enough the model has been built using an appreciably different set of bands.

It is easier to understand the inner workings of the CART procedure if we examine a pedagogical example. Suppose that the spectral signature of an observation is given by (70, 35, 110, 131, 111, 27, 105, 75, 215, 107, 115, 62, 117). The CART processes this observations as follows:

$$b7 = 105 < 129.5, \text{ which implies go left,}$$
$$b8 = 75 > 63.5, \text{ which implies go right,}$$
$$b12 = 62 < 111.5, \text{ which implies go left,}$$
$$b12 = 62 < 88.5, \text{ which implies go left, and}$$
$$b1 = 70$$

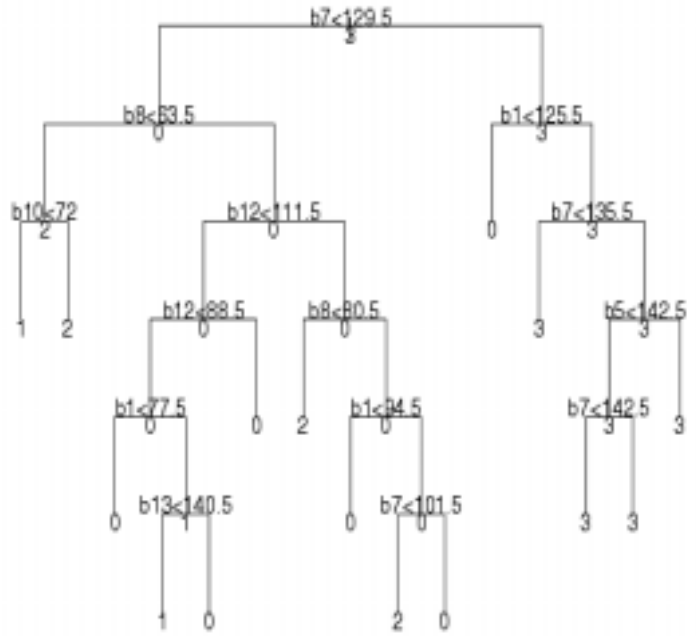

FIGURE 4.7
CART based on all 13 bands.

which implies go left which designates the observation as class 0.

Next we turn out attention to classification results obtained using the models discussed earlier. In each case we analyze the results via a confusion matrix. The first entry of the confusion matrix represents the probability of calling an observation as class 0 given that the observation came from class 0, denoted $p(c_0|c_0)$. The entry in the first row and second column is the probability of calling an observation class 0 given that it was drawn from class 1 denoted $p(c_0|c_1)$. Similarly the entry in the second row and the first column represents $p(c_1|c_0)$. So the diagonal entries represent the $p(c_i|c_i)$ for $i = 0,1,2$, and 3. We have computed confusion matrix results for the adaptive mixtures model based on the full 13 features, adaptive mixtures model based on bands 7 and 11, adaptive mixtures models based on the pseudo RGB features, the linear classifier based on all 13 features, the linear classifier based on bands 7 and 11, the linear classifier based on the RGB features, the quadratic classifier based on all 13 features, the quadratic classifier based on bands 7 and 11, the quadratic classifier based on the RGB features, the spatial CART classifier based on all 13 features, the spatial CART classifier based on RGB features, the knn classifier based on all 13 features, the reduced knn classifier based on 200 exemplars, and the

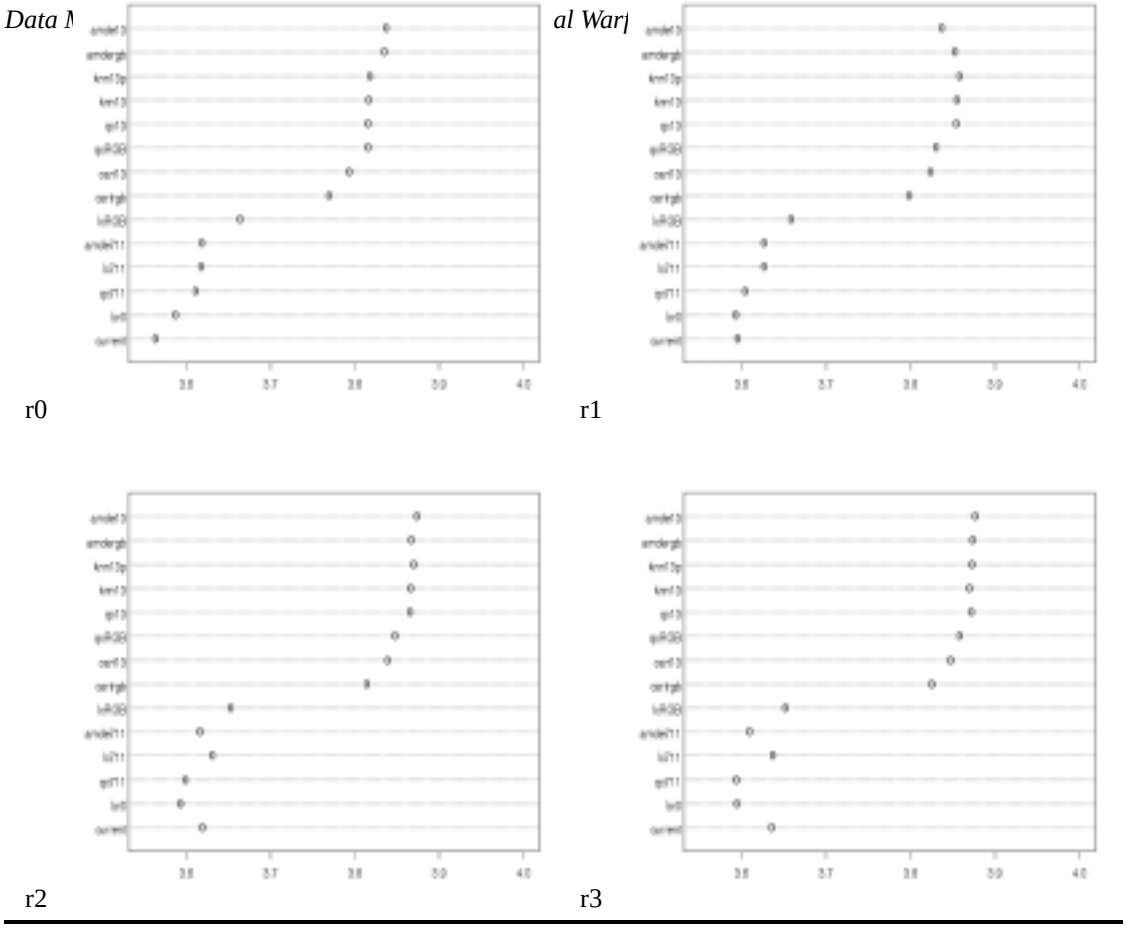

FIGURE 4.8
Sorted score for the various classification systems where the pixel radius varies between 0 and 3.

current classification system. We are however not free to provide the details of the current classification system at this time. The reader is referred to [9] for a full listing of the confusion tables.

The $r = 0$ entry in each Table treats the pixels as independent, ignoring the spatial information inherent in the original image. Since processing time is at a premium in this application, we considered the simplest method for utilizing the spatial relationship of the pixels. Initially each pixel is assigned a class as above. Then to determine the final class label for the pixel a vote is taken from all the pixels within a $(2r+1)$x$(2r+1)$ box centered at the pixel, with the final class label for the pixel being the one, which wins the vote (ties are broken arbitrarily). We refer to r as the radius. The radii used were 0-3, where a radius of 0 corresponds to the standard classifier with no spatial information.

In Figure 4.8 we present a dot chart of the results. The diagonal of the confusion matrix is summed and plotted on the x-axis. The classifiers were sorted by their performance under this metric in the case where the radius was 0 (single pixel, no spatial information), and this ordering is retained for the rest of the charts. There are several trends that are revealed in the dot chart. We first notice that those classifiers that use the full dimensionality of the data either as all 13 features or as the RGB feature set outperform the other classifiers that use the hand-selected two-dimensional projection. This occurs without fail except in the case of the simple linear classifier on all 13 features. Another thing to notice is that the semiparametric and the nonparametric classifiers outperform the parametric classifiers in general. By this we mean that the adaptive mixtures and knn-based classifiers in general perform better than the linear and quadratic classifiers. We also point out that the CART-based classifiers seem to be outperformed by the quadratic classifiers, but are better than the linear classifiers. We finally note that there is an improvement in performance as we proceed from a radius of 0 to 1 and finally 2. There is however, a leveling off of the improvement as we reach $r = 3$.

There are a few things that remain to be noted about the current approach. The current approach is bested by virtually all of the approaches at the $r = 0$ level. By the time one proceeds to the $r = 3$ level the performance of the fielded approach has improved sufficiently to allow it to outperform roughly three other classifiers. Even given this improvement the performance of the fielded system can be described as mediocre at best. This performance however may be sufficient depending on the situation at hand. This lack-luster performance is a trade-off for a need to rapidly field the system in order to be prepared for a very real threat.

4.3 Conclusions

We have attempted to evaluate the discriminant utility of these features. Our work has consisted of a data-mining phase, a model-building phases, and a model-evaluation phase. We have utilized standard statistical procedures such as histograms to provide univariate views of the data. In addition we have employed the parallel coordinates visualization framework in order to ascertain the structure of the feature set in the full thirteen-dimensional space.

We have employed high-performance probability density estimation procedures to model the distribution of the features sets in both the full space and other fortuitously reduced spaces. The density estimation/classification techniques used have included standard classification and regression trees along with adaptive mixtures, kernel estimators, k-nearest-neighbors and the recently developed shifted hats iterated procedure.

We have employed a simple scheme to incorporate spatial information into our classifier systems. We have measured the performance of the various classifiers using

the standard confusion matrix measure. In order to compare the performance of the classifiers we must turn to the confusion matrix results. As presented in the dotchart of Figure 8 we were able to show that the adaptive mixtures model obtained using all 13 features and a spatial radius of 3 proves to be the superior choice.

Assuming that one is very much limited with regard to both time and computational capabilities, then one needs to examine alternate solutions. Under these circumstances we recommend that one employ the CART model based on the full feature set with a spatial radius of 3. This system provides probability of detection that exceeds .85 while obtaining a false alarm rate less than .12. This system provides this level of performance while at the same time offering considerable speed improvements. In fact we would anticipate considerable time savings given the fact that the classifier takes the form of a simple sequence of if tests.

Acknowledgments

The authors would like to thank Greg Johnson of the Naval Surface Warfare Center for providing us with an opportunity to work on a very interesting problem. In addition we would like to thank Dr. Webster West of South Carolina University for provision of a JAVA-based parallel coordinates framework, and Dr. George Rogers of the Naval Surface Warfare Center for provision of the SHIP source code. This paper is an expansion of a part of the Keynote Lecture presented by EJW at the C. Warren Neel Conference on the New Frontiers of Statistical Data Mining, Knowledge Discovery, and E-Business. Other portions of that Keynote presentation may be found in [2].

The first author, JLS, would like to acknowledge the sponsorship of Dr. Wendy Martinez at the Office of Naval Research. The work of the second author, EJW, was completed under the sponsorship of the Air Force Office of Scientific Research under the contract F49620-01-1-0274 and the Defense Advanced Research Projects Agency through cooperative agreement 8105-48267 with Johns Hopkins University. The third author, DJM, would like to acknowledge the sponsorship of the Defense Advanced Research Projects Agency.

Reference

[1] D. J. Hand. *Construction and Assessment of Classification Rules*. John Wiley and Sons, New York, 1997.

[2] E. J.Wegman. Visual data mining. In *Statistics and Medicine*. 2002.

[3] D. J. Marchette, J. L. Solka, R. J. Guidry, and J. E. Green. The advanced distributed region of interest tool. *to appear Pattern Recognition*, 1998.

[4] G. J. McLachlan and K. E. Basford. *Mixture Models*. Marcel Dekker, New

York, 1988.

[5] C. E. Priebe. Adaptive mixtures. *J. Amer. Statist. Assoc.*, 89:796–806, 1994.

[6] C. E Priebe and D. J. Marchette. Adaptive mixture density estimation. *Pattern Recognition*, 26(5):771–785, 1993.

[7] G. W. Rogers, D. J. Marchette, and C. E. Priebe. A procedure for model complexity selection in semiparametric mixture model density estimation. *appearing in the Proceedings of and Presented at the 10th International. Conference on Mathematical and Computer Modelling and Scientific Computing*.

[8] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[9] J. L. Solka, E. J. Wegman, and D. J. Marchette. Data mining strategies for the detection of chemical warfare agents. Technical Report 182, George Mason University, 2002.

[10] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. Springer Verlag, 1994.

[11] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *J. Amer. Stat. Association*, 85:664–675, 1990.

[12] E. J. Wegman and Q. Q. Luo. High dimensional clustering using parallel coordinates and the grand tour, 1997. republished in Classification and Knowledge Organization, R. Kaluer and O. Opitz, eds.