Detection of Buried Targets Via Active Selection of Labeled Data: Application to Sensing Subsurface UXO

Yan Zhang, Xuejun Liao, *Senior Member, IEEE*, and Lawrence Carin, *Fellow, IEEE*

***Abstract*—When sensing subsurface targets, such as landmines and unexploded ordnance (UXO), the target signatures are typically a strong function of environmental and historical circumstances. Consequently, it is difficult to constitute a universal training set for design of detection or classification algorithms. In this paper, we develop an efficient procedure by which information-theoretic concepts are used to design the basis functions and training set, directly from the site-specific measured data. Specifically, assume that measured data (e.g., induction and/or magnetometer) are available from a given site, unlabeled in the sense that it is not known *a priori* whether a given signature is associated with a target or clutter. For N signatures, the data may be expressed as $\{\mathrm{x}_i, y_i\}_{i=1,N}$, where x_i is the measured data for buried object i, and y_i is the associated *unknown* binary label (target/nontarget). Let the N x_i define the set $\mathbf{X}$. The algorithm works in four steps: 1) the Fisher information matrix is used to select a set of basis functions for the kernel-based algorithm, this step defining a set of n signatures $\mathrm{B}_n \subset \mathrm{X}$ that are most informative in characterizing the signature distribution of the site; 2) the Fisher information matrix is used again to define a small subset $\mathbf{X}_s \subset \mathbf{X}$, composed of those x_i for which knowledge of the associated labels y_i would be most informative in defining the weights for the basis functions in B_n; 3) the buried objects associated with the signatures in $\mathbf{X}_s$ are excavated, yielding the associated labels y_i, represented by the set $\mathbf{Y}_s$; and 4) using $\mathrm{B}_n, \mathbf{X}_s$, and $\mathbf{Y}_s$, a kernel-based classifier is designed for use in classifying all remaining buried objects. This framework is discussed in detail, with example results presented for an actual buried-UXO site.**

***Index Terms*—Active learning, Fisher information, kernel matching pursuit, squared error, subsurface sensing, unexploded ordnance (UXO).**

I. Introduction

IT IS well known that sensor signatures of buried targets such as landmines and unexploded ordnance (UXO) are a strong function of their history and soil environment. For example, radar and seismic sensing of landmines is a strong function of the soil properties [1]. Electromagnetic induction (EMI) and magnetometer [2], [3] sensors are typically less sensitive to soil properties when the target is of high metal content, such as UXO. However, the complexity of the UXO sensing problem is strongly influenced by which ordnance are present, on how the ordnance impacted the soil, and on the surrounding man-made conducting clutter and UXO fragments. All of these issues are dependent on the history of a given UXO site.

These characteristics of the subsurface-sensing problem significantly complicate design of detection and classification algorithms, since it is difficult to define a set of landmine or UXO sensor signatures that are, for algorithm-training purposes [4], generally representative (for all landmine and UXO sites). In this paper, we investigate a technique whereby detection and classification algorithms may be designed for sensing buried landmines and UXO without requiring a separate training set of representative target and clutter signatures. The approach is based on the realization that, when sensing landmines and UXO, one will eventually excavate buried targets based on the sensor data. The approach developed here chooses which items to excavate initially, based on their importance in design of the associated detection and classification algorithm.

Let $\{\mathbf{x}_i\}_{i=1,N}$ represent the *known* measured signatures of the N subsurface objects at a given site, with the set of all $\mathbf{x}_i$ denoted as $\mathbf{X}$. Further, let $\{y_i\}_{i=1,N}$ represent the associated *unknown* binary labels (target/nontarget) of the signatures, to be determined in the detection phase. We here develop a kernel-based classifier, by which an observed signature or feature vector $\mathbf{x}$ is classified using the function

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i K(\mathbf{x}, \mathbf{b}_i) + w_o \tag{1}$$

where $\mathbf{b}_i$ is the i th basis function, w_i are scalar weights, w_o is a scalar offset or bias, and $K(\mathbf{x}, \mathbf{b}_i)$ is a general kernel that defines the similarity of $\mathbf{x}$ to $\mathbf{b}_i$. For a prescribed threshold t, $\mathbf{x}$ is deemed associated with the $+1$ class if $f(\mathbf{x}) \geq t$ and associated with the -1 class if $f(\mathbf{x}) < t$, and by varying the threshold t, one yields the receiver operating characteristic (ROC)[1] [4]. Algorithms that utilize the form in (1) include the support vector machine (SVM) [5], [6], the relevance vector machine (RVM) [7], kernel matching pursuits (KMP) [8], as well as many other related algorithms [9]–[11].

In the design of a classifier of the form in (1), the $\mathbf{b}_i$ typically come from a separate training set, for which the associated labels y_i are known. In this case, the goal is to design a classifier of the form in (1) that correctly identifies the labels of the training

Manuscript received September 2, 2003; revised August 2, 2004. This research was supported in part by the U.S. Strategic Environmental Research and Development Program under Project UX-1281 and in part by the Defense Advanced Research Project Agency under a Multidisciplinary University Research Initiative dedicated to Adaptive Multi-Modality Inverse Scattering.

The authors are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708-0291 USA (e-mail: lcarin@ee.duke.edu).

Digital Object Identifier 10.1109/TGRS.2004.836270

[1]Rigorously speaking, the ROC was originally developed for a likelihood-ratio test [4], and therefore what we consider here is arguably ROC-like. For notational convenience, throughout we refer to such as an ROC.

data (when $\mathbf{x} = \mathbf{b}_i$ for any i), with the hope that this will generalize to data $\mathbf{x}$ not observed while training. The aforementioned variability of subsurface-target signatures makes the idea of utilizing a separate training set undesirable and often impractical.

In the approach taken in this paper, the set of basis functions $\mathbf{B}_n = \{\mathbf{b}_i\}_{i=1,n}$ is selected from the set of observed data $\mathbf{X}$, i.e., $\mathbf{B}_n \subset \mathbf{X}$. This is done because such basis functions will be well matched to the data to be classified, *vis-à-vis* other data that may have come from a different site. The set $\mathbf{B}_n$ is defined by selecting those signatures from $\mathbf{X}$ that are most representative of the measured data from the site of interest, using fundamental information-theoretic considerations to be detailed below. Note that the labels (identities) of the subsurface objects associated with $\mathbf{B}_n$ are not required at this point. Having defined the basis set for (1), we must determine the associated model weights $\{w_i\}_{i=1,n}$ and w_o (denoted collectively by the vector $\mathbf{w}$), and for this task we require labeled data. Therefore, we define a subset of signatures $\mathbf{X}_s \subset \mathbf{X}$ for which knowledge of the associated labels $\mathbf{L}_s$ would be most informative in the context of defining the model weights. The set of signatures $\mathbf{X}_s$ is again determined via information-theoretic metrics detailed below. Note that the sets $\mathbf{B}_n$ and $\mathbf{X}_s$ may overlap, but they are in general distinct. After excavating the items associated with $\mathbf{X}_s$, yielding $\mathbf{L}_s$, the algorithm in (1) is trained as usual [8] and then applied to $\mathbf{x} \notin \mathbf{X}_s$. The key point is that the training set $(\mathbf{X}_s, \mathbf{L}_s)$ is determined adaptively on the observed site-dependent data, via fundamental information-theoretic metrics.

We demonstrate using measured EMI and magnetometer data from an actual UXO site that the sets $\mathbf{X}_s$ and $\mathbf{L}_s$ are often of small dimension, thereby minimizing the amount of excavation required for algorithm design. Once the algorithm has been designed, the fact that it is well matched to the environment often yields a significant reduction in the false-alarm rate, thereby ultimately reducing the total number of excavations (i.e., the false-alarm probability is reduced, and therefore less excavation is required of clutter).

The remainder of the paper is organized as follows. In Section II, we discuss the selection of basis functions $\mathbf{B}_n$ and labeled data $\mathbf{X}_s$. In Section III, we provide further theoretical analysis of the selection criteria, linking them to the minimization of squared errors. We present in Section IV example results on EMI and magnetometer detection of UXO from an actual UXO site. The work is summarized in Section V.

II. Active Classifier Design

A. Model Structure

The decision function in (1), using n basis functions, may be expressed concisely as [8]

$$f_n(\mathbf{x}) = \sum_{i=1}^{n} w_{n,i} K(\mathbf{x}, \mathbf{b}_i) + w_{n,0} = \mathbf{w}_n^T \boldsymbol{\phi}_n(\mathbf{x}) \tag{2}$$

where

$$\boldsymbol{\phi}_n(\mathbf{x}) = [1, K(\mathbf{x}, \mathbf{b}_1), K(\mathbf{x}, \mathbf{b}_2), \ldots, K(\mathbf{x}, \mathbf{b}_n)]^T \tag{3}$$

$$\mathbf{w}_n = [w_{n,0}, w_{n,1}, w_{n,2}, \ldots, w_{n,n}]^T. \tag{4}$$

Assume that the basis set $\mathbf{B}_n = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ is known. Moreover, assume that the item associated with signature $\mathbf{x}_i$ is excavated (this is termed an "experiment"), from which we learn the associated label y_i, where by construction $y_i = 1$ for one class and $y_i = -1$ for the other class (target/no-target). The label found by the experiment is related to the prediction $f_n(\mathbf{x})$ by

$$y_i = \mathbf{w}_n^T \boldsymbol{\phi}_n(\mathbf{x}_i) + \varepsilon_i \tag{5}$$

where $\varepsilon(\mathbf{x}_i)$ is the error term resulting from imperfections in the model. In algorithm design, one desires weights $\mathbf{w}$ that minimize the error observed on training data, for which the data and labels are known. If the training data are well matched to the subsequent testing data, then the algorithm is likely to constitute a robust detection procedure. As indicated above, in many subsurface-sensing problems it is impractical to have a separate training set, with this addressed by the information-theoretic techniques discussed below.

B. Selection of Basis Functions

If we assume that the ε_i in (5) are Gaussian and independent with variance σ_i^2, then the Fisher information matrix associated with $\mathbf{X}$ and $\mathbf{B}_n$ is expressed as [12]

$$\mathbf{M}_n = \sum_{i=1}^{N} \sigma_i^{-2} \boldsymbol{\phi}_n(\mathbf{x}_i) \boldsymbol{\phi}_n^T(\mathbf{x}_i) = \sum_{i=1}^{N} \sigma_i^{-2} \boldsymbol{\phi}_{n,i} \boldsymbol{\phi}_{n,i}^T \tag{6}$$

where $\phi_{n,i} \equiv \phi_n(\mathbf{x}_i)$. Note that in computing $\mathbf{M}_n$ we do not require the labels associated with $\mathbf{B}_n$ and $\mathbf{X}$ (this is a result of the fact that the model in (2) is linear in the weights $\mathbf{w}_n$). As discussed by Fedorov [12], the Fisher information matrix in (6) is associated with the uncertainty in the model weights $\mathbf{w}$, as defined through all N measured $\mathbf{x}_i$ and the basis $\mathbf{B}_n$. By appending a new basis function to $\phi_n(\cdot)$, one obtains

$$\boldsymbol{\phi}_{n+1}(\cdot) = \begin{bmatrix} \boldsymbol{\phi}_n(\cdot) \\ \phi_{n+1}(\cdot) \end{bmatrix} \tag{7}$$

where $\phi_{n+1}(\cdot) = K(\cdot, \mathbf{b}_{n+1})$ and $\mathbf{b}_{n+1} \in \mathbf{X}, \mathbf{b}_{n+1} \notin \mathbf{B}_n$. Following (2), we can write from ϕ_{n+1} the augmented classifier f_{n+1}, for which the Fisher information matrix is found to be

$$\begin{aligned} \mathbf{M}_{n+1} &= \sum_{i=1}^{N} \sigma_i^{-2} \begin{bmatrix} \boldsymbol{\phi}_{n,i} \\ \phi_{n+1,i} \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}_{n,i}^T \, \phi_{n+1,i} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{M}_n & \sum_{i=1}^{N} \sigma_i^{-2} \boldsymbol{\phi}_{n,i} \phi_{n+1,i} \\ \sum_{i=1}^{N} \sigma_i^{-2} \phi_{n+1,i} \boldsymbol{\phi}_{n,i}^T & \sum_{i=1}^{N} \sigma_i^{-2} \phi_{n+1,i}^2 \end{bmatrix} \end{aligned} \tag{8}$$

where $\phi_{n+1,i} \equiv \phi_{n+1}(\mathbf{x}_i)$. The expression in (8) is again associated with fitting the model to the N measured $\mathbf{x}_i$, but now using an $(n+1)$-dimensional basis $\mathbf{B}_{n+1}$, *vis-à-vis* the n-dimensional basis $\mathbf{B}_n$ in (6). We develop a metric which compares (6) and (8), thereby quantifying the information gain by adding the new basis $\mathbf{b}_{n+1}$.

There are many ways of comparing the information content reflected by $\mathbf{M}_n$ and $\mathbf{M}_{n+1}$, and here we employ the so-called

D-optimal procedure [12], defined as the determinant of the information matrix. The logarithm of the determinant of $\mathbf{M}$ is denoted q_n, and it may be shown that

$$q_{n+1} = q_n + \ln r(\mathbf{b}_{n+1}) \tag{9}$$

where

$$r(\mathbf{b}_{n+1}) = \sum_{i=1}^{N} \sigma_i^{-2}\phi_{n+1,i}^2 - \sum_{i=1}^{N} \sigma_i^{-2}\phi_{n+1,i}\boldsymbol{\phi}_{n,i}^T \cdot \mathbf{M}_n^{-1} \sum_{i=1}^{N} \sigma_i^{-2}\boldsymbol{\phi}_{n,i}\phi_{n+1,i}. \tag{10}$$

Since $N \geq n$ the matrix $\mathbf{M}_n$ is full rank, and its inverse exists (assuming that n of the vectors $\{\boldsymbol{\phi}_n(\mathbf{x}_i)\}_{i=1,N}$ are linearly independent). Under these conditions, it can be shown that $r > 0$, and therefore $\ln r$ in (9) is generally valid. Note that if $r(\mathbf{b}_{n+1}) = 0$, then $\mathbf{M}_{n+1}$ is rank deficient, and we delete the basis function from the candidate set and proceed to select from the remaining ones (however, we have not found $r(\mathbf{b}_{n+1}) = 0$ in practice).

It is known from information theory [13] that the inverse of $\mathbf{M}_n$ gives the Cramer–Rao lower bound (CRLB) of the covariance matrix of the estimate of $\mathbf{w}_n$ and the reciprocal of q_n lower bounds the product of its eigenvalues. The CRLB is here the actual covariance, assuming the Gaussian model. A larger q_n implies low variances of the components of $\mathbf{w}_n$. Given the nth order decision function f_n, q_n is fixed, and one relies on maximization of $\ln r(\mathbf{b}_{n+1})$ to obtain a large value of q_{n+1}. This can be achieved by conducting a "greedy" search for the new $\mathbf{b}_{n+1}$ in $\mathbf{X}$ with the previously selected support data excluded

$$\mathbf{b}_{n+1} = \arg \max_{\mathbf{b}\in\mathbf{X}, \mathbf{b}\notin\mathbf{B}_n} \ln r(\mathbf{b}). \tag{11}$$

Using the procedure outlined above, basis elements $\mathbf{b}_n$ are added until the information gain reflected in $q_{n+1} - q_n$ is no longer deemed significant. Note from (9) and (10) that evaluation of (11) does not require knowledge of the target labels y_i, and therefore, no excavation is required to determine the basis $\mathbf{B}_n$. The greedy method is suboptimal, but in practice often provides good results.

C. Selection of Labeled Data, for Model Training

Assume that the procedure discussed above selects n bases from the observed data $\mathbf{X}$. We now require labeled data to optimize the associated model weights $\mathbf{w}$. In a manner analogous to the previous discussion, we select those $\mathbf{x}_i \in \mathbf{X}$ for which knowledge of the associated labels y_i would be most informative in the context of defining $\mathbf{w}$. Those $\mathbf{x}_i$ that are so selected define a subset of signatures $\mathbf{X}_s \subset \mathbf{X}$, and these items are excavated to yield the respective set of labels $\mathbf{L}_s$. The set of signatures and labels $(\mathbf{X}_s, \mathbf{L}_s)$ are then used to define the weights $\mathbf{w}$ in a least squares sense, and the resulting model $f(\mathbf{x})$ is used to specify which of the remaining signatures $\mathbf{x} \notin \mathbf{X}_s$ are likely targets of interest.

Assume that there are J signatures in $\mathbf{X}_s$, denoted $\mathbf{X}_{s,J}$. We quantify the information context in $\mathbf{X}_{s,J}$ in the context of estimating the model weights $\mathbf{w}$ and further ask which $\mathbf{x}_i \notin \mathbf{X}_{s,J}$ would be most informative if it and its label were added for determination of $\mathbf{w}$. Analogous to (6), we have

$$\mathbf{M}_n(\mathbf{X}_{s,J}) = \sum_{i:\mathbf{x}_i\in\mathbf{X}_{s,J}} \sigma_i^{-2}\boldsymbol{\phi}_{n,i}\boldsymbol{\phi}_{n,i}^T. \tag{12}$$

The expressions in (6) and (12) both employ an n-dimensional basis set $\mathbf{B}_n \subset \mathbf{X}$. The distinction is that in (6) we are interested in defining $\mathbf{B}_n$, and we sum over all observed signatures $\{\mathbf{x}_i\}_{i=1,N}$. By contrast, in (12), the basis set $\mathbf{B}_n$ is known and fixed, and we are only summing over those signatures $\mathbf{X}_{s,J}$ for which knowledge of the associated labels is most informative in defining the model weights $\mathbf{w}$.

After adding a new signature $\mathbf{x}_i \in \mathbf{X}, \mathbf{x}_i \notin \mathbf{X}_{s,J}$, we now have $\mathbf{X}_{s,J+1}$, and $\mathbf{M}_n$ is updated as

$$\mathbf{M}_n(\mathbf{X}_{s,J+1}) = \mathbf{M}_n(\mathbf{X}_{s,J}) + \sigma_{i_{J+1}}^{-2}\boldsymbol{\phi}_{n,i_{J+1}}\boldsymbol{\phi}_{n,i_{J+1}}^T \tag{13}$$

where i_{J+1} represents the index of the new signature selected for $\mathbf{X}_{s,J+1}$. Using the matrix identity $\det(\mathbf{A}+\mathbf{F}\mathbf{F}^T) = \det(\mathbf{I}+\mathbf{F}^T\mathbf{A}^{-1}\mathbf{F})\det(\mathbf{A})$, one obtains from (13)

$$q_n(\mathbf{X}_{s,J+1}) = q_n(\mathbf{X}_{s,J}) + \ln \rho(\mathbf{x}_{i_{J+1}}) \tag{14}$$

with

$$\rho\left(\mathbf{x}_{i_{J+1}}\right) = 1 + \sigma_{i_{J+1}}^{-2}\boldsymbol{\phi}_{n,i_{J+1}}^T\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})\boldsymbol{\phi}_{n,i_{J+1}}. \tag{15}$$

Care is needed with regard to evaluating the inverse of $\mathbf{M}_n$, since if $J < n$ the matrix is rank deficient. We have considered addressing this in either of two ways. A standard approach for inversion of such matrices is to add a small diagonal term to $\mathbf{M}_n$, such that its inverse exists. Alternatively, by construction one can assume that the items associated with the basis $\mathbf{B}_n$ are all associated with $\mathbf{X}_{s,J}$, yielding a minimum of n labeled data and, therefore, assuring that the matrix is full rank. We have examined both procedures, and they yield comparable results. We use the second approach in all examples presented in Section III.

Having addressed the inverse of $\mathbf{M}_n$, one iteratively maximizes $\ln \rho(\mathbf{x}_{i_{J+1}})$ to obtain

$$\mathbf{x}_{i_{J+1}} = \arg \max_{\mathbf{x}\in\mathbf{X}, \mathbf{x}\notin\mathbf{X}_{s,J}} \ln \rho(\mathbf{x}). \tag{16}$$

Note that to define $\mathbf{x}_{i_{J+1}}$ we again do not require the signature labels. The elements of $\mathbf{X}_s$ are selected iteratively, in a "greedy" fashion as indicated in (16), until the information gain is below a prescribed threshold. After J iterations, we have defined those signatures $\mathbf{X}_{s,J}$ for which knowledge of the labels will best approximate the weights $\mathbf{w}$. These items are excavated, yielding the labels $\mathbf{L}_{s,J}$.

For the assumptions underlying the linear model in (5), and assuming knowledge of $\mathbf{B}_n$ and $(\mathbf{X}_{s,J}, \mathbf{L}_{s,J})$, the optimal estimation for the weights $\mathbf{w}$ is expressed as [8], [12]

$$\mathbf{w} = [\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}]^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{17}$$

where

$$\boldsymbol{\Sigma} = \mathrm{diag}[\sigma_{i_1}^2, \sigma_{i_2}^2, \ldots, \sigma_{i_J}^2] \tag{18}$$

where $\mathbf{y}$ represents the set of labels determined via the J excavations $\mathbf{y} = [y_{i_1}, y_{i_2}, \ldots, y_{i_J}]^T$, and and the $J \times (n+1)$ matrix $\mathbf{\Phi}$ is defined as

$$\mathbf{\Phi} = [\boldsymbol{\phi}_n(\mathbf{x}_{i_i}), \boldsymbol{\phi}_n, (\mathbf{x}_{i_2}), \ldots, \boldsymbol{\phi}_n(\mathbf{x}_{i_J})]^T \tag{19}$$

where, for example, $\mathbf{x}_{i_1}$ corresponds to y_{i_1}.

In the classification stage, we consider $\mathbf{x} \notin \mathbf{X}_{s,J}$ and compute $f(\mathbf{x})$. For a prescribed threshold t, $\mathbf{x}$ is deemed associated with the $+1$ class if $f(\mathbf{x}) \geq t$, and associated with the -1 class if $f(\mathbf{x}) < t$, and by varying the threshold t, one yields the ROC [4]. The key component of the model $f(\mathbf{x})$ is that it is linear in the weights $\mathbf{w}$, which yields a closed-form procedure for selection of $\mathbf{B}_n$ and $\mathbf{X}_{s,J}$, as indicated in the previous sections.

III. Theoretical Motivation for Classifier Design

In the previous two sections, we have presented procedures for selecting basis functions for a kernel-based classifier, based on a set of unlabeled data. After designing the basis set, we have also addressed selection of which signatures would be most informative for classifier training, if the associated signature labels were known. In this section, we provide theoretical justification for these design procedures, and in Section IV, example results are presented for UXO sensing.

A. Basis-Function Selection

To simplify notation, we utilize matrix expressions in our derivation. Let the basis functions $\boldsymbol{\phi}_n(\cdot)$ be evaluated for all initially unlabeled data points $\{\mathbf{x}_i\}_{i=1,N}$, and stacked to form the matrix $\tilde{\mathbf{\Phi}}_n = [\boldsymbol{\phi}_n(\mathbf{x}_1), \boldsymbol{\phi}_n(\mathbf{x}_2), \ldots, \boldsymbol{\phi}_n(\mathbf{x}_N)]^T$. Let the data labels be denoted $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}^T$, although these labels are not required when designing the basis functions. In this section, we assume $\sigma_i = 1$, for $i = 1, 2, \ldots, N$. This causes no loss of generality, as we can always absorb σ's into $\tilde{\mathbf{\Phi}}_n$ and $\mathbf{y}$ by making $\tilde{\mathbf{\Phi}}_n = \mathbf{\Sigma}^{-1/2}[\boldsymbol{\phi}_n(\mathbf{x}_1), \boldsymbol{\phi}_n(\mathbf{x}_2), \ldots, \boldsymbol{\phi}_n(\mathbf{x}_N)]^T$ and $\mathbf{y} = \mathbf{\Sigma}^{-1/2}[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]^T$, where $\mathbf{\Sigma}$ is as defined in (18). The difference between the true labels and those output by the classifier (2), for all $\{\mathbf{x}_i\}_{i=1,N}$ is expressed in vector form as

$$\begin{aligned}\mathbf{y} - \tilde{\mathbf{\Phi}}_n \left(\tilde{\mathbf{\Phi}}_n^T \tilde{\mathbf{\Phi}}_n\right)^{-1} \tilde{\mathbf{\Phi}}_n^T \mathbf{y} &\overset{1}{=} \left(\mathbf{I}_N - \tilde{\mathbf{\Phi}}_n \left(\tilde{\mathbf{\Phi}}_n^T \tilde{\mathbf{\Phi}}_n\right)^{-1} \tilde{\mathbf{\Phi}}_n^T\right) \mathbf{y} \\ &\overset{2}{\approx} \Big(\mathbf{I}_N - \tilde{\mathbf{\Phi}}_n \left(\lambda \mathbf{I}_{n+1} \right. \\ &\qquad \left. + \tilde{\mathbf{\Phi}}_n^T \tilde{\mathbf{\Phi}}_n\right)^{-1} \tilde{\mathbf{\Phi}}_n^T\Big) \mathbf{y} \\ &\overset{3}{=} \left(\mathbf{I}_N + \frac{1}{\lambda} \tilde{\mathbf{\Phi}}_n \tilde{\mathbf{\Phi}}_n^T\right)^{-1} \mathbf{y}\end{aligned} \tag{20}$$

where $\mathbf{I}_N$ is a $N \times N$ identity matrix ($\mathbf{I}_n$ is defined similarly), and λ is a small positive number. The equality 3 in (20) is due to the Sherman–Morrison–Woodbury formula. From (20), the squared error between the true and estimated labels is

$$e_n^2 \approx \mathbf{y}^T \left(\mathbf{I}_N + \frac{1}{\lambda} \tilde{\mathbf{\Phi}}_n \tilde{\mathbf{\Phi}}_n^T\right)^{-2} \mathbf{y}. \tag{21}$$

The expression in (21) shows that for given basis functions $\boldsymbol{\phi}_n(\cdot)$, we have approximately expressed the squared error as a quadratic form of the labels $\mathbf{y}$, with a coefficient matrix $\mathbf{C}_n^{-2}$ with $\mathbf{C}_n = \mathbf{I}_N + \tilde{\mathbf{\Phi}}_n \tilde{\mathbf{\Phi}}_n^T / \lambda$. The approximation can be made as accurate as desired by making λ sufficiently small. Without knowing $\mathbf{y}$, we prefer $\mathbf{C}_n$ to have large eigenvalues, to make the error e_n^2 small. This is accomplished by making the determinant of $\mathbf{C}_n$ large. The logarithmic determinant of $\mathbf{C}_n$ is

$$\begin{aligned}q_n^{(2)} &\overset{1}{=} \ln \det(\mathbf{C}_n) \overset{2}{=} \ln \det\left(\mathbf{I}_N + \frac{1}{\lambda} \tilde{\mathbf{\Phi}}_n \tilde{\mathbf{\Phi}}_n^T\right) \\ &\overset{3}{=} \ln \frac{\det\left(\lambda \mathbf{I}_{n+1} + \tilde{\mathbf{\Phi}}_n^T \tilde{\mathbf{\Phi}}_n\right)}{\lambda^{n+1}} \\ &\overset{4}{=} \ln \frac{\det(\lambda \mathbf{I}_{n+1} + \mathbf{M}_n)}{\lambda^{n+1}}\end{aligned} \tag{22}$$

where equality 3 is due to the property of matrix determinants, and equality 4 is due to (6). Adding a new basis function to $\boldsymbol{\phi}_n(\cdot)$, we get $\boldsymbol{\phi}_{n+1}(\cdot)$ as given in (7). The logarithmic determinant of $\mathbf{C}_{n+1} = \mathbf{I}_N + \tilde{\mathbf{\Phi}}_{n+1} \tilde{\mathbf{\Phi}}_{n+1}^T / \lambda$ is

$$q_{n+1}^{(2)} = \ln \frac{\det(\lambda \mathbf{I}_{n+2} + \mathbf{M}_{n+1})}{\lambda^{n+2}}. \tag{23}$$

Following the method of obtaining (9) and (10), we can show that $q_n^{(2)}$ and $q_{n+1}^{(2)}$ are related by

$$q_{n+1}^{(2)} = \ln q_n^{(2)} + \ln \frac{r^{(2)}(\phi_{n+1})}{\lambda} \tag{24}$$

with

$$\begin{aligned}r^{(2)}(\phi_{n+1}) = \lambda + \sum_{i=1}^{N} \phi_{n+1,i}^2 - \sum_{i=1}^{N} \phi_{n+1,i} \boldsymbol{\phi}_{n,i}^T \\ \cdot (\lambda \mathbf{I}_{n+1} + \mathbf{M}_n)^{-1} \sum_{i=1}^{N} \boldsymbol{\phi}_{n,i} \phi_{n+1,i}\end{aligned} \tag{25}$$

where $\boldsymbol{\phi}_{n,i} \equiv \phi_n(\mathbf{x}_i)$ and $\phi_{n+1,i} \equiv \phi_{n+1}(\mathbf{x}_i)$. Since we wish for a $\mathbf{C}_{n+1}$ with large determinant, we want to make $\ln(r^{(2)}(\phi_{n+1}))/(\lambda)$ or equivalently $\ln r^{(2)}(\phi_{n+1})$ large, as λ is a constant.

B. Selection of Examples for Labeling

Assume the basis functions $\boldsymbol{\phi}_n(\cdot)$ have been selected in the manner discussed above. Moreover, assume we have selected the subset $\mathbf{X}_{s,J} = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_J}\}$ of J signatures for which the associated labels will be acquired. The Fisher information matrix associated with $\mathbf{X}_{s,J}$ is $\mathbf{M}_n(\mathbf{X}_{s,J}) = \sum_{k=1}^{J} \boldsymbol{\phi}_{n,i_k} \boldsymbol{\phi}_{n,i_k}^T$. The Fisher information matrix for an augmented set $\mathbf{X}_{s,J} \cup \{\mathbf{x}\} = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_J}, \mathbf{x}\}$ is

$$\mathbf{M}_n(\mathbf{X}_{s,J} \cup \{\mathbf{x}\}) = \mathbf{M}_n(\mathbf{X}_{s,J}) + \boldsymbol{\phi}_n(\mathbf{x}) \boldsymbol{\phi}_n^T(\mathbf{x}). \tag{26}$$

Suppose we have two classifiers $f_n^{\mathbf{X}_{s,J} \cup \{\mathbf{x}\}}(\cdot)$ and $f_n^{\mathbf{X}_{s,J}}(\cdot)$, which are trained using $\mathbf{X}_{s,J} \cup \{\mathbf{x}\}$ and $\mathbf{X}_{s,J}$, respectively. We

test $f_n^{\mathbf{X}_{s,J}\cup\{\mathbf{x}\}}(\cdot)$ and $f_n^{\mathbf{X}_{s,J}}(\cdot)$ on $\mathbf{x}$ and examine how the two results are related. As given in [14, p. 121], we have

$$\left[f_n^{\mathbf{X}_{s,J}}(\mathbf{x})-y(\mathbf{x})\right]^2 = \frac{\left[f_n^{\mathbf{X}_{s,J}\cup\{\mathbf{x}\}}(\mathbf{x})-y(\mathbf{x})\right]^2}{1-\boldsymbol{\phi}_n^T(\mathbf{x})\left[\mathbf{M}_n(\mathbf{X}_{s,J})+\boldsymbol{\phi}_n(\mathbf{x})\boldsymbol{\phi}_n^T(\mathbf{x})\right]^{-1}\boldsymbol{\phi}_n(\mathbf{x})}. \quad (27)$$

By using the Sherman–Morrison–Woodbury formula, we obtain

$$\left[\mathbf{M}_n(\mathbf{X}_{s,J})+\boldsymbol{\phi}_n(\mathbf{x})\boldsymbol{\phi}_n^T(\mathbf{x})\right]^{-1} = \mathbf{M}_n^{-1}(\mathbf{X}_{s,J}) - \frac{\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})\boldsymbol{\phi}_n(\mathbf{x})\boldsymbol{\phi}_n^T(\mathbf{x})\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})}{1+\boldsymbol{\phi}_n^T(\mathbf{x})\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})\boldsymbol{\phi}_n(\mathbf{x})}$$

which is set into (27) to give

$$\left[f_n^{\mathbf{X}_{s,J}\cup\{\mathbf{x}\}}(\mathbf{x})-y(\mathbf{x})\right]^2 = \frac{\left[f_n^{\mathbf{X}_{s,J}}(\mathbf{x})-y(\mathbf{x})\right]^2}{1+\boldsymbol{\phi}_n^T(\mathbf{x})\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})\boldsymbol{\phi}_n(\mathbf{x})}. \quad (28)$$

Equation (28) shows that by including $\mathbf{x}$ in the training dataset, the squared test error on $\mathbf{x}$ will drop by a factor

$$\rho^{(2)}(\mathbf{x}) = 1+\boldsymbol{\phi}_n^T(\mathbf{x})\mathbf{M}_n^{-1}(\mathbf{X}_{s,j})\boldsymbol{\phi}_n(\mathbf{x}). \quad (29)$$

If $\rho^{(2)}(\mathbf{x})\approx 1$, we do not require the label for $\mathbf{x}$, as it does not give much decrease of the squared error. On the other hand, if $\rho^{(2)}(\mathbf{x})\gg 1$, it significantly decreases the squared error. Therefore, the $\mathbf{x}$ that maximizes $\rho^{(2)}(\mathbf{x})$ should be selected to seek the associated label y. Comparing (29) to (15), we note that $\rho^{(2)}(\mathbf{x})$ is exactly equivalent to $\rho(\mathbf{x})$, and thus, the $\mathbf{x}$ that maximizes $\rho(\mathbf{x})$ is the one that contributes the maximally to make the squared test error small.

IV. Application to UXO Detection

The active-training methodology addressed in this paper may be applied to any detection problem for which the data labels are expensive to acquire and for which there is no distinct training data. In particular, we consider the detection of buried UXO. For UXO remediation, the label of a potential target is acquired by excavation, a dangerous and time-consuming task. The overwhelming majority of UXO cleanup costs come from excavation of non-UXO items. In this context, note that *a priori* excavations are required for the procedure in Section II (to obtain labeled training data). However, if the false-alarm rate is reduced at the desired detection probability, then overall cleanup costs may diminish substantially (i.e., overall, less non-UXO items need be excavated).

The results presented here are for data collected at an actual UXO site: Jefferson Proving Ground in the United States. The technique in Section II is compared with results obtained using existing procedures. Specifically, the principal challenge in UXO sensing is development of a training set, for design of the detection algorithm. At an actual UXO site, there is often a significant quantity of UXO, UXO fragments, and man-made clutter *on the surface*. It has been recognized that the characteristics of the surface UXO and clutter is a good indicator of what will be found in the subsurface. Consequently, in practice, a subset of the surface UXO and clutter are buried, and magnetometer and induction data are collected for these items, for which the labels are obviously known. The measured data and associated labels (UXO/non-UXO) are then used for training purposes. Of course, the process of burying, collecting data, and then excavating these emplaced items is time consuming and dangerous (for the UXO items), with this procedure eliminated by the techniques outlined in Section II.

A. Magnetometer and EMI Sensors

Magnetometer and EMI sensors are widely applied in sensing buried conducting/ferrous targets, such as landmines and UXO. The magnetometer is a passive sensor that measures the change of the earth's background magnetic field due to the presence of a ferrous target. Magnetometers measure static magnetic fields. An EMI sensor actively transmits a time-varying electromagnetic field and, consequently, senses the dynamic induced secondary field from the target. To enhance soil penetration, EMI sensors typically operate at kilohertz frequencies. We here employ a frequency-domain EMI sensor that transmits and senses at several discrete frequencies [15]. Magnetometers only sense ferrous targets, while EMI sensors detect general conducting and ferrous items.

Parametric models have been developed for both magnetometer and EMI sensors [16]–[18]. The target features $\mathbf{x}$ are extracted by fitting the EMI and magnetometer models to measured sensor data. The vector $\mathbf{x}$ has parameters from both the magnetometer and EMI data, and therefore, in this sense the data from these two sensors are "fused." The one place where these two models have overlapping parameters is in specification of the target position. The magnetometer data often yield a very good estimation of the target position, and therefore, such are used in $\mathbf{x}$. In fact, the target position specified by the magnetometer data is explicitly utilized as prior information when fitting EMI data to the EMI parametric model. Details on the magnetometer and EMI models, and on the model-fitting procedure, may be found in [18].

The features employed are as in [18], and the features are centered and normalized. Specifically, using the training data, we compute the mean feature vector $\mathbf{x}_{\text{mean}}$ and the variance of each feature component (let σ_i^2 represent the variance of the i th feature). Before classification, a given feature vector $\mathbf{x}$ is shifted by implementing $\mathbf{x}_{\text{shift}}=\mathbf{x}-\mathbf{x}_{\text{mean}}$, and then the ith feature component of $\mathbf{x}_{\text{shift}}$ is divided by σ_i to effect the normalization.

B. Measured Sensor Data From the Jefferson Proving Ground

Jefferson Proving Ground (JPG) is a former military range that has been utilized for UXO technology demonstrations since 1994. We consider data collected by Geophex, Ltd. in the latest phase (Phase V) of the JPG demonstration. The goal of the JPG V is to evaluate the UXO detection and discrimination abilities under realistic scenarios, where man-made and natural clutter coexist with UXO items. Our results are presented with the GEM-3 and magnetometer data from two adjoining areas, constituting a total of approximately five acres. There are 433 potential targets detected from sensor anomalies, 40 of which are proven to be UXO and the others are clutter. The excavated

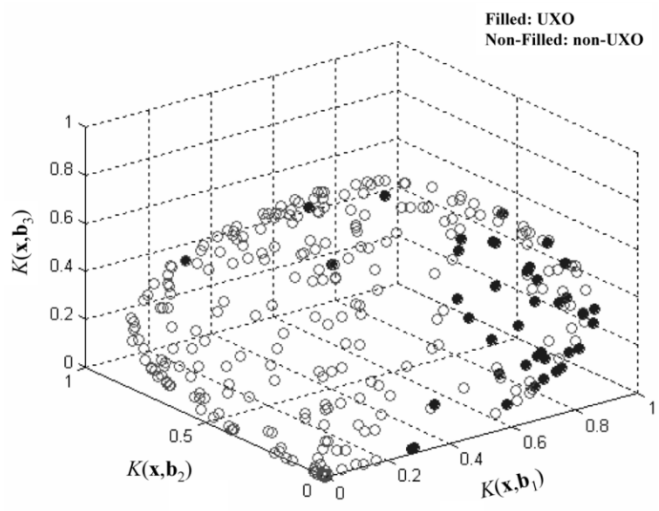

Fig. 1. For the first three basis functions selected, b1, b2, and b3, the 3-D vector $[K(\mathbf{x},\mathbf{b}1), K(\mathbf{x},\mathbf{b}2), K(\mathbf{x},\mathbf{b}3)]$ is plotted. Considered are feature vectors $\mathbf{x}$ for all UXO and non-UXO targets considered in this study.

UXO items include 4.2-in, 60-, and 81-mm mortars; 5-in, 57-, 76-, 105-, 152-, and 155-mm projectiles; and 2.75-in rockets.

This test was performed with U.S. Army oversight. One of the two JPG areas was assigned as the training area, for which the ground truth (UXO/non-UXO) was given. The trained detection algorithms are then tested on the other area, and the associated ground truth was revealed later to evaluate performance. It was subsequently recognized that several UXO types were found in equal number in each of the two areas. This indicates an effort to match the training data to the detection data, in the manner discussed above, involving burial of known UXO and non-UXO collected on the surface.

Each sensor anomaly is processed by fitting the associated magnetometer and EMI data to the parametric models [18], and the estimated parameters define $\mathbf{x}$. In addition, the model-fitting procedure functions as a prescreening tool. Any sensor anomaly failing to fit well to the model is regarded as having been generated by a clutter item. Therefore, a total of 300 potential targets remain after this prescreening stage, 40 of which are UXO. In the training area, there are 128 buried items, 16 of which are UXO.

C. Detection Results

Before presenting classification results, we examine the characteristics of the basis functions selected in the first phase of the algorithm, prior to adaptively selecting training data. In Fig. 1, we consider the first three basis functions $\mathbf{b}_1, \mathbf{b}_2$, and $\mathbf{b}_3$ selected by the first stage of the algorithm. For each feature vector $\mathbf{x}$ (from all UXO and non-UXO), we compute a three-dimensional (3-D) vector $(K(\mathbf{x},\mathbf{b}_1), K(\mathbf{x},\mathbf{b}_2), K(\mathbf{x},\mathbf{b}_3))$. By examining this 3-D vector for all $\mathbf{x}$, we may observe the degree to which UXO and non-UXO features are distinguished via the features and kernel. A radial basis function kernel is employed here, corresponding to the kernel used to select the basis functions (see discussion below concerning the selected kernel). By examining Fig. 1, we observe that the UXO and non-UXO features are relatively separated, although there is significant overlap, undermining classification performance.

The detection results are presented in the form of the receiver operating characteristic, quantifying the probability of detection (Pd) as a function of the false-alarm count. We present ROC curves using the adaptive-training approach discussed in Section II, with performance compared to results realized by training on the distinct training region discussed above (the latter approach reflects current practice). With regard to conventional training, the algorithm employed is of identical form as (2), with model weights determined iteratively using kernel matching pursuits. Details on the KMP algorithm may be found in [8] (we have employed the prefitting algorithm in [8]). To make the comparison appropriate, the adaptive training and KMP implementation employ an identical radial basis function (RBF) kernel [10]

$$K(\mathbf{x},\mathbf{b}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\|\mathbf{x}-\mathbf{b}\|_2^2/\sigma^2\right]. \qquad (30)$$

The variance σ^2 is adaptively adjusted after each basis vector is selected before the model weights are determined, and it is not related to the labeled training data. In particular, a gradient search is applied to refine σ^2 with (11), by maximizing $\ln r(\mathbf{b})$.

As indicated above, the designated training area has 128 labeled items, and conventionally classifiers are tested on the remaining signatures, in this case constituting 172 items. As a first comparison, the adaptive technique discussed in Section II is employed to select $J = 128$ items from the original 300, with these "excavated" to learn the associated labels. This, therefore, defines the set $\mathbf{X}_{s,J}$ and associated labels $\mathbf{L}_{s,J}$. The basis set $\mathbf{B}_n$ is also determined adaptively using the original 300 signatures, and here $n = 10$. The performance of the adaptive learning algorithm is then tested on the remaining 172 $\mathbf{x}_i \notin \mathbf{X}_{s,J}$, although these are generally not the same testing examples used via traditional training of the KMP algorithm (the training sets do not overlap completely). For comparison, we also show training and testing results implemented via KMP, in which the 128 training examples are selected randomly from the original 300 signatures $\mathbf{X}$. Performance comparisons are shown in Fig. 2, wherein we present results for active data selection (algorithm in Section II), KMP results using the assigned 128 training examples, and average results for randomly choosing the 128 examples for KMP training (100 random selections were performed). In addition, for the latter case, we also place error bars on the results; the length of the error bar is twice the standard derivation of the Pd for the associated false-alarm count. Therefore, if the result is Gaussian distributed, 95% of the values lie within the error bar.

Before proceeding, we note that the ROC curves are generated by varying the threshold t, as applied to the estimated label y. For the binary UXO-classification problem considered here, by design we choose the label $y = 1$ for UXO and $y = 0$ for non-UXO. In practice, one must choose one point on the ROC at which to operate. A naive choice of the operating point would be 0.5 (i.e., if the classifier maps a testing feature vector $\boldsymbol{x}$ to a label $y > 0.5$, the item is declared UXO, and otherwise it is declared non-UXO). However, we must account for the fact that in practice the number of non-UXO items is often much larger than the number of UXO. We have, therefore, invoked the following procedure.

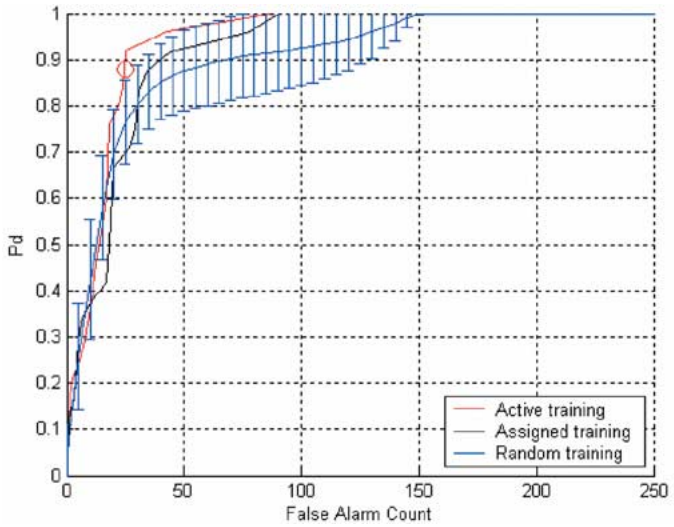

Fig. 2. ROC curves based on 128 training examples, for which the target labels were known. In one case, the training set was carefully designed *a priori*, and in the other, the training examples were chosen adaptively using the algorithm of Section II. For comparison, results are also shown when the 128 training examples are chosen randomly, 100 times. In the latter case, average results are shown, as well as the associated range of variability. The indicated point on the ROC corresponds to (32).

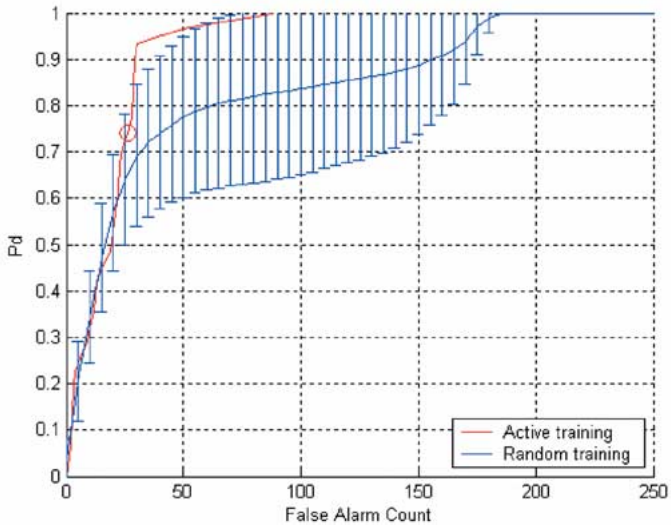

Fig. 3. As in Fig. 2, but now results are only shown for adaptive training-data selection (Section II) and for random selection. In the latter case, results are presented as in Fig. 1. Results are shown for $J = 90$ training examples.

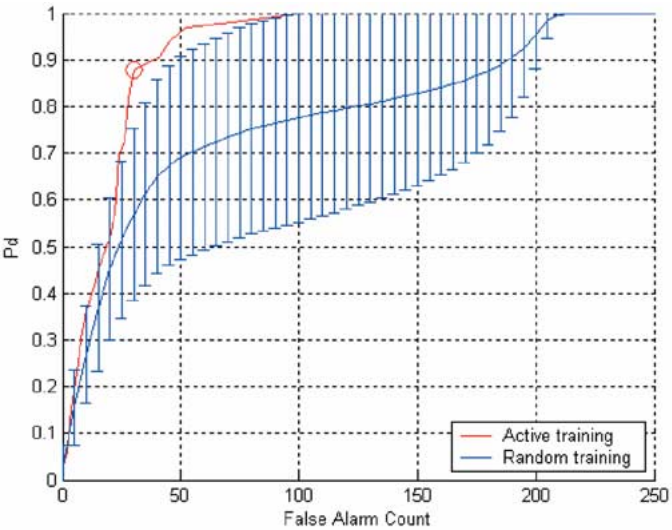

Fig. 4. As in Fig. 3, with $J = 60$.

We assume that the error (noise) between the true label ($y = 1$ or $y = 0$) and the estimated label is independentl identically distributed Gaussian with variance of σ^2, as in (5). Let N_0 and N_1 represent, respectively, the number of non-UXO and UXO items in the training set. Considering the UXO ($y = 1$) data, an unbiased estimator of the label y will yield a mean of one and a minimum variance of σ^2/N_1. Similarly, considering the non-UXO data ($y = 0$), an unbiased estimator of the label y will have zero mean and minimum variance σ^2/N_0. Let H_1 and H_0 correspond to the UXO and non-UXO hypotheses. Based upon the above discussion, we model the probability density function of y for the H_1 and H_0 hypotheses as $p(y \mid H_1) = N(1, \sigma^2/N_1)$ and $p(y \mid H_0) = N(1, \sigma^2/N_0)$. Rather than setting the threshold at $t = 0.5$, we set the threshold at that value of y for which $p(y \mid H_1) = p(y \mid H_0)$, yielding

$$t = \frac{N_1 - \sqrt{N_1^2 - (N_1 - N_0)\left(N_1 + \sigma^2 \ln \frac{N_0}{N_1}\right)}}{N_1 - N_0}. \tag{31}$$

Assuming σ is a small, we omit $\sigma^2 \ln(N_0)/(N_1)$, obtaining

$$t = \frac{\sqrt{N_1}}{\sqrt{N_1} + \sqrt{N_0}}. \tag{32}$$

From (32), the appropriate threshold is $t = 0.5$ only if $N_0 = N_1$.

For example, in Fig. 2, only 15 of the 128 actively selected training data are UXO, and therefore $N_1 = 15, N_0 = 113$. If we set the threshold to be $t = 0.5$, we detect 16% of the UXO with two false alarms. By contrast, using the procedure discussed above (for which $t = 0.27$), we detect 88% of the UXO with 25 false alarms. The operating point corresponding to $t = 0.27$ is indicated in Fig. 2. We similarly plot this point in all subsequent ROCs presented below.

We observe from the results in Fig. 2 that the active data selection procedure produces the best ROC results (for $\mathrm{Pd} > 0.7$, which is of most interest in practice), with the KMP results from the specified training area almost as good. It is observed that the average performance based on choosing the training set randomly is substantially below that of the two former approaches, with significant variability reflected in the error bars. These results demonstrate the power of the active-data-selection algorithm introduced in Section II, and also that the training data defined for JPG V is well matched to the testing data.

In the first example, we set $J = 128$ to be consistent with the size of the training area specified in the JPG V test. The algorithm in Section II can be implemented for smaller values of J, reflecting less excavation required in the training phase (for determination of target labels). It is of interest to examine algorithm performance as J is decreased from 128. In this case, training is performed using signatures and labels from the J "excavated" items, and testing is performed on the remaining 300-J examples. Results are presented for the active training procedure and for randomly choosing J training examples (100 random instantiations), as in Fig. 2. In Figs. 3–5, results are

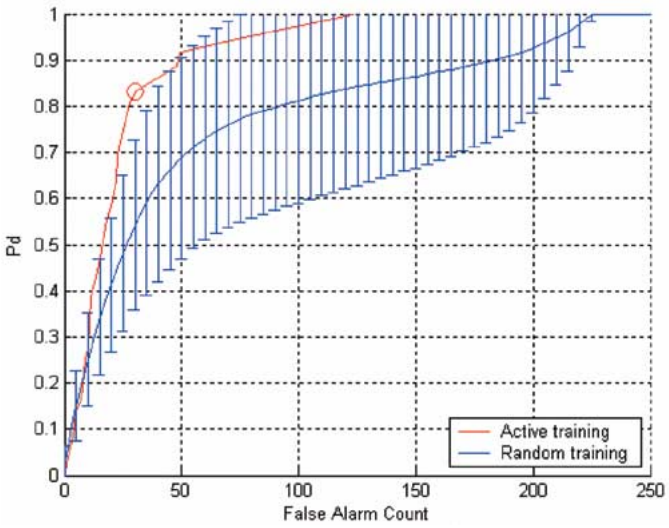

Fig. 5. As in Fig. 4, with $J = 40$.

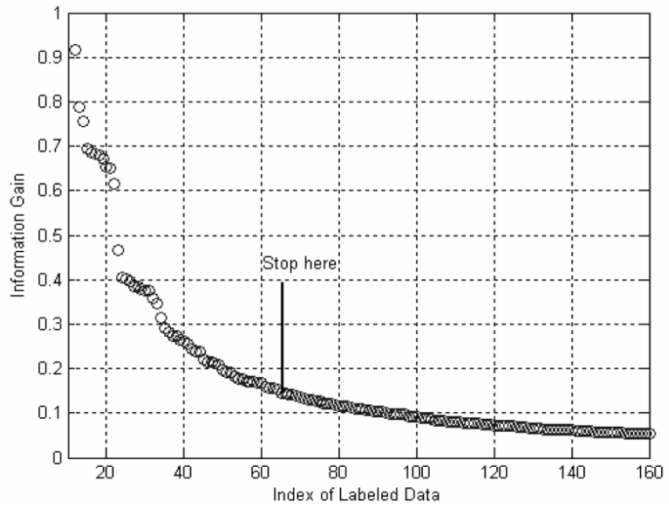

Fig. 6. Information gain of adding a new datum, as a function of the number of the training examples J, selected adaptively.

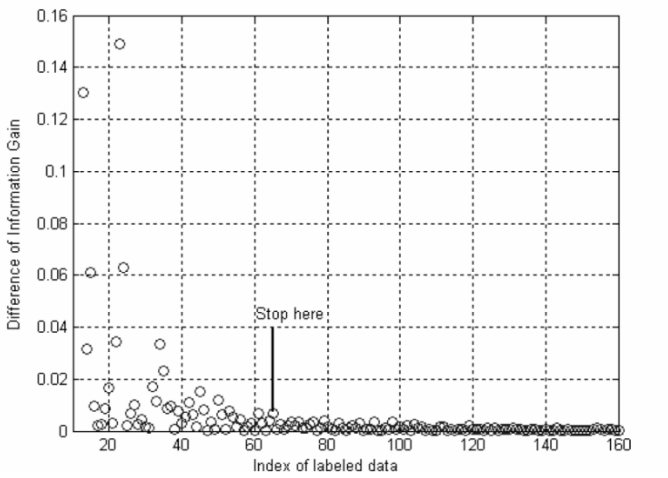

Fig. 7. Difference in the information gain, as a function of the number of training examples J.

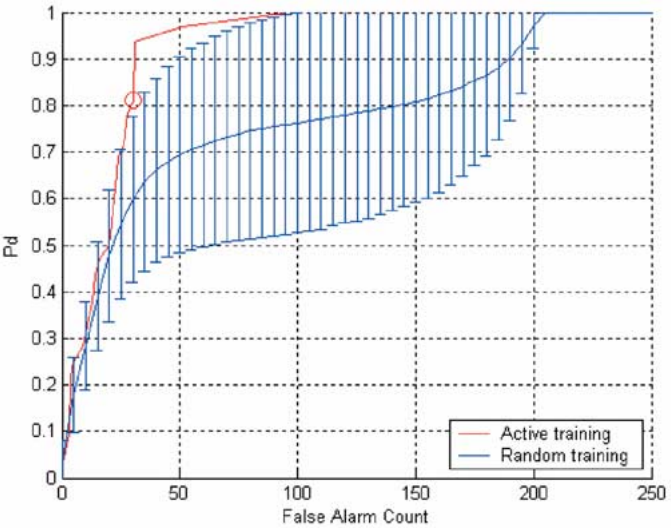

Fig. 8. ROC curves based on $J = 65$ training examples, comparing the adaptive procedure (Section II) to random training data selection. Number of training examples chosen based on Figs. 6 and 7.

presented for $J = 90, 60$, and 40. Using $J = 90$ rather than $J = 128$ results in very little degradation in ROC performance (comparing Figs. 2 and 3), with a slight performance drop for $J = 60$, and a more substantial drop for $J = 40$. It is interesting to note that with decreasing J, the number of test items 300-J increases, therefore increasing the number of false-alarm opportunities. This further highlights the quality of the results in Figs. 3–5, *vis-à-vis* Fig. 2. In all of these and subsequent examples, the size of the basis set $\mathbf{B}_n$ is $n = 10$.

In the above examples, J was specified to be matched to the size of a specified training set, or it was varied for comparison to such. However, the procedure in Section II may be employed to adaptively determine the size of the desired training set $\mathbf{X}_{s,J}$, based on the information gain as J is increased. Specifically, we track $q_n(\mathbf{X}_{s,J}) - q_n(\mathbf{X}_{s,J-1})$ for increasing J, and terminate the algorithm when the information gain is minimal. At this point, adding a new datum to the training dataset does not provide significant additional information to the classifier design.

For the JPG V data, the information gain $q_n(\mathbf{X}_{s,J}) - q_n(\mathbf{X}_{s,J-1})$ is plotted in Fig. 6 as a function of J, and the change in information gain is given in Fig. 7 for visualization assistance. Based on Figs. 6 and 7, the size of the training set is set to $J = 65$. In Fig. 8, results are shown for $J = 65$, with comparison as before to KMP results in which the $J = 65$ training examples are selected randomly. Examining the results in Fig. 8, we observe that the active selection of training data yields a detection probability of approximately 0.95 with approximately 35 false alarms; *on average* one encounters about five times this number of false alarms to achieve the same detection probability (when selecting the training data randomly).

V. Conclusion

There are many remote sensing problems for which one collects data from a given site, and the task is to specify the identity of the object responsible for each signature (e.g., detection and classification). Due to the variability and site-dependent character of many target signatures, it is often difficult to

have reliable training data *a priori* for algorithm design. In this paper, we have therefore developed an information-theoretic framework in which the training data are selected adaptively from the observed site-dependent data, without requiring an *a priori* training set. Specifically, the algorithm specifies those signatures for which knowledge of the associated labels (e.g., target/nontarget) would be most relevant in the context of detector design. An "experiment" is then performed to learn the target labels, where in the context of landmine and UXO sensing this corresponds to excavating the respective buried items. This is a reasonable procedure, since landmines and UXO need be excavated ultimately anyway, and therefore, the algorithm essentially prioritizes the order in which items are excavated, with the goal of ultimately excavating fewer nontargets (false alarms) via proper algorithm training. The algorithm has been demonstrated successfully on measured magnetometer and EMI data from an actual former bombing range, addressing the sensing of UXO.

There are several items that deserve further attention. It was demonstrated that the gain in information content is a good measure of which items should be excavated for learning of associated labels. The results in Figs. 6–8 demonstrated the effectiveness of this procedure, although the actual selection of the number of training examples, J, was determined in a somewhat *ad hoc* manner. Further work is required to make this procedure more rigorous and automated.

In addition, for the results presented here the detection algorithm was trained once using the adaptively determined training set. However, in the subsequent testing phase a "dig list" is specified for those items that are deemed to be associated with targets of interest (here UXO). Once each item is excavated, and the associated label revealed, the algorithm should be successively retrained and applied to the remaining data. The order of the dig list—and therefore the order in which we learn the labels of the testing data—is also of interest, since it may be used to further refine the algorithm sequentially, as a given site is cleaned (e.g., of landmines or UXO).

REFERENCES

[1] C. T. Schroder, W. R. Scott, and G. D. Larson, "Elastic waves interacting with buried land mines: A study using the FDTD method," *IEEE Trans. Geosci. Remote Sensing*, vol. 40, pp. 1405–1415, June 2002.

[2] B. Barrow and H. H. Nelson, "Model-based characterization of electromagnetic induction signatures obtained with the MTADS electromagnetic array," *IEEE Trans. Geosci. Remote Sensing*, vol. 39, pp. 1279–1285, June 2001.

[3] H. H. Nelson and J. R. McDonald, "Multisensor towed array detection system for UXO detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 39, pp. 1139–1145, June 2001.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[5] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[6] C. J. C. Burges, "A Tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[7] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[8] P. Vincent and Y. Bengio, "Kernel matching pursuit," *Mach. Learn.*, vol. 48, pp. 165–187, 2002.

[9] B. Schölkopf, K. K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Trans. Signal Processing*, vol. 45, pp. 2758–2765, Nov. 1997.

[10] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines and Other Kernel Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[11] B. Schölkopf and A. Smola, *Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002.

[12] V. V. Fedorov, *Theory of Optimal Experiments*. New York: Academic, 1972.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[14] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. R. Statist. Soc.*, ser. B, vol. 36, pp. 111–147, 1974.

[15] I. J. Won, D. A. Keiswetter, and D. R. Hanson, "GEM-3: A monostatic broadband electromagnetic induction sensor," *J. Environ. Eng. Geophys.*, vol. 2, pp. 53–64, Mar. 1997.

[16] N. Geng, C. E. Baum, and L. Carin, "On the low-frequency natural response of conducting and permeable targets," *IEEE Trans. Geosci. Remote Sensing*, vol. 37, pp. 347–359, Jan. 1999.

[17] L. Carin, H. Yu, Y. Dalichaouch, A. R. Perry, P. V. Czipott, and C. E. Baum, "On the wideband EMI response of a rotationally symmetric permeable and conducting target," *IEEE Trans. Geosci. Remote Sensing*, vol. 39, pp. 1206–1213, June 2001.

[18] Y. Zhang, L. M. Collins, H. Yu, C. E. Baum, and L. Carin, "Sensing of unexploded ordnance with magnetometer and induction data: Theory and signal processing," *IEEE Trans. Geosci. Remote Sensing*, vol. 41, pp. 1005–1015, May 2003.

Yan Zhang received the B.S., M.S., and Ph.D. degrees in electrical engineering from Jilin University of Technology, Changchun, China, in 1993, 1996, and 1998, respectively.

From January 1999 to July 2004, he was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Duke University, Durham, NC. In 2004, he joined the Innovation Center of Humana, Inc., Louisville, KY, as a Research Scientist. His research interests include statistical signal processing, pattern recognition, and their applications. His current research activity focuses on subsurface target detection and predictive modeling.

Xuejun Liao (SM'04) was born in Qinghai, China. He received the B.S. and M.S. degrees in electrical engineering from Hunan University, China, in 1990 and 1993, respectively, and the Ph.D. degree in electrical engineering from Xidian University, Xi'an, China, in 1999.

From 1993 to 1995, he was with the Department of Electrical Engineering, Hunan University, where he worked on electronic instruments. From 1995 to 2000, he was with the National Key Laboratory for Radar Signal Processing, Xidian University, where he worked on automatic target recognition (ATR) and radar imaging. Since May 2000, he has been working as a Research Associate with the Department of Electrical and Computer Engineering, Duke University, Durham, NC. His current research interests are in Markovian techniques in ATR, blind source separation, sensor scheduling, and machine learning.

Lawrence Carin (SM'96–F'01) was born in Washington, DC, on March 25, 1963. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1985, 1986, and 1989, respectively.

In 1989, he joined the Electrical Engineering Department, Polytechnic University, Brooklyn, NY, as an Assistant Professor, and became an Associate Professor there in 1994. In September 1995, he joined the Electrical Engineering Department, Duke University, Durham, NC, where he became a Professor in 2001. In 2003, he was named the William H. Younger Professor of Engineering at Duke University. He was the Principal Investigator (PI) on a Multidisciplinary University Research Initiative (MURI) on demining (1996–2001) and is currently the PI of a MURI dedicated to multimodal inversion. His current research interests include short-pulse scattering, subsurface sensing, and wave-based signal processing.

Dr. Carin is an Associate Editor of the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION and is a member of Tau Beta Pi and Eta Kappa Nu.