ACTIVE SELECTION OF LABELED DATA FOR TARGET DETECTION

Yan Zhang, Xuejun Liao, Esther Dura and Lawrence Carin

Department of Electrical and Computer Engineering
Duke University, Durham NC, 27708, USA

ABSTRACT

An information-theoretic approach is developed for target detection, with active selection of training set, directly from the site-specific measured data. For the proposed kernel-based algorithm, a set of basis functions are defined first to characterize the signature distribution of the site, then we determine a parsimonious set of data, for which knowledge of the associated labels would be most informative to determine the weights for the basis functions. Both of them utilize the Fisher information criteria. The proposed framework is applied to subsurface target detection, with example results presented for an actual buried unexploded ordnance site.

1. INTRODUCTION

In many target-detection problems, the target signatures are a strong function of environmental and historical circumstances. For example, in landmine and unexploded ordnance (UXO) sensing, the task is strongly influenced by which ordnance are present, on how the ordnance impacted the soil, and on the surrounding conducting clutter and UXO fragments. Therefore, for algorithm-training purposes, it is difficult to define a set of target signatures that are generally representative for different sites. In this paper we investigate a technique whereby detection and classification algorithms may be designed without requiring a separate training set of representative target and clutter signatures.

Let $\{\mathbf{x}_i\}_{i=1,N}$ represent the measured signatures of the N 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/non-target), to be determined in the detection phase. We here develop a kernel-based classifier, by which an observed feature vector $\mathbf{x}$ is classified using the function

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

where $\mathbf{b}_i$ is the center of the *ith* basis function, w_i is the associated weight, 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 $\mathbf{b}_i$. Algorithms that utilize the form in (1) include the support vector machine (SVM) [1], and many other related algorithms [2-4]. The $\mathbf{b}_i$ and w_i in (1) are typically estimated from a separate training set, for which the associated labels y_i are known. However, the variability of target signatures makes the idea of utilizing a separate training set undesirable and often impractical.

In the approach proposed 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.* $\mathbf{B}_n \subset \mathbf{X}$. 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. Note that the labels of the objects associated with $\mathbf{B}_n$ are not required at this point. Having defined the basis set for (1), we require labeled data to determine the associated model weights $\{w_i\}_{i=1,n}$ and w_o (denoted collectively by the vector $\mathbf{w}$). Then 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 $\mathbf{X}_s$ is again determined via information-theoretic metrics. Note that the sets $\mathbf{B}_n$ and $\mathbf{X}_s$ may overlap, but they are in general distinct. The determined algorithm is thereafter applied to $\mathbf{x} \notin \mathbf{X}_s$. The key point is that the training set $(\mathbf{X}_s,\mathbf{L}_s)$ is selected adaptively on the observed site-dependent data, via fundamental information-theoretic metrics, and therefore no *a priori* training data required.

The proposed approach has been applied to both subsurface UXO detection and underwater mine detection. With measured sensor data from actual test sites, we observe that the false-alarm rate is significantly reduced, as a result of the fact that the adaptively designed algorithm is well matched to the environment. In this paper, only UXO results are presented due to space limitation. At the conference we will also demonstrate results for sonar underwater mine detection.

2. ACTIVE CLASSIFIER DESIGN

The decision function in (1), using n basis functions, may be expressed concisely as [3]

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

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

$$\mathbf{w}_n = [w_{n,0}, w_{n,1}, w_{n,2}, \cdots, w_{n,n}]^T \qquad (4)$$

Assume that the item associated with signature $\mathbf{x}_i$ is recovered (this is termed an "experiment"), from which we learn the associated label y_i, where by construction y_i=1 for target and y_i=-1 for no-target or clutter. The label recovered by the experiment is related to the prediction $f_n(\mathbf{x})$ by

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

where $\varepsilon_i = \varepsilon(\mathbf{x}_i)$ is the error term resulting from imperfections in the model. In algorithm design one desires the decision function $f_n(\mathbf{x})$ that minimizes the error observed on training data, for which the data and labels are known. If the training data is well matched to the subsequent testing data, then the algorithm is

likely to constitute a robust detection procedure. However, it is impractical to have a separate training set in many target-detection problems.

2.1 Selection of Basis Functions

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

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

where $\boldsymbol{\phi}_{n,i} \equiv \boldsymbol{\phi}_n(\mathbf{x}_i)$. 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 [5], the Fisher information matrix in (6) is associated with the errors in fitting the model to all N measured $\mathbf{x}_i$, using the basis $\mathbf{B}_n$. By appending a new basis function to $\boldsymbol{\phi}_n(\cdot)$, one obtains

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

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 $\boldsymbol{\phi}_{n+1}$ the augmented classifier f_{n+1}, for which the Fisher information matrix is found to be

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

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 the (n+1)-dimensional basis $\mathbf{B}_{n+1}$, *vis-à-vis* the n-dimensional basis $\mathbf{B}_n$ in (6).

Among many ways of comparing the information content reflected by $\mathbf{M}_n$ and $\mathbf{M}_{n+1}$, we here employ the so-called D-optimal procedure [5], defined as the determinant of the information matrix. The logarithm of the determinant of $\mathbf{M}$ is denoted q_n, and it can be shown that

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

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

Since $N \geq n$ and n of the vectors $\{\boldsymbol{\phi}_n(\mathbf{x}_i)\}_{i=1,N}$ are linearly independent, the matrix $\mathbf{M}_n$ is full rank and its inverse exists. Under these conditions, it can be shown that $r > 0$, and therefore $\ln r$ in (9) is generally valid.

It is known from information theory [6] 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. 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

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

Basis elements $\mathbf{b}_n$ are added until the information gain reflected in q_{n+1}-q_n is no longer deemed significant. Note that evaluation of (11) does not require knowledge of the target labels y_i.

2.2. Selection of Labeled Data for Model Training

Labeled data are required to optimize the associated model weights $\mathbf{w}$. 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 recovered 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 then 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 content 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

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

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

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

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)

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

with

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

Care is needed with regard to evaluating the inverse of $\mathbf{M}_n$, since if $J<n$ the matrix is rank deficient. 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 in examples presented in Sec. 3. To define $\mathbf{x}_{i_{J+1}}$, one iteratively maximizes $\ln \rho(\mathbf{x}_{i_{J+1}})$ to obtain

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

We again do not require the signature labels. The elements of $\mathbf{X}_s$ are selected iteratively, in a "greedy" fashion 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 recovered, yielding the labels $\mathbf{L}_{s,J}$.

For the assumptions underlying the linear model in (5) and that $\varepsilon(\mathbf{x}_i)$ is i.i.d. over the set of i, the optimal estimation for the weights $\mathbf{w}$ with knowledge of $\mathbf{B}_n$ and ($\mathbf{X}_{s,J}$, $\mathbf{L}_{s,J}$) is expressed as

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

where $\mathbf{y}$ represents the set of labels determined via the J experiments

$$\mathbf{y} = \{y_{i_1}, y_{i_2}, ..., y_{i_J}\}^{\mathrm{T}} \tag{18}$$

and the $J \times (n+1)$ matrix $\mathbf{\Phi}$ is defined as

$$\mathbf{\Phi} = [\boldsymbol{\phi}_n(\mathbf{x}_{i_1}) \quad \boldsymbol{\phi}_n(\mathbf{x}_{i_2}) \quad \cdots \quad \boldsymbol{\phi}_n(\mathbf{x}_{i_J})]^{\mathrm{T}} \tag{19}$$

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 receiver operating characteristic (ROC).

3. 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 demonstrate the detection results of buried UXO for data collected at an actual UXO site: Jefferson Proving Ground (JPG) in the United States. 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. If at the desired detection probability, the false-alarm rate is reduced, then overall cleanup costs may diminish substantially. One principal challenge in UXO sensing is development of a training set, for design of the detection algorithm.

3.1. Magnetometer and Electromagnetic Induction Data at JPG

Magnetometer and electromagnetic induction (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. An EMI sensor actively transmits a time-varying electromagnetic field, and consequently senses the dynamic induced secondary field from the target. We here employ a frequency-domain EMI sensor that transmits and senses at several discrete frequencies. Parametric models have been developed for both magnetometer and EMI sensors [7,8]. 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". Details on the magnetometer and EMI models, and on the model-fitting procedure, may be found in [8].

Jefferson Proving Ground 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. Our results are presented with the GEM-3 (an EMI sensor) and magnetometer data from two adjoining areas, constituting a total of approximately five acres.

This test was performed with US Army oversight. One of the two JPG areas was assigned as the training area, for which the ground truth or labels (UXO/non-UXO) were 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 most UXO types were found in equal number in each of the two areas. This indicates an artificial effort to match the training data to the detection data in this demonstration, which is not always feasible in practice. There are 300 potential targets detected from sensor anomalies after the model fitting based prescreening, 40 of which are proven to be UXO and the others are clutter. The excavated UXO items include 10 different types. In the training area, there are 128 buried items, 16 of which are UXO.

3.2 Detection Results

We present ROC curves using the adaptive-training approach developed in Sec. 2, 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), which here is determined iteratively using kernel matching pursuits (KMP). Details on the KMP algorithm may be found in [3]. To make the comparison appropriate, both the adaptive training and KMP implementation employ the radial basis function (RBF) kernel [4] with variance adaptively adjusted by the algorithms.

In the first example, to be consistent with the size of the training area specified in the JPG V test, the adaptive technique in Sec. 2 is employed to select J=128 items from the original 300. The 128 "recovered" labels are utilized to build the classifier. Then the adaptive learning algorithm is tested on the remaining 172 items. The basis set $\mathbf{B}_n$ is also defined adaptively using the original 300 signatures. Here the number of n is automatically determined to be 10 via the information gain criterion, and consequently utilized for all results. Performance comparisons are shown in Fig. 1, wherein we present results for active data selection, KMP results using the assigned 128 training examples, and average results for randomly choosing the 128 examples for KMP training. For the latter case, 100 random selections were performed, and we place error bars on the results. The length of the error bar is twice the standard derivation of the Pd (detection probability) for the associated false-alarm count. We observe from the results in Fig. 1 that the active data selection procedure produces the best ROC results for Pd>0.7, which is of most interest in practice. The average performance based on choosing the training set randomly is substantially below the other two, with significant variability reflected in the error bars. These results demonstrate the power of the developed active-data-selection algorithm, and also that the training data defined for JPG V is well matched to the testing data.

The active training algorithm in Sec. 2 has been implemented with several smaller values of J down to 40, reflecting less cost for determination of target labels required in the training phase. For all cases, the performance of the active training technique upper bounds the random training selections. We only show the results of one example, where J is determined adaptively from the procedure in Sec. 2. 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. The information gain $q_n(\mathbf{X}_{s,J}) - q_n(\mathbf{X}_{s,J-1})$ is plotted in Fig. 2(a) as a function of *J*, and the change in information gain is given in Fig. 2 (b) for visualization assistance. Based on Fig. 2 the size of the training set is set to *J*=65. In Fig. 3 results are shown for *J*=65, with comparison as before to KMP results in which the *J*=65 training examples are selected randomly. From Fig. 3, 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.

4. CONCLUSIONS

Due to the variability and site-dependent character of 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 is selected adaptively from the observed site-dependent data, without requiring an *a priori* training set. The algorithm specifies those signatures for which knowledge of the associated labels (*e.g.* target/non-target) would be most relevant in the context of detector design. An "experiment" is then performed to learn the target labels. This is a reasonable procedure in many applications, including landmines/UXO detection where targets 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 non-targets via proper algorithm training.

5. REFERENCES

[1] V. Vapnik, Statistical Learning Theory, John Wiley, New York, 1998.

[2] M. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," Journal of Machine Learning Research, vol. 1, pp. 211-244, 2001.

[3] P. Vincent and Y. Bengio, "Kernel matching pursuit," Machine Learning, vol. 48, pp. 165-187, 2002.

[4] N. Cristianini and J. Shawe-Taylor, Support Vector Machines and Other Kernel Based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.

[5] V. V. Fedorov, Theory of Optimal Experiments, Academic Press, New York, 1972.

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

[7] 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. Geosc. Remote Sens., Vol. 39. pp. 1206 -1213, June 2001

[8] 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.

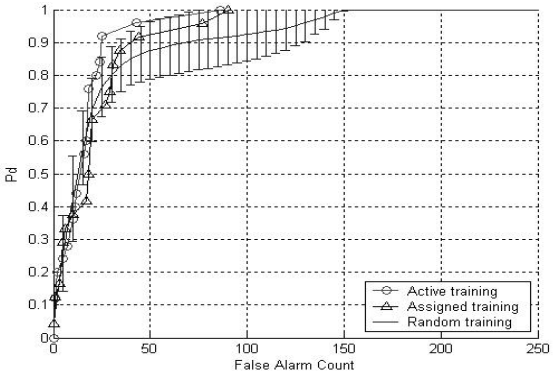

Fig. 1. ROC curves based on *J* = 128 training examples.

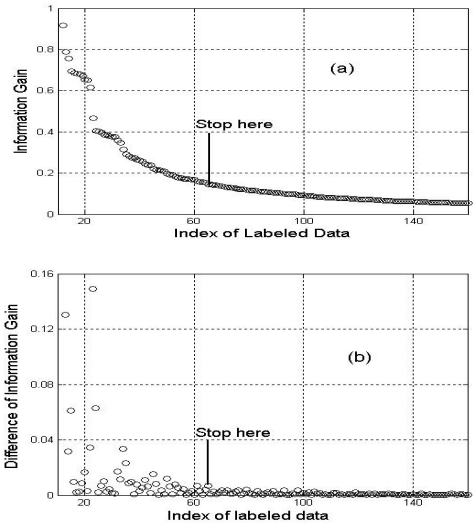

Fig. 2. Information gain of adding a new datum (a), and difference in the information gain (b), as a function of the number of the training examples *J*.

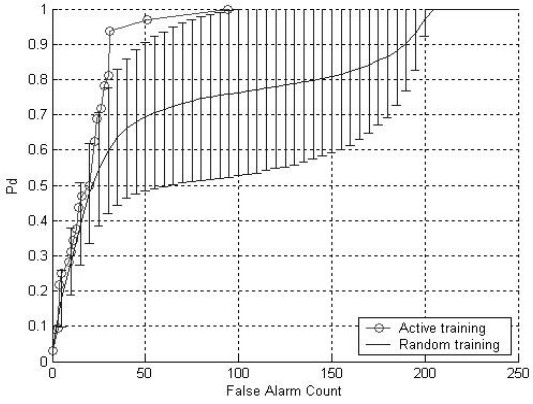

Fig. 3. ROC curves based on *J* = 65 training examples, number of training examples chosen based on Fig. 2.