# Active Data Acquisition with Incomplete Data

David Williams, Xuejun Liao, and Lawrence Carin

Duke University

Department of Electrical and Computer Engineering

Technical Report

September 2005

**Abstract**

We present a unified framework under which active data acquisition can be performed. This comprehensive framework allows for the acquisition of both labels and features. Moreover, several types of feature acquisition are permitted, including the acquisition of individual or multiple features for individual or multiple data points, which may be either labeled or unlabeled. The algorithm chooses to acquire that data for which the expected benefit — defined as the cost of acquiring the data subtracted from the expected reduction in misclassification costs if the data is possessed — is a maximum. The algorithm automatically determines the most beneficial type of data to acquire when multiple options exist. The framework also has a natural, intuitive criterion for terminating the active data acquisition process: when the expected benefit of all possible acquisitions is no longer positive. We also present a classifier that utilizes incomplete data, which is then employed in the proposed active data acquisition framework. Experimental results demonstrate the superiority of the proposed approach over random data acquisition.

## I. Introduction

The incomplete-data problem, in which certain features are missing for particular data points, exists in a wide range of fields, including social sciences, computer vision, biological systems, and remote sensing. In multi-sensor remote-sensing applications, incomplete data can result when only a subset of physical sensors (*e.g.*, radar, infrared, acoustic) are deployed at certain regions. Increasing focus in the future on using (and fusing data from) multiple sensors [1], "views" [2], or information sources [3, 4] will make such incomplete-data problems commonplace.

In many applications involving incomplete data, it is possible to acquire the missing data at a cost. In multi-sensor remote-sensing applications, data is acquired by deploying sensors to data points. (In general, we define a "sensor" to be a data-collecting instrument that produces a group of features.) In a medical diagnostic task, deciding which medical tests to administer would be equivalent to deciding which missing data to acquire. In protein classification, one can acquire

different types of data, such as protein sequences or gene expression data.

Acquiring data is usually an expensive, time-consuming task, a fact that necessitates an *active feature acquisition* process. In contrast to conventional active learning (*e.g.*, [5–7]), which selects the most beneficial labels to acquire, this process would intelligently select the most beneficial missing *features* to acquire. The major flaws that plague the scant heuristic active-feature-acquisition approaches in the literature [8–11] stem from two sources: the reliance on complete-data classification algorithms, and the disregard of data acquisition costs. In fact, if no costs are incurred by acquiring additional data, the active acquisition process is moot because all data should be collected. By ignoring the acquisition costs, the methods [8–11] are also forced to adopt *ad hoc* criteria to terminate the active feature acquisition process (*e.g.*, after a fixed, *a priori* number of features have been acquired).

The reliance on complete-data classification algorithms[1] exposes several other shortcomings of methods [8–10] in the active data acquisition literature. First, since only complete data can be handled, not all available data is utilized. Moreover, this implicit requirement that a sufficiently large number of data points are complete (*i.e.*, missing no features) is potentially crippling because it is possible for all data points to be incomplete. Furthermore, these methods [8–10] (unrealistically) require complete testing data; the methods do not apply if the testing data is missing features. Also, the only capability these methods [8–10] have is to consider acquiring *all* missing features for *single* data points. In general, many different types of data can be acquired: perfect or imperfect labels, as well as individual or multiple features for individual or multiple, labeled or unlabeled, data points. In practice, the application under investigation will usually dictate which types of data

[1] An incomplete-data classifier functions with missing data. We deem a classifier that explicitly imputes or completes missing data to *not* be an incomplete-data classifier.

acquisition are feasible.

In this work, we present an incomplete-data classification algorithm that is subsequently utilized in our proposed active data acquisition framework. The proposed active data acquisition approach does not suffer from any of the numerous limitations that plague the existing methods [8–11] in the literature. Our framework accounts for acquisition costs and has a natural, intuitive termination criterion. Moreover, if different types of data acquisition are feasible for a given application (see Figure 1), our framework automatically determines the most beneficial type of data to acquire.

The remainder of this paper is organized as follows. In Section II, we present a logistic classifier that handles incomplete data (and imperfect labels). This classifier is subsequently utilized in our active data acquisition framework presented in Section III. Experimental results of this active data acquisition framework appear in Section IV. Section V consists of a discussion. Concluding remarks are made in Section VI.

## II. Classification of Incomplete Data

The work in this paper assumes that the missing data is either missing completely at random (MCAR) or missing at random (MAR), meaning that the missing data is independent of its value (see [12, 13] for more details).

Assume we have a set of labeled incomplete data

$$\mathcal{D}_L = \{(\mathbf{x}_i, y_i, \epsilon_i, m_i) : \mathbf{x}_i \in \mathbb{R}^d, x_{ia} \quad \text{missing} \quad \forall a \in m_i\}_{i=1}^{N_L} \tag{1}$$

where $\mathbf{x}_i$ is the $i$-th data point, labeled as $y_i \in \{-1, 1\}$ with known labeling error rate $\epsilon_i \in [0, 0.5)$; the features in $\mathbf{x}_i$ indexed by $m_i$ (*i.e.,* $x_{ia}, a \in m_i$) are missing. Each $\mathbf{x}_i$ has its own (possibly unique) set of missing features, $m_i$. One special case occurs when a subset of data share common

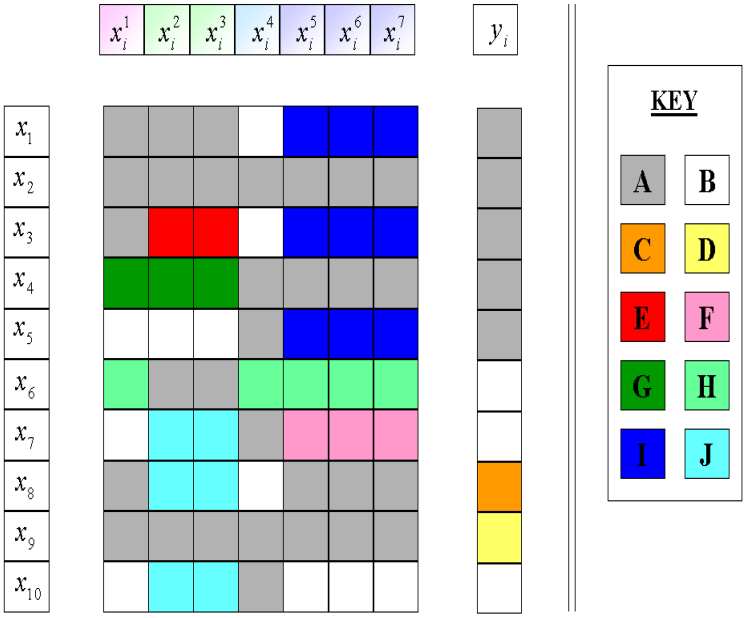


Fig. 1. Illustration of the different types of active data acquisition that are possible. Each row corresponds to one data point, while each column corresponds to one feature. The final $y_i$ column corresponds to the labels. Shown is a four sensor case, where the four sensors produce 1, 2, 1, and 3 features, respectively (note the shading in the top row). Thus, for example, the two features corresponding to the second sensor — $\mathbf{x}_i^2$ and $\mathbf{x}_i^3$ — are always either both observed or both missing. Often, data sets do not conform to the "sensor" paradigm; in such cases each individual feature can be viewed as being produced by a unique "sensor." In the figure, case $A$ corresponds to data that is observed, while case $B$ corresponds to data that is missing. All other data (*i.e.*, cases $C$–$J$) was initially missing before active data acquisition transpired. Case $C$ corresponds to the acquisition of a perfect label for an unlabeled data point, while case $D$ corresponds to the acquisition of an imperfect label for an unlabeled data point. The acquisition of the features from a single sensor that a single data point is missing, when the data point is labeled (resp. unlabeled), corresponds to case $E$ (resp. $F$). The acquisition of the features from all sensors that a single data point is missing, when the data point is labeled (resp. unlabeled), corresponds to case $G$ (resp. $H$). The acquisition of the features from the same one sensor that all labeled (resp. unlabeled) data points are missing corresponds to case $I$ (resp. $J$).

missing features, as with multi-sensor data where the common missing features result from a sensor that has not collected data.

In logistic regression (with a hyperplane classifier), the probability of label $y_i$ given $\mathbf{x}_i$ is $p\left(y_i|\mathbf{x}_i,\mathbf{w}\right) = \sigma(y_i\mathbf{w}^T\mathbf{x}_i)$, where $\sigma(\nu) = (1+\exp(-\nu))^{-1}$ is the sigmoid function and $\mathbf{w}$ constitutes a classifier. Accounting for imperfections in the labeling process arising from a known labeling error rate $\epsilon_i$, the probability of label $y_i$ given $\mathbf{x}_i$ and $\epsilon_i$ is [14]

$$p\left(y_i|\mathbf{x}_i,\epsilon_i,\mathbf{w}\right) = \epsilon_i + (1-2\epsilon_i)\sigma(y_i\mathbf{w}^T\mathbf{x}_i). \tag{2}$$

The labeling error rate is simply the probability that a true label was flipped (corrupted) to the wrong label (*e.g.*, $\{y_i^{\text{true}} = 1\} \rightarrow \{y_i = -1\}$). Imperfect labels may be an unavoidable fact (*e.g.*, human-labeled web pages), or simply a preferred shortcut (*e.g.*, a radiologist's tumor diagnosis with some level of confidence in lieu of an invasive biopsy). Note that the standard case of perfect labels is recovered when $\epsilon_i = 0$.

We partition $\mathbf{x}_i$ into its observed and missing parts, $\mathbf{x}_i = [\mathbf{x}_i^{o_i};\mathbf{x}_i^{m_i}]$ where $\mathbf{x}_i^{o_i} = [x_{ia}, a \in o_i]^T$, $\mathbf{x}_i^{m_i} = [x_{ia}, a \in m_i]^T$, and $o_i = \{1,\cdots,d\} \setminus m_i$ is the (complementary) set of observed features in $\mathbf{x}_i$. We apply the same partition to $\mathbf{w}$ to obtain $\mathbf{w} = [\mathbf{w}_{o_i};\mathbf{w}_{m_i}]$, yielding

$$p\left(y_i|\mathbf{x}_i,\epsilon_i,\mathbf{w}\right) = \epsilon_i + (1-2\epsilon_i)\sigma(y_i(\mathbf{w}_{o_i}^T\mathbf{x}_i^{o_i} + \nu_i)) \tag{3}$$

where $\nu_i = \mathbf{w}_{m_i}^T\mathbf{x}_i^{m_i}$. Because $\mathbf{x}_i^{m_i}$ (and hence $\nu_i$) is missing, (3) cannot be evaluated. By integrating out the missing data $\mathbf{x}_i^{m_i}$, the needed probability of $y_i$ given the observed features $\mathbf{x}_i^{o_i}$ can be written as

$$\begin{aligned} p\left(y_i|\mathbf{x}_i^{o_i},\epsilon_i,\mathbf{w}\right) &= \int p\left(y_i|\mathbf{x}_i^{m_i},\mathbf{x}_i^{o_i},\epsilon_i,\mathbf{w}\right) p\left(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i}\right) d\mathbf{x}_i^{m_i} & (4)\\ &= \epsilon_i + (1-2\epsilon_i)\int \sigma(y_i(\mathbf{w}_{o_i}^T\mathbf{x}_i^{o_i} + \nu_i)) p\left(\nu_i|\mathbf{x}_i^{o_i}\right) d\nu_i. & (5) \end{aligned}$$

It is important to note that the integral in (4) is in general multi-dimensional, while the integral in (5) is one-dimensional. The integration in (5) can be performed by making two minor assumptions.

First, we assume that $p\left(\mathbf{x}_i\right)$ can be modeled as a Gaussian mixture model (GMM):

$$p(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left( \begin{bmatrix} \mathbf{x}_i^{o_i} \\ \mathbf{x}_i^{m_i} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_k^{o_i} \\ \boldsymbol{\mu}_k^{m_i} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_k^{o_i o_i} & (\boldsymbol{\Sigma}_k^{m_i o_i})^T \\ \boldsymbol{\Sigma}_k^{m_i o_i} & \boldsymbol{\Sigma}_k^{m_i m_i} \end{bmatrix} \right) \tag{6}$$

where the $\pi_k$ are the non-negative mixture weights that sum to unity; necessarily $p\left(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i}\right)$ is a GMM as well. The Expectation-Maximization (EM) [15] and Variational Bayesian EM (VB-EM) [16, 17] formulations that can be used to build the required GMM with incomplete data have been developed previously [18].

Because of the linear relation $\nu_i = \mathbf{w}_{m_i}^T \mathbf{x}_i^{m_i}$, $p\left(\nu_i|\mathbf{x}_i^{o_i}\right)$ is also a GMM,

$$p\left(\nu_i|\mathbf{x}_i^{o_i}\right) = \sum_{k=1}^{K} \delta_k^i \mathcal{G}\left(\frac{\nu_i - \zeta_k^i}{\alpha_k^i}\right), \tag{7}$$

with the parameters

$$\delta_k^i = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_i^{o_i}; \boldsymbol{\mu}_k^{o_i}, \boldsymbol{\Sigma}_k^{o_i o_i}\right)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}\left(\mathbf{x}_i^{o_i}; \boldsymbol{\mu}_\ell^{o_i}, \boldsymbol{\Sigma}_\ell^{o_i o_i}\right)} \tag{8}$$

$$\zeta_k^i = \mathbf{w}_{m_i}^T \boldsymbol{\xi}_k^{m_i} \tag{9}$$

$$\alpha_k^i = \sqrt{\mathbf{w}_{m_i}^T \boldsymbol{\Omega}_k^{m_i} \mathbf{w}_{m_i}} \tag{10}$$

$$\boldsymbol{\xi}_k^{m_i} = \boldsymbol{\mu}_k^{m_i} + \boldsymbol{\Sigma}_k^{m_i o_i} (\boldsymbol{\Sigma}_k^{o_i o_i})^{-1} (\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i}) \tag{11}$$

$$\boldsymbol{\Omega}_k^{m_i} = \boldsymbol{\Sigma}_k^{m_i m_i} - \boldsymbol{\Sigma}_k^{m_i o_i} (\boldsymbol{\Sigma}_k^{o_i o_i})^{-1} (\boldsymbol{\Sigma}_k^{m_i o_i})^T \tag{12}$$

where

$$\mathcal{G}(\nu_i) = (2\pi)^{-1/2} \exp\left\{-\nu_i^2/2\right\} \tag{13}$$

is the standard univariate Gaussian density function with zero mean and unit variance.

The second assumption is that the sigmoid function can be approximated as a probit function (*i.e.,* a Gaussian cumulative distribution function)

$$\sigma(\alpha) \approx \int_{-\infty}^{\alpha} \mathcal{G}\left(\frac{z}{\beta}\right) dz \tag{14}$$

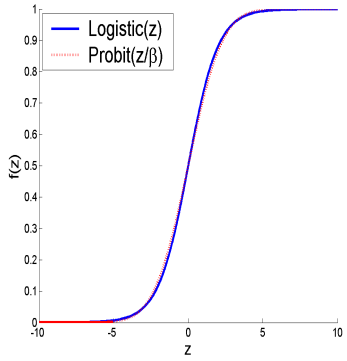


Fig. 2. Illustration of the validity of the approximation made between the logistic function and the (scaled) probit function.

where $\beta = \frac{\pi}{\sqrt{3}}$. The accuracy of this approximation is shown in Figure 2.

Substituting (7) and (14) into (5), we obtain

$$
\begin{aligned}
& p\left(y_i|\mathbf{x}_i^{o_i}, \epsilon_i, \mathbf{w}\right) \\
&\approx \epsilon_i + (1-2\epsilon_i) \int\int_{-\infty}^{y_i(\mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i} + \nu_i)} \mathcal{G}\left(\frac{z}{\beta}\right) dz \sum_{k=1}^{K} \delta_k^i \mathcal{G}\left(\frac{\nu_i - \zeta_k^i}{\alpha_k^i}\right) d\nu_i && (15) \\
&= \epsilon_i + (1-2\epsilon_i) \int\int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} \mathcal{G}\left(\frac{z' + y_i\nu_i}{\beta}\right) dz' \sum_{k=1}^{K} \delta_k^i \mathcal{G}\left(\frac{\nu_i - \zeta_k^i}{\alpha_k^i}\right) d\nu_i && (16) \\
&= \epsilon_i + (1-2\epsilon_i) \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} \int \mathcal{G}\left(\frac{z' + y_i\nu_i}{\beta}\right) \mathcal{G}\left(\frac{y_i\nu_i - y_i\zeta_k^i}{y_i\alpha_k^i}\right) d\nu_i \, dz' && (17) \\
&= \epsilon_i + (1-2\epsilon_i) \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} \mathcal{G}\left(\frac{z' + y_i\zeta_k^i}{\sqrt{(y_i\alpha_k^i)^2 + \beta^2}}\right) dz' && (18) \\
&= \epsilon_i + (1-2\epsilon_i) \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} \mathcal{G}\left(\frac{z' + y_i\zeta_k^i}{\beta} \frac{\beta}{\sqrt{(\alpha_k^i)^2 + \beta^2}}\right) dz' && (19) \\
&= \epsilon_i + (1-2\epsilon_i) \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{\frac{y_i\beta(\mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i} + \zeta_k^i)}{\sqrt{(\alpha_k^i)^2 + \beta^2}}} \mathcal{G}\left(\frac{z}{\beta}\right) dz && (20) \\
&\approx \epsilon_i + (1-2\epsilon_i) \sum_{k=1}^{K} \delta_k^i \, \sigma\left(\frac{y_i\beta(\zeta_k^i + \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i})}{\sqrt{(\alpha_k^i)^2 + \beta^2}}\right). && (21)
\end{aligned}
$$

In the derivation leading to (21), (16) results from the change of variable $z' = z - y_i\nu_i$; (17) is due to exchanging the order of integrals and summation; (18) results because the convolution

of two Gaussians is a Gaussian; (19) holds because $y_i^2 = 1$; (20) results from the change of variable $z = \frac{\beta(z' + y_i \zeta_k^i)}{\sqrt{(\alpha_k^i)^2 + \beta^2}}$; and (21) is obtained by reverting to sigmoid representation. Thus we have expressed $p\left(y_i | \mathbf{x}_i^{o_i}, \epsilon_i, \mathbf{w}\right)$ as a mixture of *sigmoids*.

Substituting (9) and (10) into (21), we obtain the probability of $y_i$ given only the observed portion of $\mathbf{x}_i$:

$$p\left(y_i | \mathbf{x}_i^{o_i}, \epsilon_i, \mathbf{w}\right) = \epsilon_i + (1 - 2\epsilon_i) \sum_{k=1}^{K} \delta_k^i \, \sigma\left( \frac{y_i \beta (\mathbf{w}_{m_i}^T \boldsymbol{\xi}_k^{m_i} + \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i})}{\sqrt{\mathbf{w}_{m_i}^T \boldsymbol{\Omega}_k^i \mathbf{w}_{m_i} + \beta^2}} \right). \tag{22}$$

For the incomplete and possibly imperfectly labeled data in (1), assuming the data points are independent of each other results in the log-likelihood function

$$\begin{aligned} \ell(\mathbf{w}) &= \log p\left( \{y_i\}_{i=1}^{N_L} | \{\mathbf{x}_i^{o_i}\}_{i=1}^{N_L}, \{\epsilon_i\}_{i=1}^{N_L}, \mathbf{w} \right) \\ &= \sum_{i=1}^{N_L} \log \left[ \epsilon_i + (1 - 2\epsilon_i) \sum_{k=1}^{K} \delta_k^i \, \sigma \left( \frac{y_i \beta (\mathbf{w}_{m_i}^T \boldsymbol{\xi}_k^{m_i} + \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i})}{\sqrt{\mathbf{w}_{m_i}^T \boldsymbol{\Omega}_k^i \mathbf{w}_{m_i} + \beta^2}} \right) \right]. \end{aligned} \tag{23}$$

The objective function (23) to be maximized is not concave for two reasons. First, the concavity is destroyed by the imperfect labels resulting from $\epsilon_i$. Even in the case of perfect labels though, (23) is not concave because of the particular form of the argument of the sigmoid function, arising from the incomplete data. Since (23) is not concave, the solution may get trapped in local maxima. A good initialization is important, so we initialize $\mathbf{w}$ as follows. We "complete" the data set by replacing the missing features $\mathbf{x}_i^{m_i}$ with the conditional mean $\mathbb{E}[\mathbf{x}_i^{m_i} | \mathbf{x}_i^{o_i}] = \sum_{k=1}^{K} \delta_k^i \boldsymbol{\xi}_k^{m_i}$, where $\delta_k^i$ and $\boldsymbol{\xi}_k^{m_i}$ are defined in (8) and (11), respectively. For the initialization, we also treat all labels as perfect, artificially setting all $\epsilon_i = 0$. This "completed," "perfectly" labeled data set is submitted to the standard logistic regression to obtain $\mathbf{w}_0$, which is then used as the initialization of $\mathbf{w}$ in maximizing (23) by gradient ascent.

Thus, the maximum-likelihood (ML) logistic regression classifier $\mathbf{w}$, in the presence of missing

data (and imperfect labels), is obtained. Thereafter, the class predictions of an unlabeled testing data point with incomplete (missing) features is computed trivially using (22) (with $\epsilon_i = 0$ since no actual labeling will have transpired).

The versatile nature of this classification algorithm, which handles incomplete data and imperfect labels, is vital for the active data acquisition framework. Specifically, the classifier permits every realistic type of data acquisition to be performed and handled under a single unified framework.

## III. Active data acquisition with incomplete data

### A. Approach

Here we introduce a comprehensive framework for active data acquisition, where the data acquired can be either labels or features. In addition to the costs involved with acquiring data, there is a different cost due to the misclassification of data, called the *risk*. It is sensible to perform active data acquisition only when the costs of acquiring additional data are outweighed by the benefit accrued. This benefit is simply the reduction in risk resulting from possessing the new data. Thus, the logical objective that should drive active data acquisition is to maximize the expected benefit derived from acquiring additional data. It should be noted that the acquisition and misclassification costs must be in the same units.

In addition to the labeled data in (1), assume we have a set of unlabeled incomplete data

$$\mathcal{D}_U = \{(\mathbf{x}_i, m_i) : \mathbf{x}_i \in \mathbb{R}^d, x_{ia} \quad \text{missing} \quad \forall a \in m_i\}_{i=N_L+1}^{N}. \tag{24}$$

Let the (estimated) risk[2] $R(\mathcal{D}_U|\mathcal{D})$ be the risk on the unlabeled data $\mathcal{D}_U$, from using a classifier

[2]This risk on the unlabeled data is the *estimated* risk because the true labels (which are unavailable) must be known to compute the *true* risk.

designed using $\mathcal{D}_L$; it is defined as

$$R(\mathcal{D}_U|\mathcal{D}_L) = \sum_{i=N_L+1}^{N} \min\{C_{[1,-1]}p(y_i = -1|\mathbf{x}_i^{o_i}, \mathbf{w}), C_{[-1,1]}p(y_i = 1|\mathbf{x}_i^{o_i}, \mathbf{w})\} \tag{25}$$

where the weights $\mathbf{w}$ are trained using $\mathcal{D}_L$, and $C_{[a,b]}$ is the cost of misclassifying a data point as belonging to class $a$ instead of the true class $b$. Let the expected risk on the unlabeled data $\mathcal{D}_U$ after acquiring new data $\mathcal{D}_*$ — which can be features or a label (perfect or imperfect) — be $\mathbb{E}_{\mathcal{D}_*}[R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*)]$. The expected benefit of acquiring data $\mathcal{D}_*$ is then the cost of acquiring the data subtracted from the expected decrease in (estimated) risk:

$$\mathbb{E}_{\mathcal{D}_*}[B(\mathcal{D}_*|\mathcal{D})] = \{R(\mathcal{D}_U|\mathcal{D}_L) - \mathbb{E}_{\mathcal{D}_*}[R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*)]\} - C(\mathcal{D}_*) \tag{26}$$

where $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$. When formulated as in (26), the active data acquisition process has a natural termination criterion: when the expected benefit of all possible acquisitions is no longer positive.[3] It should be noted that $\mathbf{w}$ within (25) is learned by using the incomplete-data classifier from Section II. This classifier — capable of handling incomplete data and imperfect labels — permits the proposed comprehensive, unified active data acquisition framework.

*1) Label Acquisition:* If the new data is a label ($\mathcal{D}_* = y_*$), the requisite expectation in (26) can be performed analytically since there are only a finite number — namely two in a binary problem — of possible values the new data can take. Specifically, two classifiers must be built for each

[3]Our construction implicitly assumes that a generous benefactor is willing to fund the data acquisition with an unlimited budget, for as long as the expected benefit is positive. If we instead have only a fixed finite budget, this additional budget constraint can be included to play a role in the termination criterion (*i.e.*, the acquisition process would continue as long as both the expected benefit was positive and the budget was not exhausted).

unlabeled data point, one for each possible label. The expected benefit is then

$$\mathbb{E}_{y_*}[B(y_*|\mathcal{D})] = R(\mathcal{D}_U|\mathcal{D}_L) - \{p(y_* = +1|\mathbf{x}_*, \mathbf{w})R(\mathcal{D}_U|\mathcal{D}_L \cup \{y_* = +1\}) \\ + p(y_* = -1|\mathbf{x}_*, \mathbf{w})R(\mathcal{D}_U|\mathcal{D}_L \cup \{y_* = -1\})\} - C(y_*) \quad (27)$$

where $p(y_* = \pm 1|\mathbf{x}_*, \mathbf{w})$ is obtained (via (22)) from the extant classifier built using $\mathcal{D}_L$.

The label for the data point with the maximum (positive) expected benefit is then acquired. Thereafter, the classifier is re-trained. The process is then repeated until the expected benefit of all possible acquisitions is no longer positive.

*2) Feature Acquisition:* If the new data is a (continuous-valued) feature, computing the new expected risk is intractable. We therefore appeal to an idea motivated by the theory of multiple imputation [19] to compute the expectation. Whereas multiple imputation will impute values for *all* missing data, our method will impute a value for only the single feature[4] under consideration, leaving the other missing data incomplete (and handled as discussed in Section II). This effectively isolates the utility of a single feature.

In multiple imputation, $M > 1$ samples are generated according to the (estimated) distribution $p(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i})$ for every missing feature, and $M$ imputed data sets are formed with the missing data completed by these samples. Standard complete-data analysis (*e.g.,* learning and classification) is then performed on each of these completed data sets. The results of each imputed data set are subsequently combined (*e.g.,* averaged) to obtain a single set of results. Theoretical work [19] has shown the proximity between an estimate's uncertainty resulting from a small number of imputations and an infinite number of imputations.

[4]For concreteness, we explain the method for the case in which the new data is a single feature for a single data point. If the new data are multiple features (rather than a single feature), the same basic framework applies. The only difference is that in this case, values will be imputed for the *group* of features under consideration.

In our approach, we impute $M$ samples from the (already possessed) conditional distribution $p(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i})$; each sample, $\mathcal{D}_*^j$, is a value for the feature under consideration. A classifier is then built using the augmented data set consisting of the original data and one of the imputed values. Each of the imputed values is used in turn so that $M$ unique classifiers are constructed. For each classifier, the (expected) risk — $R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*^j)$ — can be computed. The requisite expectation from (26) is then approximated as

$$\mathbb{E}_{\mathcal{D}_*}[R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*)] \approx \frac{1}{M}\sum_{j=1}^{M} R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*^j). \tag{28}$$

Thus, the (approximate) expected benefit of acquiring each single missing feature for every data point can be computed.

The feature with the maximum (positive) expected benefit is then acquired. Thereafter, the classifier is re-trained. The process is then repeated until the expected benefit of all possible acquisitions is no longer positive.

It should be noted that the above algorithm can be used when considering acquiring features of either labeled or unlabeled data. For a general purely supervised classification algorithm, acquiring an additional feature for an unlabeled data point does not change the classifier, so no classifier re-training must be performed[5]. However, the risk would still change because the unlabeled testing data changes. This observation highlights the fact that active data acquisition can improve performance in two distinct ways: by improving the classifier, and by enhancing the (testing) data to be classified.

[5]Our specific supervised classifier actually does change when additional unlabeled data is acquired because the GMM is re-estimated upon acquiring any new data.

*B. Computation Reduction*

The number of times a classifier must be re-trained in our active data acquisition framework grows with the number of features and labels that may be acquired. The nature of some applications may limit the types of data acquisition that are feasible, however. For example, in medical applications, it may be impossible to summon past patients (*i.e.,* data points) back to a hospital to undergo additional medical tests (*i.e.,* features). As a result, it may be feasible to acquire features only for a single specific data point (*e.g.,* a current patient). In a land mine detection task employing airborne sensors, acquiring a set of features (*i.e.,* the features for a given sensor) for *all* data points may be the only feasible type of data acquisition. It would be absurd to fly an airborne sensor to acquire data on only one data point, for example, or even to fly all relevant sensors to acquire data on only one data point.

Numerous heuristic schemes can also be employed to alleviate the computational burden involved with determining the most beneficial data to acquire. For example, groups of features (rather than individual features) can be obtained at each acquisition step. Alternatively, acquisition can be limited to a random subset of data points or features.

The number of features to be considered for acquisition can also be reduced by choosing from a subset of incomplete data points in the following manner. After a classifier is built using all data, compute the risk, $R_0$. It has been shown [20] that data is valuable in the sense that it should decrease the risk. For each incomplete data point, a unique classifier can then be built using all available data points *except* the one incomplete data point under consideration. The risk, $R_{0\setminus i}$, for each of these classifiers built without $\mathbf{x}_i$ is then computed. Since additional data should always decrease the risk, it should be true that $R_0 < R_{0\setminus i}$ for all $i$; that is, the risk for the classifier constructed

with all data should be lower than the risk for the classifiers constructed when withholding one data point. If instead $R_0 > R_{0\setminus i}$ for some $i$, the implication is that including the incomplete data point $\mathbf{x}_i$ in classifier construction adversely affects performance. This result would suggest that an inconsistency existed between our incomplete-data algorithm and the incomplete data point $\mathbf{x}_i$. We argue that these data points are precisely the ones for which additional features should be acquired, to alleviate this inconsistency. The $N_M \geq 1$ incomplete data points for which $J_i = R_0 - R_{0\setminus i}$ is largest then form the subset of data points that will subsequently be considered for the acquisition of a new feature in the standard manner.

## IV. Experimental Results

We evaluated our proposed active data acquisition algorithm on four different data sets: a synthetic data set we created, the benchmark IONOSPHERE data set from the UCI Machine Learning Repository, and two multi-sensor data sets of real (*i.e.,* measured) data. For a data set to be used to evaluate an active data acquisition method, access to the complete data (*i.e.,* with no missing features) is required. When applying such an algorithm, however, missing features must be present. Therefore, prior to beginning data acquisition, we randomly removed features from all data points (both labeled and unlabeled). In all experiments for a given data set, the same initial partition of training and testing data, and the same initial pattern of missing features were always used. This consistency allows direct comparisons among the cases of different types of data acquisition for a given data set. All of the data sets and their specific aforementioned partitions are publicly available at www.duke.edu/$\sim$dpw5/data.html. It is our hope that the data sets will serve as benchmarks against which other researchers can evaluate their active data acquisition algorithms.

We did not compare our method to the other active data acquisition methods [8–10] in the liter-

ature because those heuristic methods do not account for costs, cannot handle incomplete testing data, do not have principled terminations criteria, and can only perform one type of data acquisition (case $G$ in Figure 1). In all of the experiments, we instead compared our proposed active data acquisition method to randomly selecting which data to acquire. Our active data acquisition method terminates automatically when the expected benefit of all possible acquisitions is no longer positive. For the random data acquisition, the same number of actions was taken as in the corresponding active data acquisition case. The random method is also given the advantage that no regard is paid to costs. In all experiments, the performance of the random acquisition cases are the mean $\pm$ one standard deviation over twenty independent trials; all dotted curves in figures reflect these values for the random data acquisition. To compute the expected benefit of potential acquisitions, $M = 1$ imputation was always used (see Section III).

The proposed incomplete-data classifier of Section II is used in all experiments. It should be noted that such a classifier capable of handling incomplete data (without performing imputation) is needed for the proposed active data acquisition framework. For each data set, we also show the classification performance if *no* additional data is acquired, as well as if *all* missing data is acquired (*i.e.,* when the data set is complete). A standard logistic regression classifier is used for the latter complete-data case.

The area under a receiver operating characteristic curve (AUC) is given by the Wilcoxon statistic [21]

$$\text{AUC} = (MN)^{-1} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbf{1}_{x_m > y_n} \tag{29}$$

where $x_1, \ldots, x_M$ are the classifier outputs of data belonging to class 1, $y_1, \ldots, y_N$ are the classifier outputs of data belonging to class -1, and $\mathbf{1}$ is an indicator function. We present the results of the

active and random data acquisition algorithms in terms of the AUC.

There are no publicly available data sets that contain all the requisite costs for demonstrating the various aspects of our proposed active data acquisition framework. Although [23] provides feature acquisition costs for a few data sets, that work does not include costs of acquiring perfect or imperfect labels, or specific misclassification costs. All of the costs used in our experiments appear in the Appendix.

Table I provides a summary of the data sets when the data acquisition process begins. Additional details of the two multi-sensor data sets — UXO and LAND MINE — appear in Table II. In these two multi-sensor applications, missing features would result if sensors were deployed to only a subset of data points. To simulate this real problem, for the UXO and LAND MINE data sets, we remove *groups* of features corresponding to the features produced by a sensor, as opposed to randomly removing *individual* features. In these cases, a data point missing features from a given sensor indicates that the sensor has not been deployed to the data point. Data acquisition will then correspond to deploying sensors.

TABLE I

DETAILS OF THE FOUR DATA SETS WHEN THE DATA ACQUISITION PROCESS BEGINS.

| | NUMBER OF DATA POINTS | | NUMBER OF | FRACTION OF |
|---|---|---|---|---|
| DATA SET | LABELED | UNLABELED | FEATURES | FEATURES MISSING |
| SYNTHETIC | 20 | 180 | 3 | 0.43 |
| IONOSPHERE | 36 | 315 | 33 | 0.50 |
| UXO | 25 | 224 | 6 | 0.36 |
| LAND MINE | 36 | 677 | 41 | 0.48 |

TABLE II

DETAILS OF THE MULTI-SENSOR DATA SETS. DATA POINTS THAT ARE CLUTTER BELONG TO CLASS $-1$; DATA POINTS THAT ARE TARGETS (UXO OR MINES) BELONG TO CLASS $+1$.

| DATA SET | NUMBER OF TARGETS | NUMBER OF CLUTTER | NUMBER OF SENSORS | NUMBER OF FEATURES FROM SENSOR 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|
| UXO | 44 | 205 | 2 | 3 | 3 | — | — |
| LAND MINE | 91 | 622 | 4 | 17 | 6 | 9 | 9 |

### *A. Synthetic Data Set*

Experimental results for the SYNTHETIC data set are summarized in Table III. All eight types of data acquisition from Figure 1 are considered. A ninth case, in which any of the eight types of data acquisition can be performed at each iteration, is also considered; the progression of the performance as a function of the number of actions for this case is shown in Figure 3. This last case highlights the unified aspect of the proposed framework that allows for the acquisition of both labels and features.

### *B. Ionosphere Data Set*

Experimental results for the IONOSPHERE data set are summarized in Table IV. Four types of data acquisition from Figure 1 are considered. A fifth case, in which any of the four types of data acquisition can be performed at each iteration, is also considered; the progression of the performance as a function of the number of actions for this last case is shown in Figure 4.

TABLE III

EXPERIMENTAL RESULTS OF ACTIVE DATA ACQUISITION FOR THE SYNTHETIC DATA SET. THE DATA ACQUISITION TYPES IN THE FIRST COLUMN CORRESPOND TO THE KEY IN FIGURE 1. THE RESULTS IN THE FIRST AND SECOND ROWS CORRESPOND TO THE CASES WHEN NO ADDITIONAL DATA IS ACQUIRED, AND WHEN ALL MISSING DATA IS ACQUIRED, RESPECTIVELY. THE RESULTS IN THE LAST ROW CORRESPOND TO THE CASE WHEN ANY ACQUISITION TYPE ($C - J$) CAN BE APPLIED AT EACH ITERATION. FOR CASE $D$, THE LABELING ERROR RATE IS $\epsilon = 0.2$.

| TYPE OF DATA ACQUISITION | NUMBER OF ACTIONS | TOTAL NUMBER OF FEATURES/LABELS ACQUIRED | FINAL AUC | |
|---|---|---|---|---|
| | | | WITH ACTIVE DATA ACQUISITION | WITH RANDOM DATA ACQUISITION |
| NO MISSING FEATURES | — | 0 | 0.7522 | |
| ALL MISSING FEATURES | — | 261 | 0.9056 | |
| $C$ | 2 | 2 | 0.7980 | $0.7720 \pm 0.0321$ |
| $D$ | 3 | 3 | 0.7919 | $0.7553 \pm 0.0480$ |
| $E$ | 7 | 7 | 0.8307 | $0.7889 \pm 0.0417$ |
| $F$ | 128 | 128 | 0.8407 | $0.8438 \pm 0.0123$ |
| $G$ | 8 | 11 | 0.8299 | $0.7933 \pm 0.0343$ |
| $H$ | 47 | 79 | 0.8370 | $0.7931 \pm 0.0129$ |
| $I$ | 1 | 9 | 0.8208 | $0.7959 \pm 0.0336$ |
| $J$ | 1 | 79 | 0.7583 | $0.7948 \pm 0.0361$ |
| ANY ($C - J$) | 58 | 220/3 | 0.9008 | $0.7975 \pm 0.0522$ |

### *C. UXO Data Set*

The objective of the UXO data set is to discriminate between (*i.e.*, classify) unexploded ordnance (UXO) and clutter. The data set is composed of data from two sensors: a magnetometer

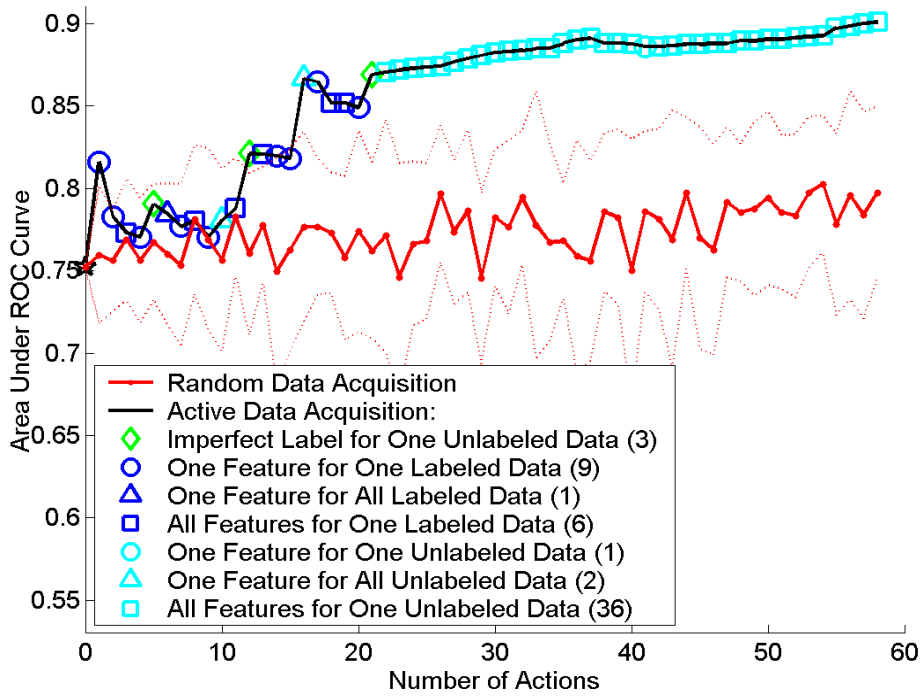


Fig. 3. The progression of the active data acquisition case from Table III for the SYNTHETIC data set, in which one action can be any of the eight actions listed in Figure 1. Numbers in parentheses in the legend indicate the number of times each acquisition type was selected.

and an electromagnetic induction (EMI) sensor. Because of the nature of the sensors, the only feasible type of data acquisition is the deployment of a single sensor to a single data point (labeled or unlabeled; cases $E$ and $F$ in Figure 1). These two cases are considered, as is a third case in which either of the two actions is allowed at each iteration; the progression of the performance as a function of the number of actions for this last case in shown in Figure 5. All experimental results for the UXO data set are summarized in Table V.

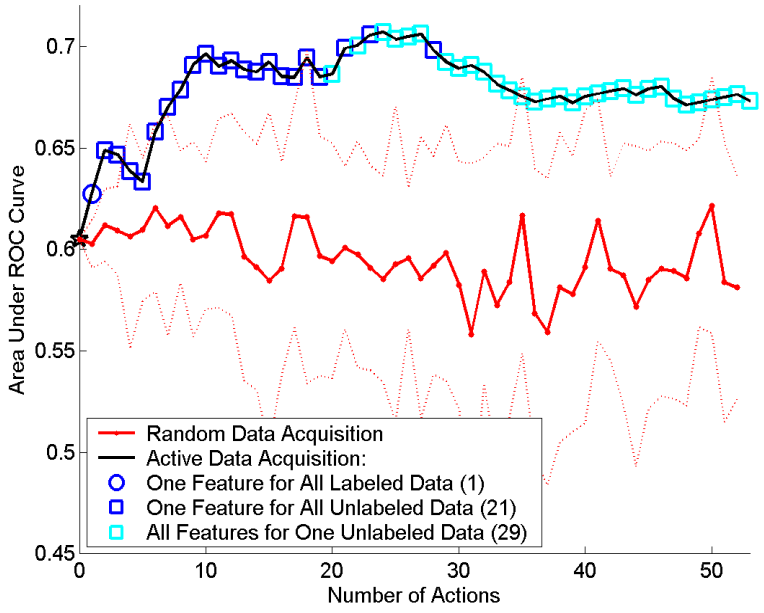


Fig. 4. The progression of the active data acquisition case from Table IV for the IONOSPHERE data set, in which one action can be the acquisition of all missing features of one labeled or unlabeled data point, or the acquisition of one feature for all labeled or unlabeled data points missing it. Numbers in parentheses in the legend indicate the number of times each acquisition type was selected.

### *D. Land Mine Data Set*

The objective of the LAND MINE data set is to discriminate between land mines and clutter. The data set is composed of data collected via four airborne sensors. The four sensors are a ground-penetrating radar (GPR) sensor, an electro-optic infrared (EOIR) sensor, a *Ku*-band synthetic aperature radar (SAR) sensor, and an *X*-band SAR sensor. Because each sensor is mounted on a different aircraft, the flight paths for each sensor are unique. This results in different areas of land — and hence different data points — missing sensor data where the flight path did not cover the location of the data point. Here data acquisition corresponds to deploying (flying) a sensor over an area of land; it is assumed that the labeled and unlabeled data are located in two different

TABLE IV

EXPERIMENTAL RESULTS OF ACTIVE DATA ACQUISITION FOR THE IONOSPHERE DATA SET. THE DATA ACQUISITION TYPES IN THE FIRST COLUMN CORRESPOND TO THE KEY IN FIGURE 1. THE RESULTS IN THE FIRST AND SECOND ROWS CORRESPOND TO THE CASES WHEN NO ADDITIONAL DATA IS ACQUIRED, AND ALL MISSING DATA IS ACQUIRED, RESPECTIVELY. THE RESULTS IN THE LAST ROW CORRESPOND TO THE CASE WHEN ANY OF THE PREVIOUS FOUR ACQUISITION TYPES CAN BE APPLIED AT EACH ITERATION.

| TYPE OF DATA ACQUISITION | NUMBER OF ACTIONS | TOTAL NUMBER OF FEATURES ACQUIRED | FINAL AUC | |
|---|---|---|---|---|
| | | | WITH ACTIVE DATA ACQUISITION | WITH RANDOM DATA ACQUISITION |
| NO MISSING FEATURES | — | 0 | 0.6048 | |
| ALL MISSING FEATURES | — | 5,803 | 0.8676 | |
| $G$ | 19 | 305 | 0.7304 | $0.5528 \pm 0.0647$ |
| $H$ | 223 | 3,697 | 0.7076 | $0.6883 \pm 0.0174$ |
| $I$ | 28 | 498 | 0.5191 | $0.5331 \pm 0.0871$ |
| $J$ | 20 | 3,197 | 0.7213 | $0.7081 \pm 0.0321$ |
| ANY ($G$,$H$,$I$,$J$) | 74 | 3,638 | 0.6734 | $0.5742 \pm 0.0424$ |

geographical locations. Because of the airborne nature of the sensors, the only feasible type of data acquisition is to collect data for *all* data points missing it, at once (cases $I$ and $J$ in Figure 1). These two cases are considered, as is a third case in which either of the two actions is allowed at each iteration; the progression of the performance as a function of the number of actions for this last case is shown in Figure 6.

TABLE V

EXPERIMENTAL RESULTS OF THE ACTIVE DATA ACQUISITION FOR THE UXO DATA SET, IN WHICH EACH ACQUISITION ACTION RESULTED IN ACQUIRING THE FEATURES OF ONE SENSOR FOR ONE DATA POINT. THE RESULTS IN THE FIRST AND SECOND ROWS CORRESPOND TO THE CASES WHEN NO ADDITIONAL DATA IS ACQUIRED, AND ALL MISSING DATA IS ACQUIRED, RESPECTIVELY.

| TYPE OF DATA ACQUIRED | NUMBER OF ACTIONS | FINAL AUC | |
|---|---|---|---|
| | | WITH ACTIVE DATA ACQUISITION | WITH RANDOM DATA ACQUISITION |
| NO MISSING FEATURES | 0 | 0.7476 | |
| ALL MISSING FEATURES | 181 | 0.8273 | |
| LABELED | 11 | 0.7899 | $0.7741 \pm 0.0232$ |
| UNLABELED | 47 | 0.7524 | $0.7314 \pm 0.0169$ |
| LABELED AND UNLABELED | 59 | 0.8156 | $0.7561 \pm 0.0431$ |

## V. DISCUSSION

### *A. Performance with Acquiring Data*

Active data acquisition can affect performance in two ways. In a supervised framework, acquiring features of labeled data will lead to a new classifier, which in turn will affect performance. Acquiring features of unlabeled (testing) data alters (enhances) the testing data, which then affects performance. (Acquiring features of unlabeled data in a semi-supervised framework will also affect the classifier and hence the performance.) Here we discuss in greater depth how acquiring additional features actually affects performance.

Loosely speaking, a classifier is considered to be *stable* [25, 26] if it does not change significantly when one data point is added or removed from the pool of training data. An unstable

TABLE VI

EXPERIMENTAL RESULTS OF ACTIVE DATA ACQUISITION FOR THE LAND MINE DATA SET. EACH ACQUISITION ACTION RESULTED IN ACQUIRING THE FEATURES OF ONE SENSOR FOR ALL (TRAINING OR TESTING) DATA POINTS. THE RESULTS IN THE FIRST AND SECOND ROWS CORRESPOND TO THE CASES WHEN NO ADDITIONAL DATA IS ACQUIRED, AND ALL MISSING DATA IS ACQUIRED, RESPECTIVELY.

| TYPE OF DATA ACQUIRED | NUMBER OF ACTIONS | FINAL AUC | |
|---|---|---|---|
| | | WITH ACTIVE DATA ACQUISITION | WITH RANDOM DATA ACQUISITION |
| NO MISSING FEATURES | 0 | 0.4993 | |
| ALL MISSING FEATURES | 8 | 0.6995 | |
| LABELED | 3 | 0.5690 | $0.5752 \pm 0.0472$ |
| UNLABELED | 1 | 0.5240 | $0.5034 \pm 0.0267$ |
| LABELED AND UNLABELED | 5 | 0.7105 | $0.5657 \pm 0.0984$ |

classifier, in contrast, has a greater propensity and flexibility to change. The idea of active data acquisition implicitly assumes that the extant classifier is unstable. If the classifier were stable, acquiring more data would affect neither the classifier nor performance significantly. Moreover, if performance would not change much, the cost of data acquisition would exceed the expected reduction in misclassification costs; in this scenario, the algorithm would abstain from acquiring data, based on the benefit criterion for acquisitions.

Incomplete data actually decreases the stability of a classifier because it removes the implicit constraints that data would naturally place on the classifier.[6] As a result, the existence of missing

[6]Missing data and training (labeled) data are opposing forces in terms of stability; as the amount of training data increases, the stability of a classifier will increase. This phenomenon has been observed in semi-supervised settings, where incorporating unlabeled data into the design of a classifier has the smallest impact when the amount of labeled training data is large [22, 24].

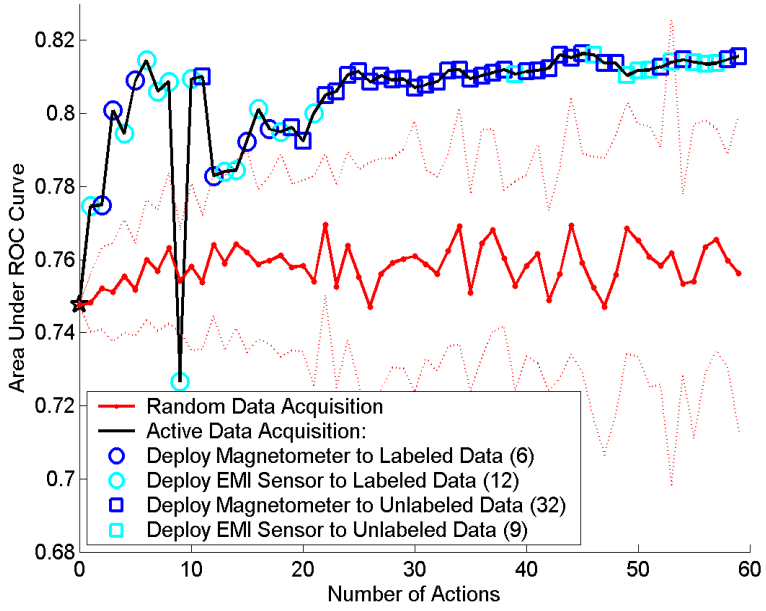


Fig. 5. The progression of the active data acquisition case from Table V for the UXO data set, in which an action can be the acquisition of the features of one missing sensor for either a single training (labeled) data point or a single testing (unlabeled) data point. Numbers in parentheses in the legend indicate the number of times each acquisition type was selected.

data — which is necessary for active data acquisition to even be possible — decreases the stability of the classifier, and hence increases the value of performing active data acquisition.

When active data acquisition is performed and missing training data is acquired, the new constraints imposed on the classifier via the new data increases the classifier's stability. A stable classifier should generalize well [25, 26], which in turn leads to improved performance. As a result, the active data acquisition usually leads to improved performance. However, it was observed that acquiring missing training data sometimes degraded performance. Similarly, it has been observed that performance with an incomplete-data classifier can exceed that of a complete-data classifier (*i.e.*, when no features are missing) [1]. We believe the underlying reason behind both of these surprising phenomena is when data is noisy in the sense that the data does not fit the assumed

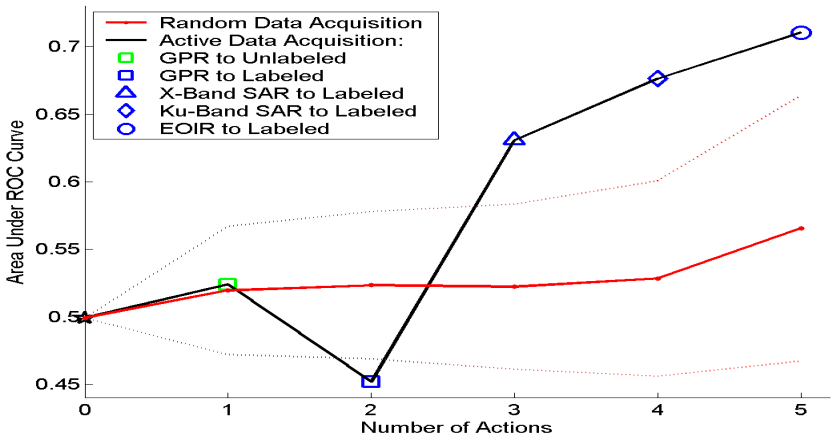


Fig. 6. The progression of the active data acquisition case from Table VI for the LAND MINE data set, in which an action can be the acquisition of the features of one missing sensor for either all training (labeled) data points or all testing (unlabeled) data points.

model. For example, in semi-supervised learning, it is generally agreed upon that having data cannot hurt. Rather, the reason that data sometimes hurts performance is because the assumed model is incorrect [27].

When a data point is incomplete, our incomplete-data classifier integrates out the missing data. This integration acts as an averaging mechanism that builds in robustness for the classifier against noisy data. As a result, a classifier built from incomplete data where some of the missing features are noisy (compared to the model) will have robustness that would be absent when those noisy features are observed. Moreover, observing the noisy features (*e.g.*, via active data acquisition) would typically degrade performance since the data is poor (worse than the effect of the integration). Thus, in these cases, the stability of the robustness owing to the integration exceeds the stability of the noisy data when present, or the instability of its absence.

When active data acquisition is performed and missing testing data is acquired, performance improves when the acquired data is good, because we have better information than the generic

averaging of the integration. However, if the acquired data is noisy, performance will degrade, because the newly acquired observed data will be a disservice compared to the integration. Since data is generally not noisy, performance should generally improve when acquiring missing testing data.

### *B. Advantages of Proposed Method over Existing Methods*

Although the few methods in the active data acquisition literature [8–11] are designed using sound intuition, we shall point out the methods' flaws after briefly describing them. In [8], missing features are acquired for misclassified incomplete training data points. In [9], missing features are acquired for data points for which the inferred missing data either has large variance or improves the model significantly. In [10], missing features are acquired for those data points that would most strongly impact the model. In [11], data acquisition is based on maximizing mutual information.

Our proposed active data acquisition approach is superior to the existing methods [8–11] in the literature for seven main reasons. First, our approach accounts for data acquisition costs, which none of the other methods [8–11] do; this aspect makes our method relevant for real-world applications. Second, our approach can consider the acquisition of labels and features in a single unified framework, which previous methods [8–11] could not. Moreover, the acquisition of imperfect labels is also possible. Third, our approach utilizes incomplete data, so it is not limited to cases (as in [8–10]) where a significant amount of data is complete. In fact, our approach works well even if *no* data points are complete. Fourth, our approach can handle incomplete testing data, another aspect that was lacking in previous approaches [8–10]. Fifth, our approach can consider several different types of feature acquisition, including acquiring all missing features for single or multiple labeled or unlabeled data points. This trait makes the algorithm relevant in a wide range of appli-

cations, where different types of data acquisition are feasible. Previous approaches [8–10] could only acquire all missing features for a single data point, which is impractical in many scenarios (*cf.* LAND MINE data set). Moreover, when faced with multiple possible types of data acquisition, our approach automatically determines which type of acquisition is the most beneficial. Sixth, our approach has a principled termination criterion, which all of the other methods [8–10] lack. Seventh, our proposed approach attempts to directly address the ultimate goal — improving performance on the testing data. Surprisingly, this obvious objective is ignored by other methods, which seek to instead improve performance on the training data [8–10], or to improve a surrogate quantity such as the uncertainty of the classifier's parameters [11].

At first thought, one may instead be inclined to use a policy-based method (such as a partially observable Markov decision process (POMDP) [28]) to perform active data acquisition. In our context, a policy-based approach is inappropriate, as we shall here explain. The purpose of active data acquisition is to improve performance. Performance can be improved in two ways: by improving the classifier, or by improving (enhancing) the testing data. A policy-based active data acquisition method would assume that the classifier is fixed, and hence would not change upon acquiring new data. This assumption would contradict one of the very purposes of acquiring data: to improve the classifier. A policy-based method might be relevant only if the goal of the data acquisition was to improve the testing data. As discussed previously, our proposed approach is part of a comprehensive framework that allows for acquiring any type of data: labels (perfect or imperfect), features of labeled data, or features of unlabeled data.

Even restricting our method to the case of improving only testing data, to compare our method to policy-based methods, policy-based methods still do not enjoy advantages over our method. For example, one attractive quality of a policy-based approach is that the main computational cost —

that of calculating the policy — can be performed "off-line." Thereafter, during data acquisition, the policy is assumed to be fixed. However, in our proposed approach, if we assume that only testing data can be acquired (which the policy-based approach assumes), then our classifier is also fixed.[7] As a result, the major computational burden of our approach can also be viewed as being performed off-line. In fact, the computation involved in our algorithm would then be trivial: sampling features from a GMM and calculating the probability of a label using (22).

The policy-based method is intended for discrete features and as such would need to discretize continuous features, whereas such quantization is unnecessary in our approach. Moreover, the policy-based method would be restricted to naïve-Bayes-type classifiers. In contrast, our proposed approach is general in that any classifier that allows for incomplete data can be utilized. This aspect allows a much richer class of classifiers to be used.

Two potential criticisms of our proposed approach — that costs are assumed to be known, and that testing data is assumed to be available (to compute the expected benefit) — would also be necessarily leveled at policy-based methods. The active data acquisition framework requires that the cost of misclassifying data and the cost of acquiring additional data both be known, but a policy-based method would also assume that these costs were known. The benefit criterion of our proposed approach is computed from the testing data, and therefore assumes testing data is available. However, in a policy-based method, the only type of data acquisition would be on testing data, so the same assumption applies. That is, if testing data was not available, the policy-based methods could not perform any data acquisition at all.

To conclude, although policy-based methods are valuable for solving certain problems, they are

[7] In our approach, if we are satisfied with the extant GMM and choose not to re-train after acquiring additional data, the classifier will indeed be fixed.

not a panacea that should be blindly applied to every problem. When dealing with discrete features in problems where only testing data can be acquired, a policy-based method may be appropriate. For problems dealing with continuous features, rich classes of classifiers, and various options of types of data to acquire, our proposed approach is appropriate.

## VI. Conclusion

Our main contribution is the development of a comprehensive unified framework under which active data acquisition can be performed. This framework is the first that allows for the acquisition of imperfect or perfect labels, as well as of features. Moreover, individual or multiple features of individual or multiple data points, which can be labeled or unlabeled, can also be acquired. As such, every possible type of data acquisition can be performed under our method; this aspect makes the algorithm suitable for any application. Furthermore, our proposed method accounts for data acquisition costs and has a principled termination criterion. All of these aspects of our approach contribute to its superiority over all existing active data acquisition methods.

The main drawback of the active data acquisition framework is the computational burden of classifier re-training. To address this issue, we have outlined several heuristic procedures that can ease the computational burden. Moreover, in practice, the particular application under investigation may also limit the types of data acquisition that are possible (*cf.* Figure 1). We believe the advantages of this principled framework, including the rigorous termination criterion, outweigh these minor drawbacks.

Future work will examine the conditions under which missing data can lead to better performance. Since data should always be helpful, we believe that the absence of some data can improve performance only when the missing data contradicts our model. Though this suggests changing the

model — not a trivial task — we will explore another interesting avenue of research: active feature *removal*. If performance is better when missing certain features, it should hold that performance can be improved by removing some (observed) features. This idea has intimate connections to feature selection.

## REFERENCES


[1] D. Williams and L. Carin, "Analytical Kernel Matrix Completion with Incomplete Multi-View Data," *Proc. Int'l Conf. Machine Learning Workshop on Learning with Multiple Views*, 2005.

[2] V. Sindhwani, P. Niyogi, and M. Belkin, "A Co-Regularization Approach to Semi-Supervised Learning with Multiple Views," *Proc. Int'l Conf. Machine Learning Workshop on Learning with Multiple Views*, 2005.

[3] K. Tsuda, S. Akaho, and K. Asai, "The *em* Algorithm for Kernel Matrix Completion with Auxiliary Data," *J. Machine Learning Research 4*, pp. 67-81, 2003.

[4] G. Lanckriet, M. Deng, N. Cristianini, M. Jordan, and W. Noble, "Kernel-Based Data Fusion and its Application to Protein Function Prediction in Yeast," *Proc. Pacific Symposium on Biocomputing 9* (pp. 300-311), 2004.

[5] D. Cohn, Z. Ghahramani, and M. Jordan, "Active Learning with Statistical Models," *J. Artificial Intelligence Research*, 4, pp. 129-145, 1996.

[6] D. MacKay, "Information-based Objective Functions for Active Data Selection," *Neural Computation*, 4(4), pp. 589-603, 1992.

[7] N. Roy and A. McCallum, "Toward Optimal Active Learning through Sampling Estimation of Error Reduction," *Proc. 18th Int'l Conf. Machine Learning*, pp. 441-448, 2001.

[8] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney, "Active Feature-Value Acquisition for Classifier Induction," *Proc. Int'l Conf. Data Mining (ICDM)*, pp. 483-486, 2004.

[9] Z. Zheng and B. Padmanabhan, "On Active Learning for Data Acquisition," *Proc. Int'l Conf. Data Mining (ICDM)*, pp. 562-570, 2002.

[10] X. Zhu and X. Wu, "Data Acquisition with Active and Impact-Sensitive Instance Selection," *Proc. Int'l Conf. Tools with Artificial Intelligence (ICTAI)*, pp. 721-726, 2004.

[11] B. Krishnapuram, D. Williams, Y. Xue, L. Carin, M. Figueiredo, and A. Hartemink "Active Learning of Features and Labels," *Proc. Int'l Conf. Machine Learning Workshop on Learning with Multiple Views*, 2005.

[12] S. Rässler, *The Impact of Multiple Imputation for DACSEIS* (DACSEIS Research Paper Series 5). University of Erlangen-Nürnberg, Nürnberg, Germany, 2004.

[13] Z. Ghahramani and M. Jordan, "Supervised Learning from Incomplete Data via the EM approach," In J. Cowan, G. Tesauro, and J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6*. San Mateo, CA: Morgan Kaufmann, 1994.

[14] M. Opper and O. Winther, "Gaussian Processes and SVM: Mean Field and Leave-One-Out," In *Advances in Large Margin Classifiers*, Eds. A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, pp. 311-326. MIT Press, 2000.

[15] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Society B 39*, pp. 1-38, 1977.

[16] M. Beal and Z. Ghahramani, "The Variational Bayesian EM Algorithm for Incomplete Data: Application to Scoring Graphical Model Structures," *Bayesian Statistics 7*, pp. 453-464, 2003.

[17] M. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Doctoral dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

[18] D. Williams, X. Liao, Y. Xue, and L. Carin, "Incomplete-Data Classification using Logistic Regression," *Proc. 22nd Int'l Conf. Machine Learning (ICML)*, pp. 977-984, 2005.

[19] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley, 1987.

[20] V. Castelli and T. Cover. "The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter," *IEEE Trans. Information Theory*, vol. 42, no. 6, pp. 2102-2117, Nov. 1996.

[21] J. Hanley and B. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology 143*, pp. 29-36, 1982.

[22] B. Krishnapuram, D. Williams, Y. Xue, L. Carin, A. Hartemink, and M. Figueriedo, "On Semi-Supervised Classification," *Advances in Neural Information Processing Systems 17*, pp. 721-728, 2004.

[23] P. Turney, "Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm," *J. Artificial Intelligence Research 2*, pp. 369-409, 1995.

[24] T. Joachims, "Transductive Learning via Spectral Graph Partitioning," *Proc. 20th Int'l Conf. Machine Learning (ICML)*, 2003.

[25] T. Poggio, R. Rifkin, S. Mukherjee and P. Niyogi, "General Conditions for Predictivity in Learning Theory," *Nature*, vol. 428, pp. 419-422, 2004.

[26] C. Tomasi, "Past Performance and Future Results," *Nature*, vol. 428, p. 378, 2004.

[27] I. Cohen, F. Cozman, N. Sebe, M. Cirelo, and T. Huang. "Semi-Supervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-Computer Interaction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1553-1567, Dec. 2004.

[28] L. Kaelbling, M. Littman and A. Cassandra, "Planning and Acting in Partially Observable Stochastic Domains," *Artificial Intelligence 101*, pp. 99-134, 1998.

## APPENDIX

### A. *Synthetic Data Set Costs*

The misclassification and data acquisition costs used for the SYNTHETIC data set appear in Table VII, while cost discount factors appear in Table VIII. The discount factors reflect the beliefs that (i) training data is more accessible, easier to examine, and hence cheaper to acquire than testing data; (ii) acquiring the same feature for multiple data points at once is cheaper because data acquisition equipment and experts must be summoned only once; and (iii) acquiring multiple features at the same time for a single data point is cheaper because data acquisition preparation must be performed only once. The discount factors mean that, for example, acquiring feature 1 for all (9, in this case) training data points for which it is missing costs $C = 9 \times 0.30 \times 0.75 \times 0.90 = 1.8225$.

TABLE VII

MISCLASSIFICATION AND DATA ACQUISITION COSTS FOR THE SYNTHETIC DATA SET.

| ACTION | COST |
|---|---|
| MISCLASSIFY $C_{[1,-1]}$ | 1.00 |
| MISCLASSIFY $C_{[-1,1]}$ | 1.00 |
| ACQUIRE PERFECT LABEL | 4.00 |
| ACQUIRE IMPERFECT LABEL ($\epsilon = 0.2$) | 2.00 |
| ACQUIRE FEATURE 1 | 0.30 |
| ACQUIRE FEATURE 2 | 0.20 |
| ACQUIRE FEATURE 3 | 0.10 |

TABLE VIII

COST DISCOUNT FACTORS ASSOCIATED WITH SPECIFIC TYPES OF DATA ACQUISITION FOR THE SYNTHETIC AND IONOSPHERE DATA SETS.

| TYPE | DISCOUNT FACTOR |
|---|---|
| ACTION PERFORMED ON TESTING DATA | 1.00 |
| ACTION PERFORMED ON TRAINING DATA | 0.75 |
| SAME SINGLE FEATURE ACQUIRED FOR MULTIPLE DATA POINTS | 0.90 |
| MULTIPLE FEATURES ACQUIRED FOR SAME SINGLE DATA POINT | 0.75 |

### *B. Ionosphere Data Set Costs*

The misclassification and data acquisition costs used for the IONOSPHERE data set appear in Table IX, while the cost discount factors are the same as those used for the SYNTHETIC data set (Table VIII).

TABLE IX

MISCLASSIFICATION AND DATA ACQUISITION COSTS FOR THE IONOSPHERE DATA SET.

| ACTION | COST |
|---|---|
| MISCLASSIFY $C_{[1,-1]}$ | 1.00 |
| MISCLASSIFY $C_{[-1,1]}$ | 1.00 |
| ACQUIRE ANY SINGLE FEATURE | 0.01 |

### *C. UXO Data Set Costs*

The misclassification and data acquisition costs used for the UXO data set appear in Table X. It is assumed that the cost of deploying a sensor to labeled and unlabeled data points is equal because

the two environments are similar.

TABLE X

MISCLASSIFICATION AND DATA ACQUISITION COSTS FOR THE UXO DATA SET. THE DEPLOYMENT OF A SENSOR RESULTS IN THE ACQUISITION OF THAT SENSOR'S FEATURES FOR A SINGLE (LABELED OR UNLABELED) DATA POINT.

| ACTION | COST |
|---|---|
| MISCLASSIFY $C_{[1,-1]}$ | 1.00 |
| MISCLASSIFY $C_{[-1,1]}$ | 1.00 |
| DEPLOY MAGNETOMETER | 0.10 |
| DEPLOY EMI SENSOR | 0.30 |

### *D. Land Mine Data Set Costs*

The misclassification and data acquisition costs used for the LAND MINE data set appear in Table XI. The asymmetric misclassification costs imply that it is much more costly to misclassify a land mine as clutter than vice versa. The costs of deploying sensors to unlabeled data are higher because the unlabeled data is located in a hostile environment.

TABLE XI

MISCLASSIFICATION AND DATA ACQUISITION COSTS FOR THE LAND MINE DATA SET. THE DEPLOYMENT OF A SENSOR RESULTS IN THE ACQUISITION OF THAT SENSOR'S FEATURES FOR *all* (LABELED OR UNLABELED) DATA.

| ACTION | COST FOR | |
|---|---|---|
| | LABELED DATA | UNLABELED DATA |
| MISCLASSIFY $C_{[1,-1]}$ | — | 1.00 |
| MISCLASSIFY $C_{[-1,1]}$ | — | 10.00 |
| DEPLOY GPR SENSOR | 5.00 | 20.00 |
| DEPLOY IR SENSOR | 10.00 | 40.00 |
| DEPLOY *Ku*-BAND SAR SENSOR | 15.00 | 60.00 |
| DEPLOY *X*-BAND SAR SENSOR | 20.00 | 80.00 |