Active Learning of Features and Labels

Balaji Krishnapuram BALAJI.KRISHNAPURAM@SIEMENS.COM
Siemens Medical Solutions USA, 51 Valley Stream Parkway, Malvern, PA 19355

David Williams DPW@EE.DUKE.EDU
Ya Xue YA.XUE@DUKE.EDU
Lawrence Carin LCARIN@EE.DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Box 90291, Durham, NC 27708-0291

Mário A. T. Figueiredo MTF@LX.IT.PT
Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal

Alexander J. Hartemink AMINK@CS.DUKE.EDU
Department of Computer Science, Duke University, Box 90129, Durham, NC 27708-0129

Abstract

Co-training improves multi-view classifier learning by enforcing internal consistency between the predicted classes of unlabeled objects based on different views (different sets of features for characterizing the same object). In some applications, due to the cost involved in data acquisition, only a subset of features may be obtained for many unlabeled objects. Observing additional features of objects that were earlier incompletely characterized, increases the data available for co-training, hence improving the classification accuracy. This paper addresses the problem of active learning of features: which additional features should be acquired of incompletely characterized objects in order to maximize the accuracy of the learned classifier? Our method, which extends previous techniques for the active learning of labels, is experimentally shown to be effective in a real-life multi-sensor mine detection problem.

1. Motivation

A fundamental assumption in the field of classifier design is that it is costly to acquire labels; after all, if label acquisition were cheap, we would have little need for classifiers because we could simply acquire labels as and when we needed them. But how does the situation change when it is also costly to acquire features? This paper aims to answer this question. We begin with a little more motivation.

In the simplest setting for classifier design, each object has been characterized by a vector of features and a label, as schematically depicted in Figure 1a. Assuming that labels are indeed costly to acquire, we can imagine relaxing this setting so that each object has been characterized by a vector of features, but only a small subset of the objects has been labeled. If we are not permitted to acquire additional labels for the unlabeled data, as shown in Figure 1b, we are in a semi-supervised learning setting (Belkin et al., 2004; Blum & Chawla, 2001; Corduneanu & Jaakkola, 2004; Inoue & Ueda, 2003; Joachims, 1999; Joachims, 2003; Krishnapuram et al., 2004; Nigam et al., 2000; Seeger, 2001; Zhu et al., 2003); on the other hand, if we *are* permitted to label some of the unlabeled data (Figure 1c), we are in an active learning setting (MacKay, 1992; Muslea et al., 2000; Krishnapuram et al., 2004; Tong & Koller, 2001).

Expanding this framework still further, sometimes the objects to be classified can be characterized by vectors of features in multiple independent ways; we will call each of these characterizations a *view*. For example, a web page may be described either using the words it contains or the set of words in the links pointing to it. A person may be identified on the basis of facial features in an image, speech patterns in an audio recording, or characteristic motions in a video. Buried mines may be investigated using radar, sonar, hyper-spectral, or other kinds of physical sensors. Assuming that only a small subset of the objects has been labeled and that no further labels may be acquired (Figure 1d), we are in the setting of the original co-training algorithm of Blum and Mitchell (1998), which has been extended in a number of interesting directions in subsequent work (Brefeld & Scheffer, 2004; Collins & Singer, 1999; Dasgupta et al., 2001; Balcan et al., 2004). In particular, we recently reformulated co-training using a prior in a

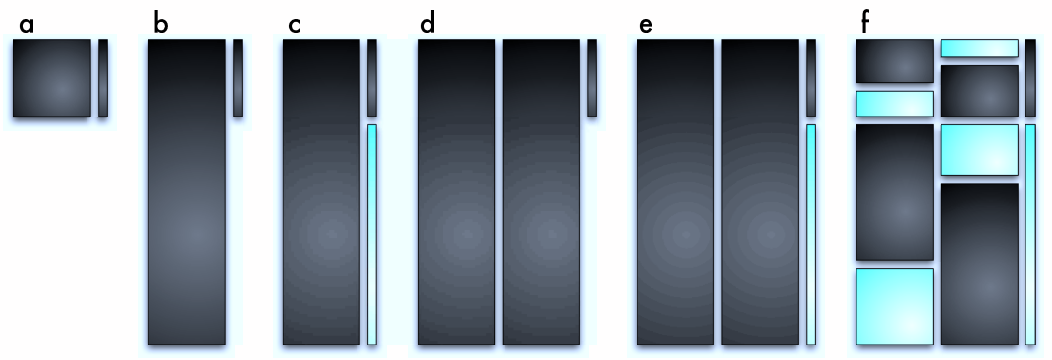


Figure 1. Schematic depiction of different settings. Throughout, rows correspond to objects, wide boxes to feature matrices, and narrow boxes to vectors of class labels; black shading indicates available data, blue shading indicates missing data that can be acquired, and whitespace indicates missing data that cannot be acquired. (a) Each object is characterized by one set of features and one label: supervised learning. (b) Some objects are missing labels that cannot be acquired: semi-supervised learning. (c) Some objects are missing labels that can be acquired: active learning of labels. (d) Objects can be characterized by more than one view, but some are missing labels that cannot be acquired: co-training. (e) Same as (d) but labels can be acquired: active learning of labels with co-training (Krishnapuram et al., 2004). (f) Some objects have not been labeled and not all objects have been characterized in all views: active learning of features and labels (this paper).

Bayesian context (Krishnapuram et al., 2004). This reformulation is based on logistic regression, yielding a convex objective function with a unique local optimum.

As shown in (Krishnapuram et al., 2004), our formulation enables us to consider active learning settings in which we are now permitted to label some of the unlabeled data, as depicted in Figure 1e. But this same formulation also enables us to consider a new setting in which each object may be characterized by only a subset of available views. This can occur in real-life when features are also costly to acquire, as is often the case when physical sensors need to be deployed for each view of an object. If new views may be acquired for any object, as depicted in Figure 1f, how should we decide which view to acquire? And what is the relative benefit of acquiring features versus labels?

In terms of previous work, while several authors have provided criteria for deciding which objects should be labeled (the setting of Figures 1c and 1e), we seek to answer a new question: which incompletely characterized objects (whether labeled or unlabeled) should be further investigated in order to most accurately learn a classifier? To the best of our knowledge, despite its clear importance, the latter question has not been formally addressed before. A few authors have developed intuitive but somewhat *ad hoc* approaches for acquiring features only for labeled objects (Melville et al., 2004; Zheng & Padmanabhan, 2002), but we believe this is the first approach for feature acquisition on both labeled and unlabeled objects.

Section 2 summarizes the probabilistic model for multi-view classifier design that we inherit from Krishnapuram et al. (2004). Section 3 explains the information-theoretic background for the criteria developed in Sections 4 and 5 for active label acquisition and active feature acquisition, respectively. Experimental results are provided in Section 6 and a summary of our conclusions in Section 7.

2. Probabilistic model

2.1. Notation

For notational simplicity, we focus on two-class problems for objects characterized by two views; the proposed methods extend naturally to multi-class and multi-view problems. Since we have only two views, we'll use dot notation to indicate them: let $\dot{\boldsymbol{x}}_i \in \mathbb{R}^{d_1}$ and $\ddot{\boldsymbol{x}}_i \in \mathbb{R}^{d_2}$ be the feature vectors obtained from the two views of the i-th object. Let $\boldsymbol{x}_i = [\dot{\boldsymbol{x}}_i^T, \ddot{\boldsymbol{x}}_i^T]^T$ be the d-dimensional ($d = d_1 + d_2$) vector containing the concatenation of the feature vectors from both views (with appropriate missing values if an object has not been characterized in both views).

In addition to the features in the two views, binary class labels are also collected for a subset of objects; the label of the i-th object is denoted as $y_i \in \{-1, 1\}$. The set of L labeled objects is $\mathcal{D}_L = \{(\boldsymbol{x}_i, y_i) : \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}\}_{i=1}^{L}$, while the set of U unlabeled objects is $\mathcal{D}_U = \{\boldsymbol{x}_i : \boldsymbol{x}_i \in \mathbb{R}^d\}_{i=L+1}^{L+U}$. Thus, the available training data is $\mathcal{D}_{\text{train}} = \mathcal{D}_L \bigcup \mathcal{D}_U$.

Let $\dot{\mathcal{S}}$, $\ddot{\mathcal{S}}$, and $\dddot{\mathcal{S}} = \dot{\mathcal{S}} \bigcap \ddot{\mathcal{S}}$ denote, respectively, the sets containing the indices of objects characterized by sensor 1, sensor 2, and both. The indices of the corresponding labeled and unlabeled objects are denoted as $\dot{\mathcal{S}}_L$, $\ddot{\mathcal{S}}_L$, $\dddot{\mathcal{S}}_L$

and $\dot{\mathcal{S}}_U, \ddot{\mathcal{S}}_U, \dddot{\mathcal{S}}_U$.

2.2. Multi-view logistic classification

In binary logistic regression, the predicted class probabilities are modeled using the well-known logistic function $\sigma(z) = (1+\exp(-z))^{-1}$. For example, in the first view,

$$P\left(y_i \;\middle|\; \dot{\boldsymbol{x}}_i, \dot{\boldsymbol{w}}\right) = \sigma\left(y_i \dot{\boldsymbol{w}}^T \dot{\boldsymbol{x}}_i\right), \tag{1}$$

where $\dot{\boldsymbol{w}}$ is the classifier weight vector for the first view. A similar expression holds for the second view. Denoting $\boldsymbol{w} = [\dot{\boldsymbol{w}}^T, \ddot{\boldsymbol{w}}^T]^T$, we can find the maximum likelihood (ML) estimate of the classifiers for both sensors $\widehat{\boldsymbol{w}}_{\text{ML}}$, by maximizing the overall log-likelihood,

$$\ell(\boldsymbol{w}) = \ell_1(\dot{\boldsymbol{w}}) + \ell_2(\ddot{\boldsymbol{w}}),$$

where

$$\ell_1(\dot{\boldsymbol{w}}) = \sum_{i \in \dot{\mathcal{S}}_L} \log P\left(y_i \;\middle|\; \dot{\boldsymbol{x}}_i, \dot{\boldsymbol{w}}\right),$$
$$\ell_2(\ddot{\boldsymbol{w}}) = \sum_{i \in \ddot{\mathcal{S}}_L} \log P\left(y_i \;\middle|\; \ddot{\boldsymbol{x}}_i, \ddot{\boldsymbol{w}}\right).$$

Given a prior $p(\boldsymbol{w})$, we can find the maximum *a posteriori* (MAP) estimate $\widehat{\boldsymbol{w}}_{\text{MAP}}$ by maximizing the log-posterior $L(\boldsymbol{w}) = \ell(\boldsymbol{w}) + \log p(\boldsymbol{w})$. Clearly, ML estimation can be accomplished by independently maximizing the log-likelihoods for each sensor, $\ell_1(\dot{\boldsymbol{w}})$ and $\ell_2(\ddot{\boldsymbol{w}})$. If the prior factorizes as $p(\boldsymbol{w}) = p_1(\dot{\boldsymbol{w}})\, p_2(\ddot{\boldsymbol{w}})$ (*i.e.* , it models $\dot{\boldsymbol{w}}$ and $\ddot{\boldsymbol{w}}$ as *a priori* independent) we can clearly still perform MAP estimation of the two classifiers separately. However, if $p(\boldsymbol{w})$ expresses some dependence between $\dot{\boldsymbol{w}}$ and $\ddot{\boldsymbol{w}}$, both classifiers must be trained simultaneously by jointly maximizing $L(\boldsymbol{w})$. In this case, the classifier learned for each sensor also depends on the data from the other sensor. This provides a Bayesian mechanism for sharing information and thus exploiting synergies in learning classifiers for different sensors.

2.3. Co-training priors

The standard means of coordinating information from both sensors is by using the concept of *co-training* (Blum & Mitchell, 1998): on the objects with indices in $\dddot{\mathcal{S}}_U$, the two classifiers should agree as much as possible. In a logistic regression framework, the disagreement between the two classifiers on the objects in $\dddot{\mathcal{S}}$ can be measured by

$$\sum_{i \in \dddot{\mathcal{S}}_U} \left(\dot{\boldsymbol{w}}^T \dot{\boldsymbol{x}}_i - \ddot{\boldsymbol{w}}^T \ddot{\boldsymbol{x}}_i\right)^2 = \boldsymbol{w}^T \mathbf{C}\, \boldsymbol{w}, \tag{2}$$

where $\mathbf{C} = \sum_{i \in \dddot{\mathcal{S}}_U} [\dot{\boldsymbol{x}}_i^T, -\ddot{\boldsymbol{x}}_i^T]^T [\dot{\boldsymbol{x}}_i^T, -\ddot{\boldsymbol{x}}_i^T]$. This suggests the following Gaussian "co-training prior"

$$p(\boldsymbol{w}) = p(\dot{\boldsymbol{w}}, \ddot{\boldsymbol{w}}) \propto \exp\left\{-(\lambda_{co}/2)\, \boldsymbol{w}^T \mathbf{C}\, \boldsymbol{w}\right\}. \tag{3}$$

This co-training prior can be combined with other *a priori* information, also formulated in the form of Gaussian priors, derived from labeled and unlabeled data using the formulation in (Krishnapuram et al., 2004). Formally,

$$p(\boldsymbol{w}|\boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{w} \;\middle|\; \mathbf{0}; \left(\boldsymbol{\Delta}_{\text{prior}}(\boldsymbol{\lambda})\right)^{-1}\right), \tag{4}$$

where the prior precision matrix $\boldsymbol{\Delta}_{\text{prior}}(\boldsymbol{\lambda})$, which is a function of a set of parameters (including λ_{co}) collected in vector $\boldsymbol{\lambda}$, is

$$\boldsymbol{\Delta}_{\text{prior}}(\boldsymbol{\lambda}) = \boldsymbol{\Lambda} + \lambda_{co}\mathbf{C} + \begin{bmatrix} \dot{\lambda}\dot{\boldsymbol{\Delta}} & \mathbf{0} \\ \mathbf{0} & \ddot{\lambda}\ddot{\boldsymbol{\Delta}} \end{bmatrix} \tag{5}$$

with $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, ..., \lambda_d\}$; finally

$$\dot{\boldsymbol{\Delta}} = \sum_{i,j \in \dot{\mathcal{S}}, i>j} \dot{K}_{ij} \left(\dot{\boldsymbol{x}}_i - \dot{\boldsymbol{x}}_j\right)\left(\dot{\boldsymbol{x}}_i - \dot{\boldsymbol{x}}_j\right)^T$$

is the precision matrix for semi-supervised learning derived in Krishnapuram et al. (2004), and $\ddot{\boldsymbol{\Delta}}$ is a similar expression. All the parameters in $\boldsymbol{\lambda}$ formally play the role of inverse variances; thus, they are given conjugate gamma hyper-priors. If we let $\dot{\lambda} = \ddot{\lambda} = \lambda_0$, then we have:

$$\begin{aligned} p\left(\lambda_0 \;\middle|\; \alpha_0, \beta_0\right) &= \text{Ga}\left(\lambda_0 \;\middle|\; \alpha_0, \beta_0\right), \\ p\left(\lambda_i \;\middle|\; \alpha_1, \beta_1\right) &= \text{Ga}\left(\lambda_i \;\middle|\; \alpha_1, \beta_1\right), \\ p\left(\lambda_{co} \;\middle|\; \alpha_{co}, \beta_{co}\right) &= \text{Ga}\left(\lambda_{co} \;\middle|\; \alpha_{co}, \beta_{co}\right). \end{aligned}$$

Under this formulation, it is possible to interpret $\boldsymbol{\lambda}$ as a hidden variable and write a generalized EM (GEM) algorithm for obtaining an MAP estimate $\widehat{\boldsymbol{w}}_{\text{MAP}}$. It is easy to check that the complete-data log-likelihood is linear with respect to $\boldsymbol{\lambda}$; thus, in each iteration of the GEM algorithm, the E-step reduces to the computation of the conditional expectation of $\boldsymbol{\lambda}$ given the current parameter estimate and the observed data (this can be done analytically due to conjugacy). The (generalized) M-step then consists of maximizing a lower bound on the complete log-likelihood (see (Böhning, 1992)) to obtain the new classifier estimate. The steps are repeated until some convergence criterion is met.

3. Information-theoretic criteria for active data acquisition

This section is devoted to answering the following question: what additional information should be added to $\mathcal{D}_{\text{train}}$ so that the classifier parameters $\boldsymbol{w}$ are learned most accurately, at minimum expense? Observe that there are several ways in which $\mathcal{D}_{\text{train}}$ can be augmented: (1) label information y_i for a previously unlabeled object $\boldsymbol{x}_i \in \mathcal{D}_U$; (2) features from sensor 1 for an *unlabeled* object $i \in \ddot{\mathcal{S}}_U \setminus \dddot{\mathcal{S}}_U$ (*i.e.,* such that sensor 2 has been acquired, but 1 has not);

(3) features from sensor 2 for an *unlabeled* object $i \in \mathcal{S}_U^1 \setminus \ddot{\dot{\mathcal{S}}}_U$; (4) and (5) same as (2) and (3), but for labeled objects. In this section, we show how information-theoretic tools can be used to choose the best object to be queried for further information under each scenario.

3.1. Laplace approximation for the posterior density

Ignoring the hyper-priors on the regularizer $\boldsymbol{\lambda}$ (*i.e.* , assuming a fixed $\boldsymbol{\lambda}$), after estimating a classifier $\widehat{\boldsymbol{w}}_{\text{MAP}}$ from training data $\mathcal{D}_{\text{train}}$, a Laplace approximation models the posterior density $p(\boldsymbol{w}|\mathcal{D}_{\text{train}})$ as a Gaussian

$$p(\boldsymbol{w}|\mathcal{D}_{\text{train}}) \approx \mathcal{N}\left(\boldsymbol{w}|\widehat{\boldsymbol{w}}_{\text{MAP}}; (\boldsymbol{\Delta}_{\text{post}})^{-1}\right). \tag{6}$$

Under the logistic log-likelihood and the Gaussian prior (4) herein considered, the posterior precision matrix of the Laplace approximation is given by:

$$\boldsymbol{\Delta}_{\text{post}} = \boldsymbol{\Delta}_{\text{prior}}(\boldsymbol{\lambda}) + \boldsymbol{\Psi} \tag{7}$$

where $\boldsymbol{\Delta}_{\text{prior}}(\boldsymbol{\lambda})$ is the prior precision matrix in (5) and $\boldsymbol{\Psi} = \text{block-diag}\{\dot{\boldsymbol{\Psi}}, \ddot{\boldsymbol{\Psi}}\}$ is the Hessian of the negative log-likelihood (see, *e.g.* , (Böhning, 1992)) where

$$\dot{\boldsymbol{\Psi}} = \sum_{i \in \dot{\mathcal{S}}_L} \dot{p}_i \left(1 - \dot{p}_i\right) \dot{\boldsymbol{x}}_i \dot{\boldsymbol{x}}_i^T,$$

with $\dot{p}_i = \sigma\left(\dot{\boldsymbol{w}}^T \dot{\boldsymbol{x}}_i\right)$; a similar expression holds for $\ddot{\boldsymbol{\Psi}}$.

The differential entropy of the Gaussian posterior under the Laplace approximation is thus ($|\cdot|$ denotes determinant)

$$h(\boldsymbol{w}) = -\frac{1}{2} \log \frac{|\boldsymbol{\Delta}_{\text{post}}|}{2\pi e}. \tag{8}$$

3.2. Mutual information

After estimating a classifier $\widehat{\boldsymbol{w}}_{\text{MAP}}$ from $\mathcal{D}_{\text{train}}$, the (un)certainty in the label y_i predicted for an unlabeled object $\boldsymbol{x}_i \in \mathcal{D}_U$ is given by the logistic model (1): $P(y_i|\boldsymbol{x}_i, \widehat{\boldsymbol{w}}_{\text{MAP}})$. For a object (labeled or not) for which we have $\dot{\boldsymbol{x}}_i$ but not $\ddot{\boldsymbol{x}}_i$ ($i \in \dot{\mathcal{S}} \setminus \ddot{\dot{\mathcal{S}}}$), the uncertainty in the latter can be modeled by some representation of $p(\ddot{\boldsymbol{x}}_i|\dot{\boldsymbol{x}}_i)$ learned from the training objects in $\ddot{\dot{\mathcal{S}}}$.

The mutual information (MI) between $\boldsymbol{w}$ and y_i is the *expected* decrease in entropy of $\boldsymbol{w}$ when y_i is observed,

$$\begin{aligned} I\left(\boldsymbol{w}; y_i\right) &= h(\boldsymbol{w}) - \mathbb{E}\left[h(\boldsymbol{w}|y_i)\right] \\ &= \mathbb{E}\left[\log \left|\boldsymbol{\Delta}_{\text{post}}^{y_i}\right|\right] - \log \left|\boldsymbol{\Delta}_{\text{post}}\right|, \end{aligned} \tag{9}$$

where the expectation is w.r.t y_i with probability distribution $P(y_i|\boldsymbol{x}_i, \widehat{\boldsymbol{w}}_{\text{MAP}})$, while $\boldsymbol{\Delta}_{\text{post}}^{y_i}$ is the posterior precision matrix of the re-trained classifier after observing y_i.

Similarly, the MI between $\boldsymbol{w}$ and a previously unobserved feature $\ddot{\boldsymbol{x}}_i$ (for $i \in \dot{\mathcal{S}} \setminus \ddot{\dot{\mathcal{S}}}$) is given by

$$\begin{aligned} I\left(\boldsymbol{w}; \ddot{\boldsymbol{x}}_i\right) &= h(\boldsymbol{w}) - \mathbb{E}\left[h(\boldsymbol{w}|\ddot{\boldsymbol{x}}_i)|\dot{\boldsymbol{x}}_i\right] \\ &= \mathbb{E}\left[\log \left|\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}\right| \middle| \dot{\boldsymbol{x}}_i\right] - \log \left|\boldsymbol{\Delta}_{\text{post}}\right|, \end{aligned} \tag{10}$$

where the expectation is over the uncertainty $p(\ddot{\boldsymbol{x}}_i|\dot{\boldsymbol{x}}_i)$ and $\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}$ is the posterior precision matrix of the retrained classifier after seeing features from sensor 2 for object i.

The maximum MI criterion has been used before to identify the "best" unlabeled object for which to obtain an additional label (MacKay, 1992):

$$i^* = \underset{i:\boldsymbol{x}_i \in \mathcal{D}_U}{\arg\max}\ I\left(\boldsymbol{w}; y_i\right) = \underset{i:\boldsymbol{x}_i \in \mathcal{D}_U}{\arg\max}\, \mathbb{E}\left[\log \left|\boldsymbol{\Delta}_{\text{post}}^{y_i}\right|\right]. \tag{11}$$

Based on the same criterion, the best object for which to acquire sensor 2 features—among $\dot{\mathcal{S}} \setminus \ddot{\mathcal{S}}$ for which we have features from sensor 1, but not sensor 2—would be

$$i^\dagger = \underset{i \in \dot{\mathcal{S}} \setminus \ddot{\dot{\mathcal{S}}}}{\arg\max}\, \mathbb{E}\left[\log \left|\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}\right| \middle| \dot{\boldsymbol{x}}_i\right]. \tag{12}$$

3.3. Upper bound on mutual information

Unfortunately, $\mathbb{E}[\log |\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}|]$ is very difficult to compute for our models. Alternatively, we compute an upper bound and use it in the maximum MI criterion just presented. Since the function $\log|\mathbf{X}|$ is concave (Boyd & Vandenberghe, 2003), by Jensen's inequality we obtain

$$\mathbb{E}\left[\log|\mathbf{X}|\right] \leq \log\left|\mathbb{E}\left[\mathbf{X}\right]\right|. \tag{13}$$

Hence, our sample selection criterion will be

$$i^\dagger = \underset{i \in \dot{\mathcal{S}} \setminus \ddot{\dot{\mathcal{S}}}}{\arg\max} \left|\mathbb{E}\left[\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i} \middle| \dot{\boldsymbol{x}}_i\right]\right|, \tag{14}$$

instead of the original (12). Intuitively, we try to maximize the expected posterior precision of the parameters.

3.4. Simplifying assumptions

We make two simplifying assumptions, fundamental in making our approach practical for real-life problems.

Assumption 1: Let the posterior density of the parameters, given the original training data $\mathcal{D}_{\text{train}}$, be $p(\boldsymbol{w}|\mathcal{D}_{\text{train}})$. Consider that we obtain additional features $\ddot{\boldsymbol{x}}_i$, for some $i \in \dot{\mathcal{S}} \setminus \ddot{\dot{\mathcal{S}}}$ and retrain the classifier, obtaining a new posterior $p(\boldsymbol{w}|\mathcal{D}_{\text{train}}, \ddot{\boldsymbol{x}}_i)$. When computing the utility of $\ddot{\boldsymbol{x}}_i$, we assume that the the modes of $p(\boldsymbol{w}|\mathcal{D}_{\text{train}}, \ddot{\boldsymbol{x}}_i)$ and $p(\boldsymbol{w}|\mathcal{D}_{\text{train}})$ coincide, although their precision matrices may not. It turns out that it will be possible to obtain the new precisions, without actually re-training, which would be very computationally expensive. It is important to highlight that,

after a "best" index $i^\dagger$ is chosen (under this simplifying assumption), we actually observe $\ddot{\boldsymbol{x}}_{i^\dagger}$ and re-train the classifier, thus updating the mode of the posterior. Since this re-training is done only once for each additional feature acquisition, tremendous computational effort is saved.

The same assumption is made for label acquisition.

Assumption 2: For the purpose of computing the utility of acquiring some new data (a label or a set of features), we treat $\boldsymbol{\lambda}$ as deterministic, and fixed at the value of its expectation after convergence of the GEM algorithm mentioned in Section 2.3. This value is substituted in (7) to compute the entropy and the mutual information.

4. Acquiring additional labels

For the sake of completeness, we now review the approach in Krishnapuram et al. (2004) for acquiring labels.

According to Assumption 1, the MAP estimate $\widehat{\boldsymbol{w}}_{\text{MAP}}$ does not change when $\mathcal{D}_{\text{train}}$ is augmented with a new label y_i; consequently, the class probability estimates are also unchanged. Based on (7), if we obtain the label y_i, for some $\boldsymbol{x}_i \in \mathcal{D}_U$, regardless of whether $y_i = -1$ or $y_i = 1$, the posterior precision matrix becomes

$$\begin{aligned}\boldsymbol{\Delta}_{\text{post}}^{y_i} = \boldsymbol{\Delta}_{\text{post}} &+ \dot{p}_i\,(1-\dot{p}_i)\begin{bmatrix}\dot{\boldsymbol{x}}_i\\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\dot{\boldsymbol{x}}_i\\ \mathbf{0}\end{bmatrix}^T \\ &+ \ddot{p}_i\,(1-\ddot{p}_i)\begin{bmatrix}\mathbf{0}\\ \ddot{\boldsymbol{x}}_i\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \ddot{\boldsymbol{x}}_i\end{bmatrix}^T\end{aligned} \tag{15}$$

The unlabeled object maximizing $|\boldsymbol{\Delta}_{\text{post}}^{y_i}|$ is thus queried for its label. Intuitively, this favors objects with uncertain class probability estimates ($\dot{p}_i$ and/or $\ddot{p}_i$ close to $1/2$).

5. Acquiring additional features

In this section we show how to compute $\mathbb{E}[\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}|\dot{\boldsymbol{x}}_i]$, which is needed to implement the criterion in (14). Due to symmetry, $\mathbb{E}[\boldsymbol{\Delta}_{\text{post}}^{\dot{\boldsymbol{x}}_i}|\ddot{\boldsymbol{x}}_i]$ is computed in a similar fashion, and hence will not be explicitly described. Two different cases must be studied: when $\boldsymbol{x}_i$ is labeled or unlabeled.

5.1. Additional features for unlabeled objects

Equation (7) shows that if we acquire $\ddot{\boldsymbol{x}}_i$ on a object previously characterized by $\dot{\boldsymbol{x}}_i$, matrix $\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}$ becomes

$$\begin{aligned}\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i} = \boldsymbol{\Delta}_{\text{post}} &+ \ddot{\lambda}\sum_{j\in\ddot{\mathcal{S}}}\ddot{K}_{ij}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{S}_{ij}\end{bmatrix} \\ &+ \lambda_{co}\begin{bmatrix}\dot{\boldsymbol{x}}_i\dot{\boldsymbol{x}}_i^T & -\dot{\boldsymbol{x}}_i\ddot{\boldsymbol{x}}_i^T\\ -\ddot{\boldsymbol{x}}_i\dot{\boldsymbol{x}}_i^T & \ddot{\boldsymbol{x}}_i\ddot{\boldsymbol{x}}_i^T\end{bmatrix},\end{aligned} \tag{16}$$

where

$$\begin{aligned}\mathbf{S}_{ij} = \ddot{\boldsymbol{x}}_i\ddot{\boldsymbol{x}}_i^T - \ddot{\boldsymbol{x}}_i\ddot{\boldsymbol{x}}_j^T \\ - \ddot{\boldsymbol{x}}_j\ddot{\boldsymbol{x}}_i^T + \ddot{\boldsymbol{x}}_j\ddot{\boldsymbol{x}}_j^T.\end{aligned} \tag{17}$$

To compute the conditional expectation $\mathbb{E}[\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}|\dot{\boldsymbol{x}}_i]$ (see (14)) we need a model for $p(\ddot{\boldsymbol{x}}_i|\dot{\boldsymbol{x}}_i)$. To this end, we use a Gaussian mixture model (GMM) to represent the joint density:

$$p(\ddot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_i) = \sum_c \pi_c \mathcal{N}\left(\dot{\boldsymbol{x}}|\dot{\boldsymbol{\mu}}_c, \dot{\boldsymbol{\Sigma}}_c\right)\mathcal{N}\left(\ddot{\boldsymbol{x}}|\ddot{\boldsymbol{\mu}}_c, \ddot{\boldsymbol{\Sigma}}_c\right).$$

Notice that, although using component-wise independence, this joint GMM globally models the dependency between $\dot{\boldsymbol{x}}$ and $\ddot{\boldsymbol{x}}$. From this joint GMM, it is straightforward to derive the conditional $p(\ddot{\boldsymbol{x}}_i|\dot{\boldsymbol{x}}_i)$, which is also a GMM, with weights that depend on $\dot{\boldsymbol{x}}$:

$$p(\ddot{\boldsymbol{x}}|\dot{\boldsymbol{x}}) = \sum_c \pi_c'(\dot{\boldsymbol{x}})\,\mathcal{N}\left(\ddot{\boldsymbol{x}}|\ddot{\boldsymbol{\mu}}_c, \ddot{\boldsymbol{\Sigma}}_c\right). \tag{18}$$

Further, the $\dot{K}_{ij}$ and $\ddot{K}_{ij}$ are set to Gaussian kernels; *e.g.* ,

$$\dot{K}_{ij} = \mathcal{N}\left(\dot{\boldsymbol{x}}_i|\dot{\boldsymbol{x}}_j, \boldsymbol{\Sigma}_\kappa\right). \tag{19}$$

Using (18), (19) and standard Gaussian identities, the required expectations are obtained analytically:

$$\mathbb{E}\left[\ddot{\boldsymbol{x}}_i|\dot{\boldsymbol{x}}_i\right] = \sum_c \pi_c'(\dot{\boldsymbol{x}}_i)\;\ddot{\boldsymbol{\mu}}_c = \boldsymbol{m}_1$$

$$\begin{aligned}\mathbb{E}\left[\ddot{\boldsymbol{x}}_i\ddot{\boldsymbol{x}}_i^T\middle|\,\dot{\boldsymbol{x}}_i\right] \\ = \sum_c \pi_c'(\dot{\boldsymbol{x}}_i)\left(\ddot{\boldsymbol{\mu}}_c\,\ddot{\boldsymbol{\mu}}_c^T + \ddot{\boldsymbol{\Sigma}}_c\right) = \mathbf{M}_2\end{aligned}$$

$$\mathbb{E}\left[\ddot{K}_{ij}\middle|\,\dot{\boldsymbol{x}}_i\right] = \sum_c \pi_c'(\dot{\boldsymbol{x}}_i)\;z_{cj} = m_{3j}$$

$$\mathbb{E}\left[\ddot{K}_{ij}\ddot{\boldsymbol{x}}_i\middle|\,\dot{\boldsymbol{x}}_i\right] = \sum_c \pi_c'(\dot{\boldsymbol{x}}_i)\;z_{cj}\;\boldsymbol{\mu}_{cj} = \boldsymbol{m}_{4j}$$

$$\begin{aligned}\mathbb{E}\left[\ddot{K}_{ij}\ddot{\boldsymbol{x}}_i\ddot{\boldsymbol{x}}_i^T\middle|\,\dot{\boldsymbol{x}}_i\right] \\ = \sum_c \pi_c'(\dot{\boldsymbol{x}}_i)\;z_{cj}\;\left(\ddot{\boldsymbol{\mu}}_{cj}\ddot{\boldsymbol{\mu}}_{cj}^T + \boldsymbol{\Lambda}_c\right) = \mathbf{M}_{5j}.\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\Lambda}_c &= \left(\ddot{\boldsymbol{\Sigma}}_c^{-1} + \boldsymbol{\Sigma}_\kappa^{-1}\right)^{-1} \\ \boldsymbol{\mu}_{cj} &= \boldsymbol{\Lambda}_c\left(\ddot{\boldsymbol{\Sigma}}_c^{-1}\ddot{\boldsymbol{\mu}}_c + \boldsymbol{\Sigma}_\kappa^{-1}\ddot{\boldsymbol{x}}_j\right)\end{aligned}$$

and

$$z_{cj} = (2\pi)^{-d/2} \, |\boldsymbol{\Lambda}_c|^{1/2} |\ddot{\boldsymbol{\Sigma}}_c|^{-1/2} |\boldsymbol{\Sigma}_\kappa|^{-1/2} \exp\left\{ -\frac{\ddot{\boldsymbol{\mu}}_c^T \ddot{\boldsymbol{\Sigma}}_c^{-1} \ddot{\boldsymbol{\mu}}_c + \ddot{\boldsymbol{x}}_j^T \boldsymbol{\Sigma}_\kappa^{-1} \ddot{\boldsymbol{x}}_j - \boldsymbol{\mu}_{cj}^T \boldsymbol{\Lambda}_c^{-1} \boldsymbol{\mu}_{cj}}{2} \right\}.$$

Finally,

$$\mathbb{E}\left[\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i} \middle| \dot{\boldsymbol{x}}_i \right] = \boldsymbol{\Delta}_{\text{post}} + \ddot{\lambda} \sum_{j \in \ddot{\mathcal{S}}} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{ij} \end{bmatrix} + \lambda_{co} \begin{bmatrix} \dot{\boldsymbol{x}}_i \dot{\boldsymbol{x}}_i^T & -\dot{\boldsymbol{x}}_i \boldsymbol{m}_1^T \\ -\boldsymbol{m}_1 \dot{\boldsymbol{x}}_i^T & \mathbf{M}_2 \end{bmatrix}, \quad (20)$$

where

$$\mathbf{S}_{ij} = m_{3j} \ddot{\boldsymbol{x}}_j \ddot{\boldsymbol{x}}_j^T - \ddot{\boldsymbol{x}}_j \boldsymbol{m}_{4j}^T - \boldsymbol{m}_{4j} \ddot{\boldsymbol{x}}_j^T + \mathbf{M}_{5j}.$$

Substituting (20) into (14) gives us our selection criterion.

5.2. Additional features for labeled objects

From (7), we can derive $\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}$ for the case when $\boldsymbol{x}_i$ is a labeled object ($\boldsymbol{x}_i \in \mathcal{D}_L$):

$$\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i} = \boldsymbol{\Delta}_{\text{post}} + \ddot{p}_i (1 - \ddot{p}_i) \begin{bmatrix} \mathbf{0} \\ \ddot{\boldsymbol{x}}_i \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \ddot{\boldsymbol{x}}_i \end{bmatrix}^T + \lambda_{co} \begin{bmatrix} \dot{\boldsymbol{x}}_i \dot{\boldsymbol{x}}_i^T & -\dot{\boldsymbol{x}}_i \ddot{\boldsymbol{x}}_i^T \\ -\ddot{\boldsymbol{x}}_i \dot{\boldsymbol{x}}_i^T & \ddot{\boldsymbol{x}}_i \ddot{\boldsymbol{x}}_i^T \end{bmatrix} + \ddot{\lambda} \sum_{j \in \ddot{\mathcal{S}}} \ddot{K}_{ij} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{ij} \end{bmatrix}. \quad (21)$$

Using standard Gaussian identities and the approximation

$$\sigma(z)\sigma(-z) \approx \mathcal{N}\left(z \,\middle|\, 0, \frac{8}{\pi} \right),$$

we can show that,

$$\mathbb{E}\left[\ddot{p}_i (1 - \ddot{p}_i) \, \ddot{\boldsymbol{x}}_i \ddot{\boldsymbol{x}}_i^T \middle| \dot{\boldsymbol{x}}_i \right] = \sum_c \pi_c'(\dot{\boldsymbol{x}}_i) \left(\boldsymbol{u}_c \boldsymbol{u}_c^T + \mathbf{U}_c \right) l_c = \mathbf{M}_6, \quad (22)$$

where

$$l_c = \mathcal{N}\left(\ddot{\boldsymbol{w}}^T \ddot{\boldsymbol{\mu}}_c \,\middle|\, 0, \frac{8}{\pi} + \ddot{\boldsymbol{w}}^T \ddot{\boldsymbol{\Sigma}}_c \ddot{\boldsymbol{w}} \right),$$

$$\mathbf{U}_c = \left(\ddot{\boldsymbol{\Sigma}}_c^{-1} + \frac{\pi}{8} \, \ddot{\boldsymbol{w}} \ddot{\boldsymbol{w}}^T \right)^{-1},$$

and $\boldsymbol{u}_c = \mathbf{U}_c \, \ddot{\boldsymbol{\Sigma}}_c^{-1} \ddot{\boldsymbol{\mu}}_c$. Finally, we can compute

$$\mathbb{E}\left[\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i} \middle| \dot{\boldsymbol{x}}_i \right] = \boldsymbol{\Delta}_{\text{post}} + \ddot{\lambda} \sum_{j \in \ddot{\mathcal{S}}} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{ij} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_6 \end{bmatrix} + \lambda_{co} \begin{bmatrix} \dot{\boldsymbol{x}}_i \dot{\boldsymbol{x}}_i^T & -\dot{\boldsymbol{x}}_i \boldsymbol{m}_1^T \\ -\boldsymbol{m}_1 \dot{\boldsymbol{x}}_i^T & \mathbf{M}_2 \end{bmatrix}$$

and substitute it into (14).

5.3. Sample requirements and practical approximations

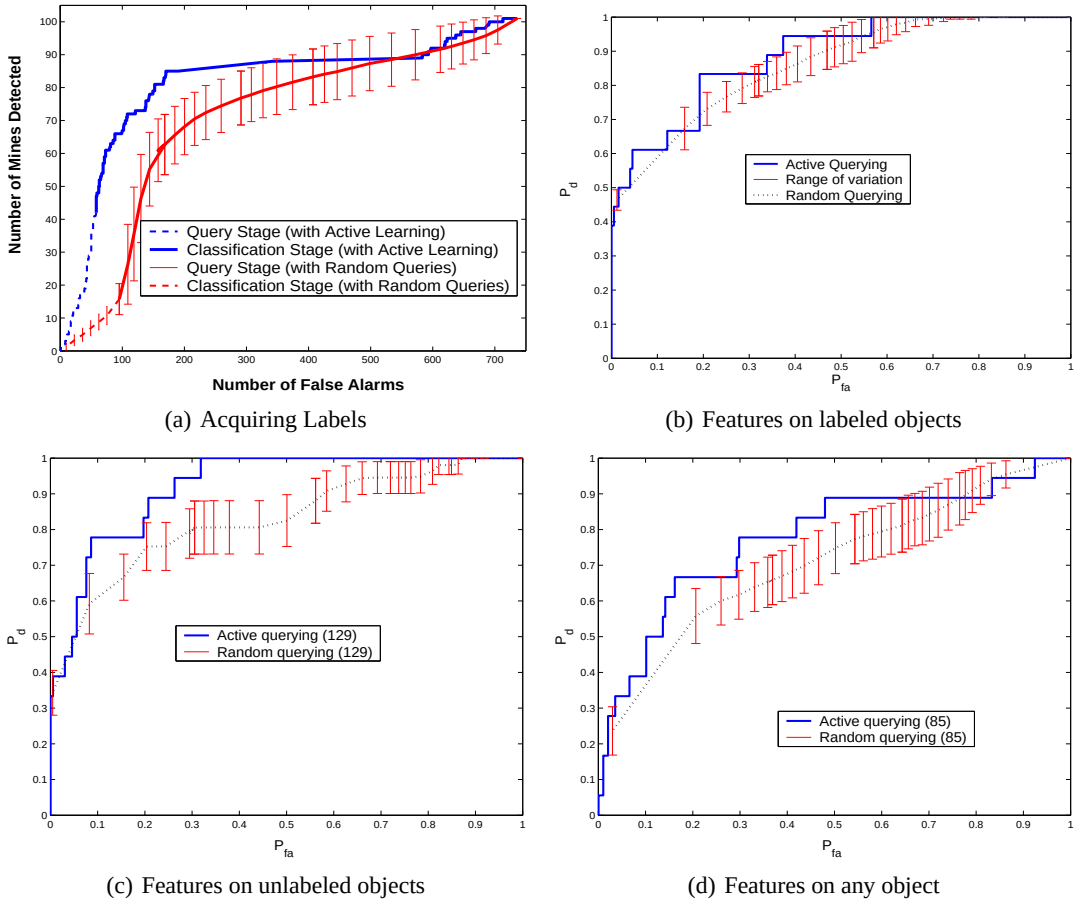
The conditional distribution (18) used to compute $\mathbb{E}[\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i} | \dot{\boldsymbol{x}}_i]$ in Sections 5.1 and 5.2 relies on a Gaussian mixture model (GMM) for $p(\ddot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_i)$. Unfortunately, fitting an accurate GMM demands a large number of samples; *i.e.* , $\dot{\ddot{\mathcal{S}}}$ must be large relative to $d_1 + d_2$. While our (unreported) studies on simulated data confirmed that the statistical methods proposed above work well when a sufficient number of samples is already available in $\dot{\ddot{\mathcal{S}}}$, in many real-life problems each sensor provides a large number of features, and the above requirement may not be satisfied (especially in early stages of the active learning process). The estimation of covariances is particularly problematic in these small-sample cases.

Due to this difficulty, in the results presented in the next section we use an alternative surrogate for $\mathbb{E}[\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}]$. Specifically, in the formulae for $\boldsymbol{\Delta}_{\text{post}}^{\ddot{\boldsymbol{x}}_i}$ ((16) and (21)) we simply replace $\ddot{\boldsymbol{x}}_i$ with $\boldsymbol{m}_1 = \mathbb{E}[\ddot{\boldsymbol{x}}_i | \dot{\boldsymbol{x}}_i]$—which can still be reliably estimated from limited data, since it does not involve covariances—and subsequently compute the determinant of the resulting matrix. As demonstrated in the next section, this approximation still yields very good experimental results as compared to the random acquisition of additional features.

6. Experiments: Multi-view feature acquisition vs. label acquisition

To evaluate the methods proposed in this paper, we use the same data used in Krishnapuram et al. (2004) to study the performance of co-training and active label acquisition algorithms. Mirroring their experimental setup, we also operate our algorithms transductively, testing the accuracy of the classifier on the same unlabeled data used for semi-supervised training. In brief, the goal was to detect surface and subsurface land mines, using two sensors: (1) a 70-band hyper-spectral electro-optic (EOIR) sensor which provides 420 features; and (2) an X-band synthetic aperture radar (SAR) which provides 9 features. Our choice of dataset was influenced by two factors: lack of other publicly available multi-sensor datasets; a need to compare the benefits of the proposed active feature acquisition strategy against the benefits of adaptive label querying methods.

The results for active feature acquisition on the unlabeled samples (Section 5.1), and on the labeled samples (Section 5.2) are shown in Figure 2. Additionally we let the algorithm automatically decide whether to query additional features on labeled or unlabeled data at each iteration, based on the bound on mutual information for the best candidate query in each case. The results for this are also pro-



(a) Acquiring Labels

(b) Features on labeled objects

(c) Features on unlabeled objects

(d) Features on any object

Figure 2. Multi-sensor adaptive data acquisition with EOIR and SAR features. (a) (dotted) Number of land mines detected during the querying for 100 labels (solid) ROC for the remaining objects. Reproduced from Krishnapuram et al. (2004). (b) ROC after acquiring 27 additional feature sets for incompletely characterized labeled objects. (c) ROC after acquiring 129 additional feature sets for incompletely characterized unlabeled objects. (d) ROC after acquiring 85 features for either labeled or unlabeled objects. Error bars represent one s.d. from the mean.

vided in Figure 2. In all cases, for a baseline comparison, we also provide average ROCs for 100 trials with random querying, with error bars representing one standard deviation from the mean. For additional insight, we also reproduce the results from Krishnapuram et al. (2004) for active label query selection (Section 4) on the same data.

Analysis of results: all the adaptive data acquisition algorithms show significant benefits over the baseline random methods. Nevertheless, as compared to random sample query selection, active learning exhibits maximum additional benefits in two scenarios: label query selection and additional feature acquisition on the unlabeled samples.

Since labeled data is more valuable than unlabeled data, the intelligent choice of a small set of additional label queries improves the classifier performance most. The acquisition of additional features on the unlabeled data also serves to disambiguate the most doubtful test objects, in addition to improving the classifier itself. Since the labeled data do not need further disambiguation, we expect active acquisition of features for labeled objects to exhibit a smaller (but still statistically significant) improvement in accuracy, especially in a transductive experimental setting. We have verified these intuitions by experimentally querying a varying number of objects in each case, although we present only one example result in Figure 2.

7. Conclusions

Using simple but practical approximations, this paper relies on an information-theoretic criterion to answer the question: Which feature sensor should be used to make measurements on objects in order to accurately design multi-sensor classifiers? Since a sensor may be used to obtain

more than one feature simultaneously, this is a more general problem than that of greedily choosing which feature must be obtained in a myopic way, although it subsumes the latter problem as a special case (especially in supervised settings when co-training effects are ignored by fixing $\lambda_{co} = 0$). Despite the potentially wide applicability, we have not seen this question addressed systematically in the literature. Results on measured data indicate that the proposed criterion for adaptive characterization of unlabeled objects significantly improves classifier accuracy; results using the corresponding criterion for labeled objects are less impressive though.

In learning a classifier, one attempts to minimize the error rate on an infinite set of future test samples drawn from the underlying data-generating distribution. However, in transductive settings, one may sometimes only care about classifying the unlabeled training samples. Future work includes extensions of the ideas proposed here to automatically select the sensor whose deployment will most improve the accuracy on *the remaining unlabeled training samples*, instead of attempting to learn accurate classifiers.

We will also consider non-myopic active learning strategies that evaluate the benefits of improved classification accuracy in a setting that explicitly considers both the cost of obtaining class labels and the costs involved in using various sensors to make feature measurements. This would allow us to automatically decide which of the following is the best course of action in any situation: (a) obtain many individually less effective feature measurements (with regard to improving the classification accuracy) using a cheap sensor; or (b) obtain fewer, but more useful feature measurements using an alternative, costlier sensor; or (c) obtain a small number of additional class labels at a significant cost.

References

Balcan, M.-F., Blum, A., & Yang, K. (2004). Co-training and expansion: Towards bridging theory and practice. *NIPS*. Cambridge, MA: MIT Press.

Belkin, M., Niyogi, P., & Sindhwani, V. (2004). *Manifold learning: a geometric framework for learning from examples* (Technical Report). Dept. Computer Science, U. of Chicago.

Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. *Proceedings of the 18th International Conference on Machine Learning*.

Blum, A., & Mitchell, T. (1998). Combining labelled and unlabelled data with co-training. *Proc. Eleventh Annual Conference on Computational Learning Theory (COLT) 1998*.

Böhning, D. (1992). Multinomial logisitic regression algorithm. *Annals of the Institute of Statistical Mathematics*, *44*, 197–200.

Boyd, S., & Vandenberghe, L. (2003). *Convex Optimization*. Cambridge University Press.

Brefeld, U., & Scheffer, T. (2004). Co-EM support vector learning. *Proceedings of the Twenty-First International Conference on Machine Learning – ICML*.

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Corduneanu, A., & Jaakkola, T. (2004). Distributed information regularization on graphs. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Dasgupta, S., Littman, M., & McAllester, D. (2001). Pac generalization bounds for co-training. *Proc. Neural Info. Processing Systems NIPS*.

Inoue, M., & Ueda, N. (2003). Exploitation of unlabelled sequences in hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 1570–1581.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proceedings of the 16th International Conference on Machine Learning* (pp. 200–209). San Francisco: Morgan Kaufmann.

Joachims, T. (2003). Transductive learning via spectral graph partitioning. *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*.

Krishnapuram, B., Williams, D., Xue, Y., Hartemink, A., Carin, L., & Figueiredo, M. (2004). On semi-supervised classification. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 589–603.

Melville, P., Saar-Tsechansky, M., Provost, F., & Mooney, R. J. (2004). Active feature acquisition for classifier induction. *Proceedings of the Fourth International Conference on Data Mining (ICDM-2004)* (p. (to appear)).

Muslea, I., Minton, S., & Knoblock, C. (2000). Selective sampling with redundant views. *Proc. of National Conference on Artifical Intelligence* (pp. 621–626).

Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning Journal*, *39*, 103–134.

Seeger, M. (2001). *Learning with labelled and unlabelled data* (Technical Report). Institute for Adaptive and Neural Computation, University of Edinburgh, Edinburgh, UK.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*, 45–66.

Zheng, Z., & Padmanabhan, B. (2002). On active learning for data acquisition. *Proceedings of the Second International Conference on Data Mining (ICDM-2002)*.

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). *Semi-supervised learning: From Gaussian fields to Gaussian processes* (Technical Report CMU-CS-03-175). School of CS, CMU.