Active Sensing

Shipeng Yu, Balaji Krishnapuram, Romer Rosales, R. Bharat Rao
CAD and Knowledge Solutions, Siemens Medical Solutions USA, Inc.
{firstname.lastname}@siemens.com

Abstract

Labels are often expensive to get, and this motivates *active learning* which chooses the most informative samples for label acquisition. In this paper we study *active sensing* in a multi-view setting, motivated from many problems where grouped features are also expensive to obtain and need to be acquired (or *sensed*) actively (e.g., in cancer diagnosis each patient might go through many tests such as CT, Ultrasound and MRI to get valuable features). The strength of this model is that one actively sensed (sample, view) pair would improve the *joint* multi-view classification on all the samples. For this purpose we extend the Bayesian co-training framework such that it can handle missing views in a principled way, and introduce two criteria for view acquisition. Experiments on one toy data and two real-world medical problems show the effectiveness of this model.

1 Introduction

Labeled data can be expensive to obtain in a variety of machine learning problems. *Active learning* addresses the problem of efficiently choosing data samples to be labeled in order to improve overall learning performance. From a cancer diagnosis perspective, this is equivalent to choosing patients to do a biopsy such that the tumor is correctly diagnosed (benign/malignant). In this paper we consider a related but different problem, now motivated by the fact that features may also be expensive to obtain (much in the same way as labels). More generally we consider subsets of features; we refer to them as *views*. In cancer diagnosis, features could come from different imaging modalities such as CT, Ultrasound and MRI. In problems where there exist different views of the data, some of these views could be missing for certain samples (due to, e.g., high cost or limited budget). We call *active sensing* the process of efficiently choosing what views *and* samples to additionally acquire to improve the overall learning performance (cf. Fig. 1).

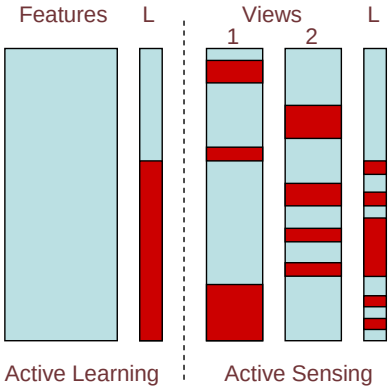

Figure 1: Settings for active learning (left) and active sensing with 2 views (right). The "L" column denotes the labels (outputs). Light blue blocks denote observed data, and red blocks denote missing data.

Examples of the active sensing setting described above are abundant. For land mine detection in a sensor network, we may have different types of sensors (as different views) deployed at one location, but some sensors may not be available for all locations due to high cost. So the interest is to decide which location and which type of sensor we should additionally consider to achieve better detection accuracy. In the medical diagnosis scenario, our motivating application, specialists rely on different sets of medical factors, such as demographics, imaging, and bio-markers, to make clinical decisions. A patient does not undergo all possible tests

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

at once (due to various side effects such as radiation and contrast), but these tests are selected based on the evidence collected up to a particular point. [1]

It is seen that standard active learning would not work in this setting. When some views are missing, one solution is to learn a model using those samples for which all the views are available. Another solution is to impute the missing features using the observed features. However, as the main motivating factor for *multi-view learning* approaches, the information provided by the combined set of views (taken at once) is in general larger than that provided by any algorithm that considers the views separately (e.g., views can reinforce each other).

We provide two approaches for efficiently choosing the (sample, view) pair, based on the mutual information (involving various random variables) and on the predictive uncertainty, respectively. We formalize these within the recently proposed Bayesian co-training framework (Yu et al., 2008), with an important extension to account for data with missing view information. This provides an undirected graphical model representation of the active sensing problem. We also provide methods for addressing density modeling and approximated inference sub-problems arising in this probabilistic setting. Empirical studies using one toy data and two real-world medical problems clearly show the effectiveness of this model.

The rest of the paper is organized as follows. We survey the related literature in Section 2. The Bayesian co-training model is extended to handle missing views in Section 3. Section 4 describes two methods for active sensing, i.e., deciding which incomplete samples should be further characterized, and which sensors should be deployed on them. Experimental results are provided in Section 5. We conclude with a brief discussion and future work in Section 6.

2 Related Work

Active sensing provides a new scenario in active data acquisition. The present formulation benefits from previous work in experiment design (Lindley, 1956; Fedorov, 1972), active learning (MacKay, 1992; Seung et al., 1992), and sensor placement (Krause et al., 2008). While there exist a considerable body of work on the general notion of active data acquisition, to the best of our knowledge, this is the first paper to focus on this notion of active sensing—feature acquisition—specifically for improving multi-view learning *jointly* for all unlabeled samples.

Feature acquisition was addressed in, e.g., (Melville et al., 2004; Bilgic and Getoor, 2007), but there is a clear difference to active sensing. Previous feature acquisition only considers one sample at a time, i.e., when one sample is in consideration, the other samples are not affected. But in active sensing, one actively acquired (sample, view) pair will improve the classification performance of *all* the unlabeled samples via a co-training setting. A related yet different problem was considered to identify the optimal spatial locations for placing a single type of sensor to model spatially varying phenomena (Krause et al., 2008); however, this work addressed the use of a single type of sensor, and did not consider the scenario of multiple views.

Co-training (Blum and Mitchell, 1998) is based on the idea that the error rate on unseen test samples can be upper bounded by the disagreement between the classification-decisions obtained from independent characterizations (views) of the data (Dasgupta et al., 2001). Recently, Bayesian co-training (Yu et al., 2008) was proposed which defines an undirected graphical model for co-training and provides a principled solution to multi-view learning. However it can only handle data without missing views, and in this paper we extend it such that it provides the basis for active sensing.

One of our criteria for active sensing is to choose the (sample, view) pair which provides the maximum mutual information (MI) (Cover and Thomas, 1991) about the non-parametric classification function. In order to accomplish this, we use the D-optimality criterion, while other choices such as A-optimality and E-optimality are also available (Flaherty et al., 2006). Apart from MI maximization, other objective criteria for active learning include *uncertainty sampling* (Lewis and Gale, 1994; Cohn et al., 1996) and performance optimization (e.g., (Roy and McCallum, 2001)).

In this paper we make two principal contributions. First, we extend Bayesian co-training to allow for missing views, accommodating incompletely characterized objects. Deploying additional sensors to characterize an object would naturally help improving classification accuracy. Our second contribution is to identify which objects should be characterized using additional sensors in order to improve the classification of *all the unlabeled data*. This is significantly different from previous feature acquisition work.

3 Bayesian Co-Training with Missing Views

Bayesian co-training defines an undirected graphical model for semi-supervised multi-view learning (Yu

[1] This is normally referred to as differential diagnosis.

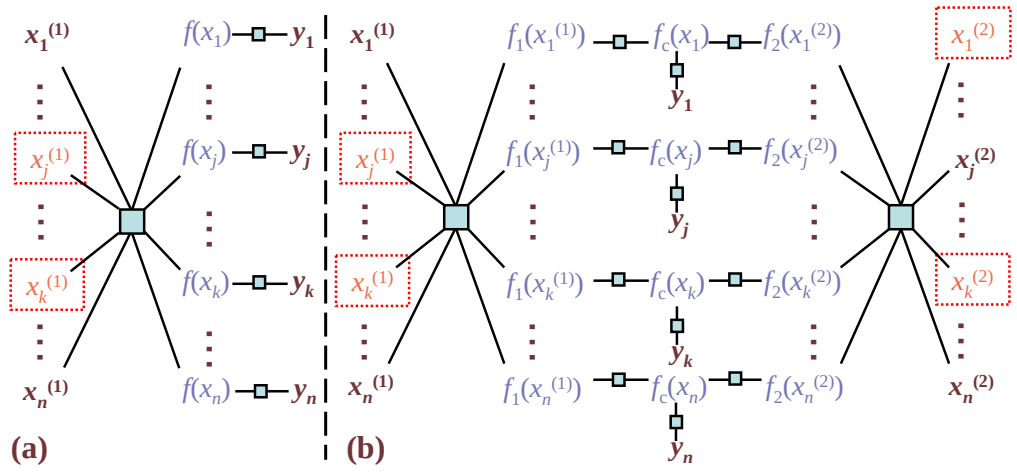

Figure 2: Bayesian co-training factor graphs for (a) one-view and (b) two-view problems, with missing views. Observed variables are marked as dark/bold, and unobserved ones are marked as red/non-bold, including functions f_1, f_2, f_c (blue/non-bold). Unobserved variables in a dotted box (such as $\mathbf{x}_j^{(1)}$) are potential observations for active sensing. All labels y are denoted as observed in the graph, but this is not required.

et al., 2008). The original work assumes that the input data are complete, i.e., all the views are observed for every data sample, but for active sensing we need to define a co-training strategy for data with *incomplete* or *missing views*. In this section we extend Bayesian co-training to the case where there are missing (sample, view) pairs in the input data. The same notations as in (Yu et al., 2008) are preserved unless otherwise mentioned.

Suppose we have m different views of a set of n data samples. Let $\mathbf{x}_i^{(j)} \in \mathbb{R}^{d_j}$ be the features for the ith sample obtained using the jth view, where d_j is the dimensionality of the input space for view j.[2] Let each view j be observed for a subset of n_j samples, and let $\mathbb{I}_j$ denote the indices of these samples in the whole sample set. Finally let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ denote the labels for these samples. In this paper we consider a binary classification scenario where each $y_i \in \{-1, +1\}$.

In Bayesian co-training, let f_j denote the latent function for the jth view, and $f_j \sim \mathcal{GP}(0, \kappa_j)$ be its GP prior in view j. The consensus function f_c is defined to ensure conditional independence between the output y and the m latent functions $\{f_j\}$ (Yu et al., 2008) (cf. Fig. 2 for the factor graph). The undirected graphical model leads to the following joint probability:

$$p(\mathbf{y}, \mathbf{f}_c, \mathbf{f}_1, \ldots, \mathbf{f}_m) = \frac{1}{Z} \prod_{i=1}^{n} \psi(y_i, f_c(\mathbf{x}_i)) \prod_{j=1}^{m} \psi(\mathbf{f}_j) \psi(\mathbf{f}_j, \mathbf{f}_c), \quad (1)$$

where $\mathbf{f}_c = \{f_c(\mathbf{x}_i)\}_{i=1}^n$ and $\mathbf{f}_j = \{f_j(\mathbf{x}_i^{(j)})\}_{i \in \mathbb{I}_j}$ are column vectors of length n and n_j, respectively. Note that unlike in (Yu et al., 2008), $\mathbf{f}_j$ is only realized on a subset of samples (as denoted in $\mathbb{I}_j$) and is of length n_j (instead of n). The *within-view potential* $\psi(\mathbf{f}_j)$ is defined via the GP prior,

$$\psi(\mathbf{f}_j) = \exp(-\frac{1}{2} \mathbf{f}_j^\top \mathbf{K}_j^{-1} \mathbf{f}_j),$$

where $\mathbf{K}_j \in \mathbb{R}^{n_j \times n_j}$ is the covariance matrix for view j; the *consensus potential* $\psi(\mathbf{f}_j, \mathbf{f}_c)$ describes how each latent function $\mathbf{f}_j$ is related to the consensus function $\mathbf{f}_c$, which we define as follows:

$$\psi(\mathbf{f}_j, \mathbf{f}_c) = \exp\left(-\|\mathbf{f}_j - \mathbf{f}_c(\mathbb{I}_j)\|^2 \Big/ 2\sigma_j^2\right). \quad (2)$$

Note that $\mathbf{f}_c(\mathbb{I}_j)$ takes the length-n_j subset of vector $\mathbf{f}_c$ with indices given by $\mathbb{I}_j$. The idea here is to define the consensus potential for view j using only the data samples observed in view j. As in (Yu et al., 2008), $\sigma_j > 0$ quantifies how far the latent function $\mathbf{f}_j$ is apart from $\mathbf{f}_c$, and the output potential $\psi(y_i, f_c(\mathbf{x}_i))$ is defined as $\lambda(y_i f_c(\mathbf{x}_i))$ with logistic function $\lambda(z) = (1 + \exp(-z))^{-1}$.

3.1 Co-Training Kernel with Missing Views

As in (Yu et al., 2008), we can also derive a co-training kernel $\mathbf{K}_c$ by integrating out all the latent functions $\{\mathbf{f}_j\}$ in (1). It is calculated as $\mathbf{K}_c = \mathbf{\Lambda}_c^{-1}$, $\mathbf{\Lambda}_c = \sum_{j=1}^m \mathbf{A}_j$, and each $\mathbf{A}_j$ is a $n \times n$ matrix as

$$\mathbf{A}_j(\mathbb{I}_j, \mathbb{I}_j) = (\mathbf{K}_j + \sigma_j^2 \mathbf{I})^{-1}, \text{ and 0 otherwise.} \quad (3)$$

That is, $\mathbf{A}_j$ is an expansion of one-view information matrix $(\mathbf{K}_j + \sigma_j^2 \mathbf{I})^{-1}$ to the full size $n \times n$, with the

[2] Note that subscripts index the data sample, and superscripts (with round brackets) index the view.

other (unindexed) entries filled with 0. It is easily seen that such a kernel $\mathbf{K}_c$ is indeed positive definite as long as each one-view kernel $\mathbf{K}_j$ is positive definite. Very importantly, we note that *one additional observation of a (sample, view) pair will affect all the elements of the co-training kernel.* This is exactly the property we would like to have in active sensing.

3.2 Co-Regularization with Missing Views

To be complete we also give the marginalization result for co-regularization. Ignoring the output $\mathbf{y}$ for the moment, integrating out the consensus view $\mathbf{f}_c$ leads to the following joint prior:

$$p(\mathbf{f}_1,\ldots,\mathbf{f}_m) = \frac{1}{Z}\exp\left\{-\frac{1}{2}\sum_{j=1}^{m}\mathbf{f}_j^\top\mathbf{K}_j^{-1}\mathbf{f}_j - \frac{1}{2}\sum_{j<k}\sum_{\mathbf{x}\in\mathbb{I}_j\wedge\mathbb{I}_k}\left[\frac{[f_j(\mathbf{x})-f_k(\mathbf{x})]^2}{\sigma_j^2\sigma_k^2}\Big/\sum_{\mathbb{I}_\ell\ni\mathbf{x}}\frac{1}{\sigma_\ell^2}\right]\right\}.$$

The first part regularizes the functional space of each view, and the second part constrains that every pair of views need to agree on the outputs for *co-observed* samples (inversely weighted by view variances and the sum of precisions of the views in which the sample is observed).

4 Active Sensing

In active sensing, we are interested in selecting the best unobserved (sample, view) pair for sensing, or for view acquisition, which will improve the overall classification performance. In this section we mainly discuss an approach based on the mutual information framework, which measures the expected information gain after observing an additional (sample, view) pair. Another approach based on the predictive uncertainty is also briefly discussed in Section 4.5. In the following let $\mathcal{D}_O$ and $\mathcal{D}_U$ denote the observed and unobserved (sample, view) pairs, respectively.

4.1 Laplace Approximation

To calculate the mutual information we need to calculate the differential entropy of the consensus view function $\mathbf{f}_c$. With co-training kernel and the logistic regression loss, Laplace approximation can be applied to approximate the posterior distribution of $\mathbf{f}_c$ as a Gaussian distribution. In particular, let the prior of the consensus view take the GP prior with co-training kernel, i.e., $\mathbf{f}_c \sim \mathcal{N}(0,\mathbf{K}_c)$. With the logistic regression loss, the *a posteriori* distribution of $\mathbf{f}_c$, $p(\mathbf{f}_c|\mathcal{D}_O,\mathbf{y})$, is approximately

$$\mathcal{N}(\hat{\mathbf{f}}_c, (\boldsymbol{\Delta}_{\text{post}})^{-1}), \tag{4}$$

where $\hat{\mathbf{f}}_c$ is the maximum *a posteriori* (MAP) estimate of $\mathbf{f}_c$, and the posterior precision matrix $\boldsymbol{\Delta}_{\text{post}} = \mathbf{K}_c^{-1} + \boldsymbol{\Phi}$, with $\boldsymbol{\Phi}$ the Hessian of the negative log-likelihood. It turns out that $\boldsymbol{\Phi}$ is a diagonal matrix, with $\boldsymbol{\Phi}(i,i) = \eta_i(1-\eta_i)$ where $\eta_i = \lambda(\hat{\mathbf{f}}_c(\mathbf{x}_i))$. The differential entropy of $\mathbf{f}_c$ under this Laplace approximation is

$$H(\mathbf{f}_c) = -\frac{n}{2}\log(2\pi e) - \frac{1}{2}\log\det(\boldsymbol{\Delta}_{\text{post}}),$$

where $\det(\cdot)$ denote matrix determinant.

4.2 Mutual Information for Active Sensing

Remind that $\mathbf{x}_i^{(j)}$ denote the features in the jth view for the ith sample. In active sensing, the mutual information (MI) between the consensus view function $\mathbf{f}_c$ and the unobserved (sample, view) pair $\mathbf{x}_i^{(j)} \in \mathcal{D}_U$ is the *expected decrease in entropy of* $\mathbf{f}_c$ when $\mathbf{x}_i^{(j)}$ is observed,

$$\begin{aligned} I(\mathbf{f}_c, \mathbf{x}_i^{(j)}) &= \mathbb{E}[H(\mathbf{f}_c)] - \mathbb{E}[H(\mathbf{f}_c|\mathbf{x}_i^{(j)})] \\ &= -\frac{1}{2}\log\det(\boldsymbol{\Delta}_{\text{post}}) + \frac{1}{2}\,\mathbb{E}\left[\log\det(\boldsymbol{\Delta}_{\text{post}}^{x(i,j)})\right], \end{aligned}$$

where the expectation is with respect to $p(\mathbf{x}_i^{(j)}|\mathcal{D}_O,\mathbf{y})$, the distribution of the unobserved (sample, view) pair given all the observed pairs and available outputs. $\boldsymbol{\Delta}_{\text{post}}^{x(i,j)}$ is the *a posteriori* precision matrix, derived from Section 4.1, after one pair $\mathbf{x}_i^{(j)}$ is observed.

The maximum MI criterion has been used before to identify the "best" unlabeled sample in active learning (MacKay, 1992). Here we adopt this criterion and choose the unobserved pair which maximizes MI:

$$\begin{aligned} (i^*, j^*) &= \arg\max_{\mathbf{x}_i^{(j)}\in\mathcal{D}_U} I(\mathbf{f}_c, \mathbf{x}_i^{(j)}) \\ &= \arg\max_{\mathbf{x}_i^{(j)}\in\mathcal{D}_U} \mathbb{E}\left[\log\det(\boldsymbol{\Delta}_{\text{post}}^{x(i,j)})\right]. \end{aligned} \tag{5}$$

4.3 Density Modeling

In order to calculate the expectation in (5), we need a conditional density model for the unobserved pairs, i.e., $p(\mathbf{x}_i^{(j)}|\mathcal{D}_O,\mathbf{y})$. This of course depends on the type of the features in each view, and in this paper we use a special Gaussian mixture model (GMM). Let the joint input density be

$$\begin{aligned} p(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}) &= p(y=+1)p(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}|y=+1) \\ &\quad + p(y=-1)p(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}|y=-1), \end{aligned}$$

and each conditional density takes a *component-wise factorized* GMM form, e.g., for positive class,

$$p(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}|y=+1) = \sum_c \pi_c^+ \prod_j \mathcal{N}(\mathbf{x}^{(j)}|\boldsymbol{\mu}_c^{+(j)}, \boldsymbol{\Sigma}_c^{+(j)}).$$

Here $\boldsymbol{\mu}_c^{+(j)}$ and $\boldsymbol{\Sigma}_c^{+(j)}$ are the mean and covariance for view j in component c, and $\pi_c^+ > 0$, $\sum_c \pi_c^+ = 1$ are the mixture weights for the positive class. Note that although the conditional density for each mixture component is decoupled for different views, the joint conditional density is not.[3] Under this model, the joint density $p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)})$ is also a GMM, and any marginal (conditioned on y or not) density is still a GMM, e.g., $p(\mathbf{x}^{(j)}|y=+1) = \sum_c \pi_c^+ \mathcal{N}(\mathbf{x}^{(j)}|\boldsymbol{\mu}_c^{+(j)}, \boldsymbol{\Sigma}_c^{+(j)})$.

Now it is easy to calculate $p(\mathbf{x}_i^{(j)}|\mathcal{D}_O, \mathbf{y})$. Let $\mathbf{x}_i^{(O)}$ be the set of observed views for $\mathbf{x}_i$, we need to distinguish two different settings. When the label y_i is available, e.g., $y_i = +1$, we have

$$\begin{aligned} p(\mathbf{x}_i^{(j)}|\mathcal{D}_O, \mathbf{y}) &= p(\mathbf{x}_i^{(j)}|\mathbf{x}_i^{(O)}, y_i = +1) \\ &= \sum_c \pi_c^{+(j)}(\mathbf{x}_i^{(O)}) \cdot \mathcal{N}(\mathbf{x}_i^{(j)}|\boldsymbol{\mu}_c^{+(j)}, \boldsymbol{\Sigma}_c^{+(j)}), \quad (6) \end{aligned}$$

which is again a GMM model, with the mixing weights

$$\pi_c^{+(j)}(\mathbf{x}_i^{(O)}) = \pi_c^+ \frac{\prod_{k\in O} \mathcal{N}(\mathbf{x}_i^{(k)}|\boldsymbol{\mu}_c^{+(k)}, \boldsymbol{\Sigma}_c^{+(k)})}{p(\mathbf{x}_i^{(O)}|y_i = +1)}.$$

When the label y_i is not available, we need to integrate out the labeling uncertainty and compute

$$\begin{aligned} p(\mathbf{x}_i^{(j)}|\mathcal{D}_O, \mathbf{y}) &= p(\mathbf{x}_i^{(j)}|\mathbf{x}_i^{(O)}) \\ &= p(y_i = +1)p(\mathbf{x}_i^{(j)}|\mathbf{x}_i^{(O)}, y_i = +1) \\ &\quad + p(y_i = -1)p(\mathbf{x}_i^{(j)}|\mathbf{x}_i^{(O)}, y_i = -1), \end{aligned}$$

which is a GMM model as well, as seen from (6).

4.4 Expectation Calculation

We are now ready to compute the expectation in (5). The *a posteriori* precision matrix after one (sample, view) pair $\mathbf{x}_i^{(j)}$ is observed, $\boldsymbol{\Delta}_{\text{post}}^{x(i,j)}$, is calculated as

$$\begin{aligned} \boldsymbol{\Delta}_{\text{post}}^{x(i,j)} &= \boldsymbol{\Phi} + (\mathbf{K}_c^{x(i,j)})^{-1} \\ &= \boldsymbol{\Phi} + \mathbf{A}_j^{x(i,j)} + \sum_{k\neq j} \mathbf{A}_k, \quad (7) \end{aligned}$$

where $\mathbf{K}_c^{x(i,j)}$ and $\mathbf{A}_j^{x(i,j)}$ are the new $\mathbf{K}_c$ and $\mathbf{A}_j$ matrices after the new pair is observed. Based on (3), to calculate $\mathbf{A}_j^{x(i,j)}$ we need to recalculate the kernel for the jth view, $\mathbf{K}_j$, after an additional pair $\mathbf{x}_i^{(j)}$ is observed. This is simply done by adding one more row and column to the old $\mathbf{K}_j$ as:

$$\mathbf{K}_j^{x(i,j)} = \begin{bmatrix} \mathbf{K}_j & \mathbf{b}_j \\ \mathbf{b}_j^\top & a_j \end{bmatrix},$$

where $a_j = \kappa_j(\mathbf{x}_i^{(j)}, \mathbf{x}_i^{(j)}) \in \mathbb{R}$, and $\mathbf{b}_j \in \mathbb{R}^{n_j}$ has the ℓth entry as $\kappa_j(\mathbf{x}_\ell^{(j)}, \mathbf{x}_i^{(j)})$. Then from (3), the non-zero part of $\mathbf{A}_j^{x(i,j)}$ is calculated as

$$\begin{aligned} \left(\mathbf{K}_j^{x(i,j)} + \sigma_j^2\mathbf{I}\right)^{-1} &= \begin{bmatrix} \mathbf{K}_j + \sigma_j^2\mathbf{I} & \mathbf{b}_j \\ \mathbf{b}_j^\top & a_j + \sigma_j^2 \end{bmatrix}^{-1} \quad (8) \\ &= \begin{bmatrix} \boldsymbol{\Gamma}_j + \lambda_j \boldsymbol{\Gamma}_j \mathbf{b}_j \mathbf{b}_j^\top \boldsymbol{\Gamma}_j & -\lambda_j \boldsymbol{\Gamma}_j \mathbf{b}_j \\ -\lambda_j \mathbf{b}_j^\top \boldsymbol{\Gamma}_j & \lambda_j \end{bmatrix}, \end{aligned}$$

using the block-matrix inverse formula, where $\boldsymbol{\Gamma}_j = (\mathbf{K}_j + \sigma_j^2\mathbf{I})^{-1}$ and $\lambda_j = \frac{1}{a_j + \sigma_j^2 - \mathbf{b}_j^\top \boldsymbol{\Gamma}_j \mathbf{b}_j}$.

As seen from (7) and (8), it is difficult to directly calculate the expectation in (5). Since for any matrix $\mathbf{Q}$, $\mathbb{E}[\log\det(\mathbf{Q})] \leq \log\det(\mathbb{E}[\mathbf{Q}])$ due to the concavity of $\log\det(\cdot)$, we alternatively take the upper bound $\log\det(\mathbb{E}[\boldsymbol{\Delta}_{\text{post}}^{x(i,j)}])$ as the selection criteria. From (7) and (8), this reduces to computing $\mathbb{E}[\lambda_j], \mathbb{E}[\lambda_j \mathbf{b}_j]$ and $\mathbb{E}[\lambda_j \mathbf{b}_j \mathbf{b}_j^\top]$, where the expectations are with respect to $p(\mathbf{x}_i^{(j)}|\mathcal{D}_O, \mathbf{y})$, a GMM model (cf. Section 4.3). In general one needs to calculate these expectations numerically, as different kernel functions lead to different integrals. As another approximation one might assume each of the GMM component is a point-mass such that the mean is used for the calculation.

4.5 Discussion

The mutual information based approach directly measures the expected information gain for every (sample, view) pair. A different (and simpler) approach is based on the predictive uncertainty, in which the most *uncertain* sample (after the current classifier is trained) is selected for view acquisition (see also (Melville et al., 2004)). This uncertainty (i.e., predictive variance) is estimated as the diagonal entries of the *a posteriori* covariance matrix $(\boldsymbol{\Delta}_{\text{post}})^{-1}$, as seen from (4). However it is not clear what view to acquire for this sample (if more than one view is missing for the sample). The advantage of this approach is that no density modeling is necessary for unobserved views.

5 Empirical Study

For the following experiments we are given a classification task with missing views. At each iteration we are allowed to select an unobserved (sample, view) pair for sensing (i.e., feature acquisition). We compare the classification performance on unlabeled data using the following three sensing approaches:

- **Active Sensing MI**: The pair is selected based on the mutual information criteria (5).
- **Active Sensing VAR**: A sample is selected first which has the maximal predictive variance and has

[3] A straightforward EM algorithm can be derived to estimate all these parameters. When labels are only available for a very limited number of samples, one might assume a full generative GMM model neglecting the dependency on labels (instead of a conditional GMM model).

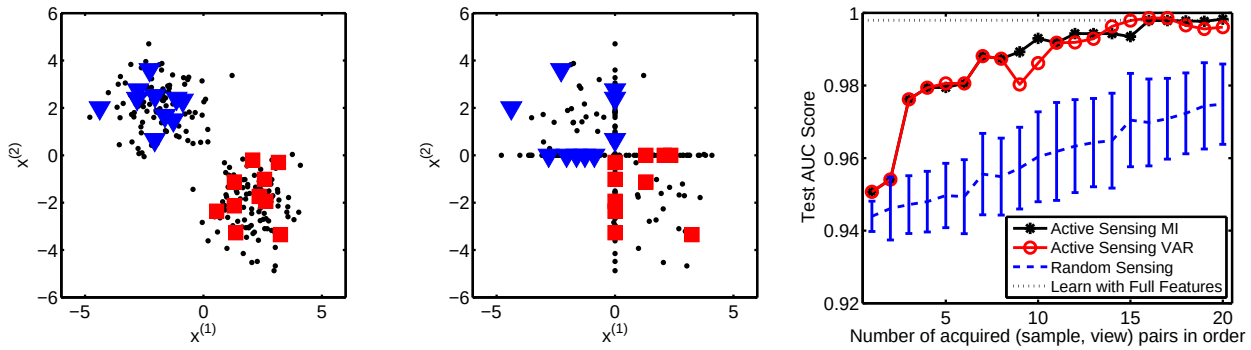

Figure 3: Toy example for active sensing (left). Big red-square/blue-triangular markers denote $+1/-1$ labeled points; remaining points are unlabeled. Data are sampled from two Gaussians with mean $(2,-2)$, $(-2,2)$ and unit variance. After "hiding" some of the features the data look like (middle) with removed features replaced with 0. Comparison of active sensing with random sensing is shown on the right.

missing views, and then one of the missing views is randomly selected for sensing.

• **Random Sensing**: A random unobserved (sample, view) pair is selected for sensing.

After the pair is acquired in each iteration, learning is done using the Bayesian co-training model. Note that for all the three approaches, the acquired (sample, view) pair will affect all the samples in the next iteration (via the co-training kernel). In active sensing with MI, we use EM algorithm to learn the GMM structure with missing entries, and the GMM model is re-estimated after each pair is selected and filled in (this is fast thanks to the incremental updates in the EM algorithm).

5.1 Toy Data

We first illustrate active sensing with a toy example. Fig. 3 (left) shows a well separated two-class problem which was used in (Yu et al., 2008), with big squares and triangles representing the labeled positive and negative samples, and black dots denoting unlabeled points. To simulate our active sensing experiment, we randomly "hide" one of the two features of each sample with 40% probability each, and with 20% probability observe both features. The final incomplete training data are shown in Fig. 3 (middle) with the incomplete samples shown along the first or second axis. It can be seen that only 2 fully observed positive and negative samples are available. For active sensing MI we use the Gaussian kernel with width 0.5, and let the GMM choose the number of clusters automatically. Standard transductive setting is applied where all the unlabeled data are available for co-training kernel calculation. In Fig. 3 (right) we compare active sensing with random sensing, using the Area Under the ROC Curve (AUC) for the unlabeled data. The x-axis labels each acquired pair in order. This indicates that active sensing is much better than random sensing in improving the classification performance. The Bayes optimal accuracy (reachable when there is no missing data) is reached by the 16th query by active sensing whereas random sensing improves much slower with the number of acquired pairs. The two active sensing algorithms show similar results.

5.2 Survival Prediction for Lung Cancer

We consider 2-year survival prediction for advanced non-small cell lung cancer (NSCLC) patients treated with (chemo-)radiotherapy. This is currently a very challenging problem in clinical research, since the prognosis of this group of patients is very poor (less than 40% survive two years). Currently most models in the literature rely on various clinical factors of the patient such as gender and the WHO performance status. Very recently, imaging-related factors such as the size of the tumor and the number of positive lymph node stations are shown to be better predictors (Dehing-Oberije et al., 2009). However, it is expensive to obtain the images and to manually measure these factors. Therefore we study how to select the best set of patients to go through imaging to get additional features. All the relevant factors are listed in Fig. 4 (left) with short descriptions. These factors are all known to be predictive (Dehing-Oberije et al., 2009). From Bayesian co-training point of view we have 2 views, with 3 features in the first (clinical feature) view and 2 features in the second (imaging-based feature) view.

Our study contains 233 advanced NSCLC patients treated at the MAASTRO Clinic in the Netherlands from 2002 to 2006, among which 77 survived 2 years (labeled $+1$). All the features are available for these patients, and are normalized to have zero mean and

Features for NSCLC 2-years Survival Prediction

Feature	Description	View
GENDER	1-Male, 2-Female	1st
WHO	WHO performance status	1st
FEV1	Forced expiratory volume in 1 second	1st
GTV	Gross tumor volume	2nd
NPLN	Number of positive lymph node stations	2nd

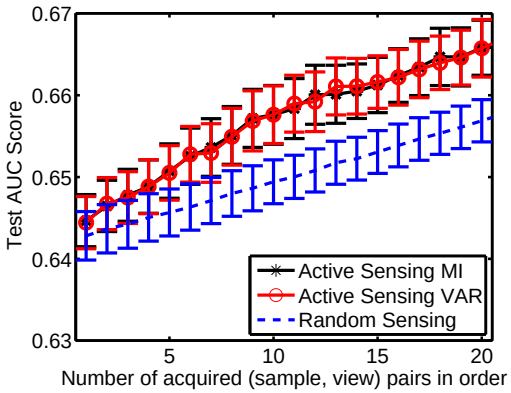

Figure 4: Experiments on NSCLC survival prediction. The features for the 2 views are listed in the left table, and the performance comparison of active sensing and random sensing is shown in the right figure. As baselines, training with full features (i.e., no sensing needed) yields 0.73; training with mean imputation (i.e., using the mean of each feature to fill in the missing entries) yields 0.62.

unit variance before training. We randomly choose 30% of the patients as training samples (with labels known), and the rest 70% as unlabeled samples. We use linear kernel for each view, and let the GMM algorithm automatically choose the number of clusters. As the active sensing setup, the first view is available for all the patients, and the second view is available only for randomly chosen 50% patients. So our goal is to sequentially select patients to acquire features in view 2, such that the overall classifier performance is maximized. Fig. 4 (right) shows the test AUC scores (with error-bars) of active sensing and random sensing, with different number of acquired pairs. Performance is averaged over 20 runs with randomly chosen 50% patients at the start. Active sensing in general yields better performance, and is significantly better after 5 first pairs. Active sensing based on MI and VAR again yield very similar results. We have also tested other experimental settings, and the comparison is not sensitive to this setup.

5.3 Pathological Complete Response (pCR) Prediction for Rectal Cancer

Our second example is to predict tumor response after chemo-radiotherapy for locally advanced rectal cancer. This is important in individualizing treatment strategies, since patients with a pathologic complete response (pCR) after therapy, i.e., with no evidence of viable tumor on pathologic analysis, would need less invasive surgery or another radiotherapy strategy instead of resection. Most available models combine clinical factors such as gender and age, and pre-treatment imaging-based factors such as tumor length and SUV_{max} (from CT/PET imaging), but it is expected that adding imaging data collected *after* therapy would lead to a better predictive model (though with a higher cost). In this study we show how to effectively select patients to go through pre-treatment and post-treatment imaging to better predict pCR.

We use the data from (Capirci et al., 2007) which contains 78 prospectively collected rectal cancer patients. All patients underwent a CT/PET scan before treatment and 42 days after treatment, and 21 of them had pCR (labeled +1). We split all the features into 3 views (clinical, pre-treatment imaging, post-treatment imaging), and the features are listed in Fig. 5 (left). For active sensing, we assume that all the (labeled or unlabeled) patients have view 1 features available, 70% of the patients have view 2 features available, and 40% of the patients have view 3 features available. This is to account for the fact that view 3 features are most expensive to get. All the other settings are the same as the NSCLC survival prediction study. Fig. 5 (right) shows the performance comparison of active sensing with random sensing, and it is seen that after about 18 pair acquisitions, active sensing is significantly better than random sensing. Active sensing MI and VAR share a similar trend, and the MI based active sensing is overall better than VAR based active sensing. The difference is however not statistically significant. The optimal AUC (when there are no missing features) is shown as a dotted line, and we see that with around 34 actively acquired pairs, active sensing can almost achieve the optimum. It takes however much longer for random sensing to reach this performance.

6 Conclusion and Future Work

This paper makes two primary contributions. First of all, for the purpose of active sensing we extend

Features for pCR Prediction in Rectal Cancer

Feature	Description	View
GENDER	1-Male, 2-Female	1st
AGE	Age in years	1st
STAGE	Staging of cancer	1st
LENGTH	Max diameter of the tumor	2nd
SUVPre	SUV_{max} before treatment	2nd
ΔSUV	Absolute difference of SUV_{max} before and after treatment	3rd
RI	Response Index, ΔSUV in %	3rd

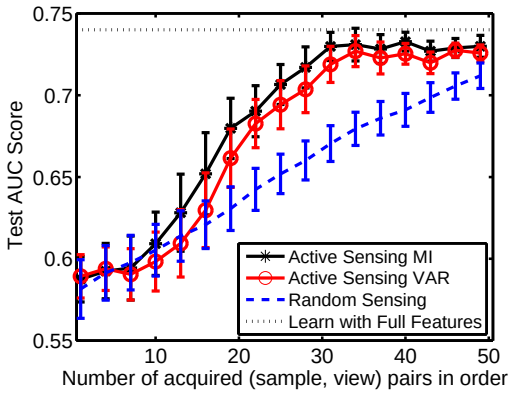

Figure 5: Experiments on pCR prediction for rectal cancer. The features for the 3 views are listed in the left table, and the performance comparison of active sensing and random sensing is shown in the right figure. As baselines, training with full features (i.e., no sensing needed) yields 0.74 (shown as a dotted line); training with mean imputation (i.e., using the mean of each feature to fill in the missing entries) yields 0.55 (not shown).

the Bayesian co-training framework to handle real-life data where objects are often incompletely characterized, i.e., only a subset of views are available for certain samples. Second, we introduce two approaches for active sensing, based on mutual information and predictive variance, respectively, which automatically decides which (sample, view) pair should be acquired further to get the most benefit. Note that one actively acquired pair would improve the overall multi-view classification performance for all the unlabeled samples. Experimental results on two real medical classification problems indicate that the proposed approach is indeed more accurate than randomly acquiring unobserved (sample, view) pairs.

As part of the future work, we will take into account the actual cost involved in the view acquisition for better decision making. This might be important, for instance, in medical diagnosis where Ultrasound and MRI induce quite different costs. Another step is to combine active sensing with active learning, such that one can query both an unobserved (sample, view) pair, and an unobserved label.

References

M. Bilgic and L. Getoor. VOILA: Efficient Feature-value Acquisition for Classification. In *AAAI*, 2007.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

C. Capirci, L. Rampin, P. Erba, F. Galeotti, G. Crepaldi, E. Banti, M. Gava, S. Fanti, G. Mariani, P. Muzzio, and D. Rubello. Sequential FDG-PET/CT reliably predicts response of locally advanced rectal cancer to neo-adjuvant chemo-radiation therapy. *Eur J Nucl Med Mol Imaging*, 34, 2007.

D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Interscience, 1991.

S. Dasgupta, M. Littman, and D. McAllester. PAC generalization bounds for co-training. In *NIPS*, 2001.

C. Dehing-Oberije, S. Yu, D. De Ruysscher, S. Meerschout, K. van Beek, Y. Lievens, J. van Meerbeeck, W. de Neve, G. Fung, B. Rao, S. Krishnan, H. van der Weide, and P. Lambin. Development and external validation of a prediction model for 2-year survival of non-small cell lung cancer patients treated with (chemo) radiotherapy. *To appear in Int J Radiat Oncol Biol Phys*, 2009.

V. Fedorov. *Theory of Optimal Experiments*. Academic Press, 1972.

P. Flaherty, M. Jordan, and A. Arkin. Robust design of biological experiments. In *NIPS*, 2006.

A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes:theory, efficient algorithms and empirical studies. *JMLR*, 9:235–284, 2008.

D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, pages 3–12, 1994.

D. Lindley. On a measure of the information provided by an experiment. *Ann. Math. Stat*, 27:986–1005, 1956.

D. MacKay. Information-based objective functions for active data selection. *Neural Comp.*, 4:590–604, 1992.

P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. Active feature-value acquisition for classifier induction. In *ICDM*, 2004.

N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, pages 444–448, 2001.

S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Fifth Workshop on Computational Learning Theory*, pages 287–94, 1992.

S. Yu, B. Krishnapuram, R. Rosales, H. Steck, and B. Rao. Bayesian co-training. In *NIPS*, 2008.