Instance Label Prediction by Dirichlet Process Multiple Instance Learning

Melih Kandemir
Heidelberg University HCI/IWR
Germany

Fred A. Hamprecht
Heidelberg University HCI/IWR
Germany

Abstract

We propose a generative Bayesian model that predicts instance labels from weak (bag-level) supervision. We solve this problem by simultaneously modeling class distributions by Gaussian mixture models and inferring the class labels of positive bag instances that satisfy the multiple instance constraints. We employ Dirichlet process priors on mixture weights to automate model selection, and efficiently infer model parameters and positive bag instances by a constrained variational Bayes procedure. Our method improves on the state-of-the-art of instance classification from weak supervision on 20 benchmark text categorization data sets and one histopathology cancer diagnosis data set.

1 INTRODUCTION

Automated data acquisition has reached unprecedented scales. However, annotation of ground-truth labels is still manual in many applications, lagging behind the massive increase in observed data. This fact makes learning from partially labeled data emerge as a key problem in machine learning. Multiple instance learning (MIL) tackles this problem by learning from labels available only for instance groups, called *bags* [7]. A negatively labeled bag indicates that all instances have negative labels. In a positively labeled bag, there is at least one positively labeled instance; however, which of the instances are positive is not specified. We refer to these bag labeling rules as *multiple instance constraints*. A positive bag instance with a positive label is called a *witness*, and one with a negative label a *non-witness*.

The classical MIL setup involves both bag-level training and bag-level prediction. The mainstream MIL algorithms are developed and evaluated under this classical setup. The harder problem of *instance*-level prediction from *bag*-level training has been addressed in a comparatively smaller volume of studies [16, 17, 32]. A group of existing models, such as Key Instance SVM (KI-SVM) [16] and CkNN-ROI [32] aim to identify a single positive instance from each positive bag, the so called *key instance*, that determines the bag label, and discard the other instances. In a recent work, Liu et al. [17] generalize this approach by a voting framework (VF) that learns an arbitrary number of key instances from each positive bag. While KI-SVM extends the MI-SVM formulation [2] with binary variables indicating key instances, CkNN-ROI and VF are built on the Citation k-NN method [26].

1.1 Contribution

Our central assumption is that all instances belonging to the same Gaussian / cluster share the same class label. By performing simultaneous assignment of instances to one class or the other and clustering instances within each class, our method effectively captures non-witnesses within the positive bags from their clustering relationships to other instances. Figure 1 illustrates this idea.

We discover the latent positive bag instance labels by non-parametrically modeling the distributions of both classes, while simultaneously assigning the positive bag instances to the most appropriate class. To capture almost arbitrarily complex data distributions, we model the class distributions as mixture of a potentially very large (determined by data and the Dirichlet process prior) number of Gaussians with full covariance. The Dirichlet process prior on the mixture weights addresses the model selection problem, which is in our context the question of how many clusters to use.

We infer the class distribution parameters and positive bag instance labels by an efficient constrained variational inference procedure. For a fixed configuration of positive bag instance labels, we update class distribution parameters as in variational inference of standard Dirichlet process mixtures of Gaussians. Then keeping class distribution parameters fixed, we assign each positive bag instance to the class that maximizes the total variational lower bound of

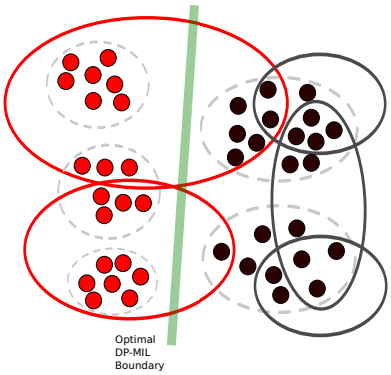


Figure 1: Dots, solid ellipses, and dashed ellipses indicate instances, bags, and clusters in a two dimensional feature space, respectively. Positive class is shown as red and negative class as black. DPMIL infers the label of a positive bag instance based on the class of the cluster that explains it best.

class distributions. This way, an increase in lower bound is guaranteed for all coordinate ascent updates, providing fast convergence.

We evaluate our method on 20 benchmark text categorization data sets, and on a novel application: finding Barrett's cancer tumors in histopathology tissue images from bag labels. Our method improves the state-of-the-art in both of these applications in terms of instance-level prediction performance. Furthermore, differently from many existing MIL methods, the inferred data modes and cluster weights of our method enable enhanced interpretability. The source code of our method is publicly available [1].

2 PRIOR ART

There exist several strategies for learning from weak supervision. One is *semi-supervised learning*, which suggests using large volumes of unlabeled data along with the limited labeled data to improve supervised learning performance [6]. *Active learning* is an alternative strategy that proposes learning from the smallest possible set of training samples selected by the model itself [24]. Another strategy is *self-taught learning* where abundant unlabeled data are available from a different but related task than the actual learning problem to be solved [20].

Multiple instance learning also aims to solve the weakly supervised learning problem by allowing supervision only for *groups of instances*. This learning setup has been first introduced by Dietterich et al. [7]. The authors propose detecting witnesses from the assumption that they lie in a single axis parallel rectangle (APR) in the feature space.

MIL methods are built upon different heuristics. A group of methods iteratively choose one instance from each bag as a representative, and infer model parameters from this selected instance set. Based on the new model parameters, a new representative set is selected in the next iteration. Seminal examples of this approach are EMDD [30] and MI-SVM [2]. While the former learns a Gaussian density kernel on the representative instances, the latter trains a support vector machine (SVM) on them.

Another group of MIL methods calculate similarities between bag pairs by bag-level kernels, and train standard kernel learners, such as SVM, based on these bag similarities. MI Kernel [10] and mi-Graph [31] are seminal examples of this approach. The common property of these models is that they assume non-i.i.d. relationships between instances belonging to the same bag. There have been recent attempts to exploit within-bag correlations in more elaborate ways, such as Ellipsoidal MIL [15] and MIMN [11]. The former method represents each bag as an ellipsoid and learns a max-margin classifier that obeys the multiple instance constraints. The latter models the within-bag relationships by a Markov Random Field whose unary potentials are determined by the output of a linear instance-level classifier and clique (bag) potentials are calculated from the unary potentials subject to the multiple instance constraints. These methods are typically both effective and efficient. However, they are not applicable to instance level prediction due to the central non-i.i.d bag instances assumption.

MIL as semi-supervised learning. MIL can be formulated as a semi-supervised learning problem by assigning latent variables to positive bag instances and inferring them subject to the multiple instance constraints [8]. mi-SVM [2] applies this principle to the SVM formulation. GPMIL [14] and Bayesian Multiple Instance RVM [21] apply it to the Gaussian process classifier and the relevance vector machine, respectively, by adapting the likelihood function to MIL.

Generative MIL models. The semi-supervised learning approach has also been adopted by some generative methods that model the class distributions and infer the label of each positive bag instance based on which of these two distributions explain that instance with higher likelihood [1,8]. Foulds et al. [8] model each class distribution by a Gaussian density with isotropic or diagonal covariance, and learn the latent positive bag instances without employing the multiple instance constraints on the training data. Adel et al. [1], on the other hand, provide a generic framework that enforces the multiple instance constraint in the hard assignment of instances to classes. They model class distributions by a Gaussian density and Gaussian copula. We fol-

[1] http://hci.iwr.uni-heidelberg.de/Staff/mkandemi/

low this line of research, and extend the existing work by i) using a richer family of distributions (potentially infinite mixtures of Gaussians with full covariance), while ii) keeping the multiple instance constraints and also providing an efficient variational inference procedure, and iii) making instance rather than bag level predictions.

Applications. Recent applications of MIL include diabetic retinopathy screening [19], visual saliency estimation [27] as well as content-based object detection and tracking [23]. MIL is also useful in drug activity prediction where each molecule constitutes a bag, each configuration of a molecule an instance, and binding of any of these configurations to the desired target is treated as a positive label, as first introduced by Dietterich et al. [7]. More recent applications of MIL to this problem include finding the interaction of proteins with Calmodulin molecules [18], and finding bioactive conformers [9]. Xu et al. [28, 29] apply MIL to tissue core (bag) level diagnosis of prostate cancer from histopathology images, where they combine multi-instance boosting [25] and clustering. There does not exist any prior work that focuses on locating tumors from tissue core level supervision, which we do in this paper as a case study.

Instance-level MIL prediction. There exist few studies focusing on instance prediction within the MIL setting. The first principled attempt towards this direction has been made by Zhou et al. [32]. The authors introduce a variant of Citation k-NN, called CkNN-ROI. This method chooses one instance from each positive bag as the *key instance* that determines the bag label based on how well it predicts the training bag labels by nearest neighbor matching, and ignores the other instances. Li et al. [16] detect key instances by a large margin method called KI-SVM. This method extends MI-SVM by binary latent variables assigned to each positive bag instance, which identify strictly one key instance per positive bag, and filter other instances out. The authors propose two variants of their method: i) Bag KI-SVM that has one slack variable per negative bag, and ii) Instance KI-SVM that has one slack variable per negative bag *instance*. Liu et al. [17] later propose detecting multiple key instances per positive bag by another variant of Citation kNN that learns a voting function from training bags. These models are shown to be effective in region-of-interest detection in natural scene images and text categorization. In this paper, we target the same learning problem, and empirically show that rich modeling of class distributions leads to better prediction performance.

3 THE MODEL

Let $\mathbf{X}$ be a data set consisting of B bags $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_B]$ indexed by b, and $\mathbf{y} = [y_1, \cdots, y_B]$ be the vector of the corresponding binary bag labels $y_b \in \{-1, +1\}$. Each bag $\mathbf{X}_b = [\mathbf{x}_{b1}, \cdots, \mathbf{x}_{bN_b}]$ consists of N_b instances. We assume that each instance is associated with a binary latent variable $r_{bn} \in \{-1, +1\}$ representing the label of the instance. We further assume that the positive instances in the data set ($r_{bn} = +1$) come from distribution $p(\mathbf{x}_{bn}|\boldsymbol{\theta}_{+1})$, and the negative instances ($r_{bn} = -1$) come from distribution $p(\mathbf{x}_{bn}|\boldsymbol{\theta}_{-1})$, parameterized by $\boldsymbol{\theta}_{+1}$ and $\boldsymbol{\theta}_{-1}$, respectively. Both of these two distributions are Gaussian mixture models with full covariance and with Dirichlet process priors on mixture weights. The generative process of our model is

$$
\begin{aligned}
p(\mathbf{v}_l) &= \prod_{k=1}^{K} Beta(v_{lk}|1, \alpha), && \forall l \\
p(z_{lbn}|\mathbf{v}_l) &= Mult(z_{lbn}|\pi_{l1}, \cdots, \pi_{lK}), && \forall l, b, n \\
p(\boldsymbol{\Lambda}_{lk}) &= \mathcal{W}(\boldsymbol{\Lambda}_{lk}|\mathbf{W}_0, \nu_0), && \forall l, k \\
p(\boldsymbol{\mu}_{lk}|\boldsymbol{\Lambda}_{lk}) &= \mathcal{N}(\boldsymbol{\mu}_{lk}|\mathbf{m}_0, (\beta_0\boldsymbol{\Lambda}_{lk})^{-1}), && \forall l, k, \\
p(\mathbf{x}_{bn}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, z_{lbn}, r_{bn}) &= \\
\prod_{l\in\{-1,+1\}} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_{bn}|\boldsymbol{\mu}_{lk}, \boldsymbol{\Lambda}_{lk}^{-1})^{\mathbf{1}(z_{lbn}=k)\cdot\mathbf{1}(r_{bn}=l)}, & && \forall b, n, \\
p(y_b = +1|\mathbf{r}) &= 1 - \prod_{n=1}^{N_b} \left(1 - \mathbf{1}(r_{bn} = +1)\right), && \forall b
\end{aligned}
$$

where the hyperparameters of the model are $\{\nu_0, \mathbf{W}_0, \mathbf{m}_0, \beta_0, \alpha\}$. The function $\mathbf{1}(\cdot)$ is the indicator function which returns 1 if its argument is true, and 0 otherwise. Mult$(\cdot|\cdots)$, Beta$(\cdot|\cdot,\cdot)$, $\mathcal{N}(\cdot|\cdot,\cdot)$ and $\mathcal{W}(\cdot|\cdot,\cdot)$ denote the multinomial mass function, and Beta, Gaussian and Wishart distribution densities, respectively. K is the number of clusters, and k is the related index; $l \in \{-1, +1\}$ indexes the two class densities; $\pi_{lk} = v_{lk}\prod_{j=1}^{k-1}(1 - v_{lj})$ is the stick breaking prior over cluster assignments z_{lbn}. The vector $\mathbf{Z}_l$ contains cluster-assignment weights z_{lbn}. The sets $\boldsymbol{\mu} = \{\boldsymbol{\mu}_{-11}, \cdots, \boldsymbol{\mu}_{-1K}, \boldsymbol{\mu}_{+11}, \cdots, \boldsymbol{\mu}_{+1K}\}$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_{-11}, \cdots, \boldsymbol{\Lambda}_{-1K}, \boldsymbol{\Lambda}_{+11}, \cdots, \boldsymbol{\Lambda}_{+1K}\}$ contain the mean and inverse covariance of all clusters in the model, respectively. The vector $\mathbf{r}$ has class-assignment variables for all instances in its entries, and $\mathbf{r}_{-r_{bn}}$ has the same for all instances except r_{bn}. The set $\mathbf{r}_b$ has the class-assignment variables of bag b. If $y_b = -1$ is observed, it is also observed that $r_{bn} = -1$ for all instances of bag b. If $y_b = +1$ is observed, r_{bn} for bag instances of b are latent, hence are inferred from data. We refer to this model as *Dirichlet process multiple instance learning (DPMIL)*. Figure 2 illustrates the model in plate notation.

3.1 Inference

Following the probabilistic paradigm, for inference of the model above, we aim to maximize the marginal likelihood $p(\mathbf{X}, \mathbf{y}|\mathbf{z})$ with respect to the class assignments $\mathbf{z}$ subject

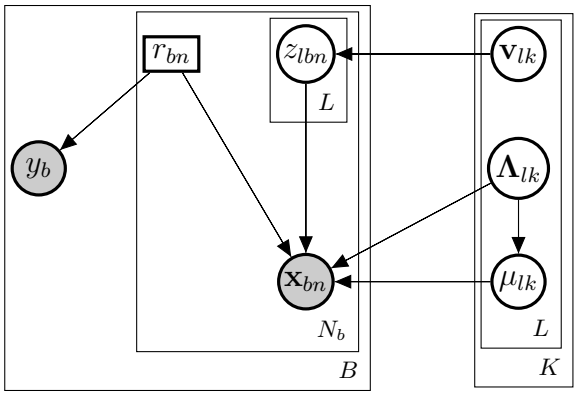


Figure 2: The generative process of DPMIL in plate notation. Shaded nodes denote observed, and unshaded notes denote latent variables that are inferred by constrained variational Bayes. Note that r_{bn} is a discrete binary latent variable without a prior. Hence it is denoted by a rectangle.

to the multiple instance constraints

$$\begin{aligned} &\underset{\mathbf{r}}{\text{maximize}} \quad p(\mathbf{X}, \mathbf{y}|\mathbf{r}) \\ &\text{s.t.} \qquad \max(\mathbf{r}_b) = y_b, \qquad \forall b. \end{aligned} \tag{1}$$

Let $\mathbf{r}_*$ be a solution to the optimization problem (1), we can define the divergence from the optimal configuration $\mathbf{r}_*$ as

$$D(\mathbf{r}) = \log \ \ p(\mathbf{X}, \mathbf{y}|\mathbf{r}_*) - \log \ \ p(\mathbf{X}, \mathbf{y}|\mathbf{r}).$$

It is easy to see that $D(\mathbf{r}) \geq 0$ for any $\mathbf{r}$ and $D(\mathbf{r}) = 0$ if $\mathbf{r} = \mathbf{r}_*$.

For a given configuration $\mathbf{r}$, calculating $p(\mathbf{X}, \mathbf{y}|\mathbf{r})$ is intractable. Hence, we approximate the posterior p a factorized distribution q

$$\begin{aligned} &p(\mathbf{Z}, \boldsymbol{\mu}, \mathbf{\Lambda}, \mathbf{v}_{-1}, \mathbf{v}_{+1}|\mathbf{X}, \mathbf{r}) \\ &= \left(\prod_{l \in \{-1,+1\}} \prod_{b=1}^{B} \prod_{n=1}^{N_b} q(z_{lbn}|\mathbf{r}) \right) \\ &\times \left(\prod_{l \in \{-1,+1\}} \prod_{k=1}^{K} q(\boldsymbol{\mu}_{lk}, \mathbf{\Lambda}_{lk}|\mathbf{r}) q(\mathbf{v}_{lk}|\mathbf{r}) \right). \end{aligned}$$

Let $\boldsymbol{\theta} = \boldsymbol{\theta}_{-1} \cup \boldsymbol{\theta}_{+1}$ denote the set of all parameters and latent variables of both class distributions. Following the standard variational Bayes formulation we can decompose $p(\mathbf{X}, \mathbf{y}|\mathbf{r})$ as

$$\log p(\mathbf{X}, \mathbf{y}|\mathbf{r}) = \mathcal{L}(\boldsymbol{\theta}|\mathbf{r}) + KL(q||p)$$

where

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{r}) = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}|\mathbf{r})] - \mathbb{E}_q[\log q(\boldsymbol{\theta}|\mathbf{r})]$$

is the variational lower bound and $KL(\cdot||\cdot)$ is the Kullback-Leibler divergence between the true posterior p and the approximate posterior q. Similarly to above, $KL(q||p) \geq 0$ for all q and $KL(q||p) = 0$ if and only if $q = p$. Combining these two facts, we have

$$\log p(\mathbf{X}, \mathbf{y}|\mathbf{r}_*) = \mathcal{L}(\boldsymbol{\theta}|\mathbf{r}) + \underbrace{KL(q||p) + D(\mathbf{r})}_{E(q,\mathbf{r})}$$

where the divergence term $E(q, \mathbf{r})$ approaches 0 as q and $\mathbf{r}$ approach optimal values. Hence, we can perform inference by

$$\begin{aligned} &\underset{\mathbf{r}, \boldsymbol{\theta}}{\text{maximize}} \quad \mathcal{L}(\boldsymbol{\theta}|\mathbf{r}) \\ &\text{s.t.} \qquad \max(\mathbf{r}_b) = y_b, \qquad \forall b. \end{aligned}$$

which has the same global optimum as the optimization problem (1). This problem can be solved by coordinate ascent. Keeping $\mathbf{r}$ fixed, model parameters $\boldsymbol{\theta}$ can be updated as in standard variational Bayes. Let $\boldsymbol{\psi}_j \subset \boldsymbol{\theta}$ be a subset of model parameters corresponding to a factor of q, the best possible update for this factor can be calculated by

$$\frac{\partial \mathcal{L}}{\partial q(\boldsymbol{\psi}_j)} = \mathbb{E}_{q(\boldsymbol{\theta}_{-\boldsymbol{\psi}_j})}[\log p(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}|\mathbf{r})] - \log q(\boldsymbol{\psi}_j) - 1 = 0.$$

Hence, the update rule becomes

$$q(\psi_j) = \exp\left\{ \mathbb{E}_{q(\boldsymbol{\theta}_{-\boldsymbol{\psi}_j})}[\log p(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}|\mathbf{r})] \right\}. \tag{2}$$

Consequently, keeping $\boldsymbol{\theta}$ fixed, $\mathbf{r}$ can be updated by

$$r_{bn}^{(t+1)} = \underset{l \in \{-1,+1\}}{\text{argmax}} \ \mathcal{L}(\boldsymbol{\theta}|\mathbf{r}_{-bn}^{(t)}, r_{bn} = l). \tag{3}$$

The cases that violate the multiple instance constraint $\max\left(\mathbf{r}_b^{(t+1)}\right) = y_b$ can be resolved by flipping one of the instances of bag b that had a positive label at iteration (t) back to positive. The fact that Equations (2) and (3) both increase $\mathcal{L}$ and that $E(q, \mathbf{r}) \geq 0$ bring out fast convergence to a local maximum in practice, as experimented in Section 4.3. The overall inference procedure is given in Algorithm 1, and the detailed update equations are available in Appendix 1.

3.2 Prediction

For a new bag $\mathbf{X}_b^* = [\mathbf{x}_{b1}^*, \cdots, \mathbf{x}_{bN_b}^*]$, instance-level prediction can be done by

$$\hat{r}_{bn} \leftarrow \underset{l \in \{-1,+1\}}{\text{argmax}} \ p(\mathbf{x}_{bn}^*|\mathbf{X}, \mathbf{y}, \mathbf{r}, y_{bn}^* = l),$$

where

$$p(\mathbf{x}_{bn}^*|\mathbf{X}, \mathbf{y}, \mathbf{r}, y_{bn}^* = l) = \int q(\boldsymbol{\theta}_l|\mathbf{X}, \mathbf{y}, \mathbf{r}) p(\mathbf{x}_{bn}^*|\boldsymbol{\theta}_l) d\boldsymbol{\theta}_l,$$

which corresponds to the standard predictive density for DP Gaussian mixtures as given in [4]. The extended formula of the predictive density for fixed $\mathbf{r}$ is given in Appendix 1.

Algorithm 1 Constrained variational inference for DPMIL

Input: Data $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_B]$,
Bag labels $\mathbf{y} = \{y_1, \cdots, y_b\}$
repeat
\\ Initialize instance class labels
$r_{bn} = y_b, \quad \forall b, n$
\\ Update the class distributions given the current $\mathbf{r}$
for $\psi_j \in \boldsymbol{\theta}$ **do**
$q(\psi_j|\mathbf{r}) \leftarrow \exp\left\{(\mathbb{E}_{q(\boldsymbol{\theta}_{-\psi_j})}[\log\ p(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}|\mathbf{r})]\right\}$
end for
\\ Update $\mathbf{r}$ given the class distributions
for $b \in \{j|y_j = +1\}$ **do**
for $n = 1$ **to** N_B **do**
$r_{bn}^{(t+1)} \leftarrow \underset{l \in \{-1,+1\}}{\operatorname{argmax}}\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{r}_{-r_{bn}}^{(t)}, r_{bn} = l)$
end for
\\ Resolve constraint violation
if $\max(\mathbf{r}_b) = -1$ **then**
$r_{bj}^{(t+1)} \leftarrow +1$, for any $j \in \{r_{bj}^{(t)} = +1\}$
end if
end for
until $convergence$

3.3 Relationship to existing models

DPMIL has the following connections to some of the existing methods:

- **mi-SVM** [2]: DPMIL and mi-SVM can be viewed as generative-discriminative pairs [12]. The two models find similar labels for positive bag instances when classes are separable. DPMIL additionally finds the clusters of both positive and negative instances.

- **EMDD** [30]: EMDD learns a class-conditional distribution $p(y_b = +1|\mathbf{X}_b)$ in a *discriminative* manner by applying a *single* Gaussian kernel on the most representative *subset* of training instances. DPMIL explains the *generative* process of *all* training instances by *multiple* Gaussian densities.

- **QDA**: Our method extends Quadratic Discriminant Analysis (QDA) in three aspects: i) DPMIL fits multiple Gaussians on each class distribution, while QDA fits only one. ii) DPMIL employs priors over mean and covariance, while QDA performs maximum likelihood estimation, following the frequentist paradigm. iii) DPMIL explains bag labels keeping the multiple instance constraints, while QDA performs single-instance learning.

- **MIMM** [8]: This model is a special case of DPMIL. In particular, when $K = 1$, uninformative priors are used for mixture coefficients $\mathbf{Z}$ and multiple instance constraints are ignored, DPMIL reduces to MIMM. Quadratic Discriminant Analysis (QDA) is the single-instance version of MIMM.

4 RESULTS

We evaluate the instance prediction performance of our method on two applications: i) web page categorization, and ii) Barrett's cancer diagnosis. For both experiments, we set cluster count K to 20 (per class), ν_0 to $D+1$, where D is the dimensionality of the data, $\mathbf{W}_0$ to the inverse empirical covariance of the data, $\mathbf{m}_0$ to the empirical mean of the data, β_0 to 1, and the concentration parameter α to 2, which is chosen as the smallest integer larger than the uninformative case ($\alpha = 1$). This value is not manually tuned. Other choices of α are observed not to affect the outcome significantly. We set maximum iteration count to 100.

We compare DPMIL to three MIL and two key instance detection algorithms: mi-SVM [2], MI-SVM [2], GPMIL [14], Bag KI-SVM [16], and Instance KI-SVM [16]. Models such as mi-Graph [31], iAPR [7], EMDD [30], Citation k-NN [26], MILBoost [25], and MIMM [8] are observed to perform worse than the list above, hence are not reported in detail. For all kernelizeable models, the radial basis function (RBF) kernel is used. Hyperparameters of the competing models are learned by cross-validation.

4.1 20 text categorization data sets

As a benchmarking study, we evaluate DPMIL on the public *20 Newsgroups* database that consists of 20 text categorization data sets. Each data set consists of 50 positive and 50 negative bags. Positive bags have on average 3 % of their instances from the target category, and the rest from other categories. Each instance in a bag is the top 200 TF-IDF representation of a post. We reduce the dimensionality to 100 by Kernel Principal Component Analysis (KPCA) with an RBF kernel with a length scale of $\sqrt{100}$, following the heuristic of Chang et al [5]. We evaluate the generalization performance using 10-fold cross validation with the standard data splits. We use Area Under Precision-Recall Curve (AUC-PR) as the performance measure due to its insensitivity to class imbalance. Table 1 lists the performance scores of models in comparison for the 20 data sets. We report average AUC-PR of two comparatively recent methods, VF and VFr, on the same database from [17] Table 5 [2], for which public source code is not available. Our method gives the highest instance prediction performance in 18 of the 20 data sets, and its average performance throughout the database is 3 percentage points higher than the state-of-the-art VF method.

Table 1: Area Under Precision-Recall Curve (AUC-PR) scores of methods on the *20 Newsgroups* database for instance prediction. DPMIL outperforms the other MIL models in 18 out of 20 data sets. B-KI-SVM and I-KI-SVM stand for Bag KI-SVM and Instance KI-SVM, respectively.

Data set	DPMIL	VF	VFr	B-KISVM	miSVM	I-KISVM	GPMIL	MISVM
alt.atheism	0.67	-	-	**0.68**	0.53	0.46	0.44	0.38
comp.graphics	**0.79**	-	-	0.47	0.65	0.62	0.49	0.07
comp.os.ms-windows.misc	**0.51**	-	-	0.38	0.42	0.14	0.36	0.03
comp.sys.ibm.pc.hardware	**0.67**	-	-	0.31	0.57	0.38	0.35	0.10
comp.sys.mac.hardware	**0.76**	-	-	0.39	0.56	0.64	0.54	0.27
comp.windows.x	**0.73**	-	-	0.37	0.56	0.35	0.36	0.04
misc.forsale	**0.45**	-	-	0.29	0.31	0.25	0.33	0.10
rec.autos	**0.76**	-	-	0.45	0.51	0.42	0.38	0.34
rec.motorcycles	**0.69**	-	-	0.52	0.09	0.61	0.46	0.27
rec.sport.baseball	**0.74**	-	-	0.52	0.18	0.41	0.38	0.22
rec.sport.hockey	**0.91**	-	-	0.66	0.27	0.64	0.43	0.75
sci.crypt	**0.68**	-	-	0.47	0.57	0.26	0.31	0.32
sci.electronics	**0.90**	-	-	0.42	0.83	0.65	0.71	0.34
sci.med	**0.73**	-	-	0.55	0.37	0.44	0.32	0.44
sci.space	**0.70**	-	-	0.51	0.46	0.33	0.32	0.20
soc.religion.christian	**0.72**	-	-	0.53	0.05	0.45	0.45	0.40
talk.politics.guns	**0.64**	-	-	0.43	0.57	0.32	0.38	0.01
talk.politics.mideast	**0.80**	-	-	0.60	0.77	0.49	0.46	0.60
talk.politics.misc	0.60	-	-	0.50	**0.61**	0.38	0.29	0.30
talk.religion.misc	**0.51**	-	-	0.32	0.08	0.34	0.32	0.04
Average	**0.70**	0.67	0.59	0.47	0.45	0.43	0.40	0.26

Table 2: Barrett's cancer diagnosis accuracy and F1 score of models in comparison. DPMIL outperforms the second best model by 6 percentage points in accuracy and 3 percentage points in F1 score. Instance level supervision performance is provided in the bottom row for reference.

Method	Accuracy (%)	F1 Score
DPMIL	**71.8**	**0.74**
GPMIL	65.8	0.54
I-KISVM	65.4	0.45
B-KISVM	64.7	0.48
mi-SVM	62.7	0.71
MISVM	46.9	0.64
SVM	83.5	0.82

4.2 Barrett's cancer diagnosis

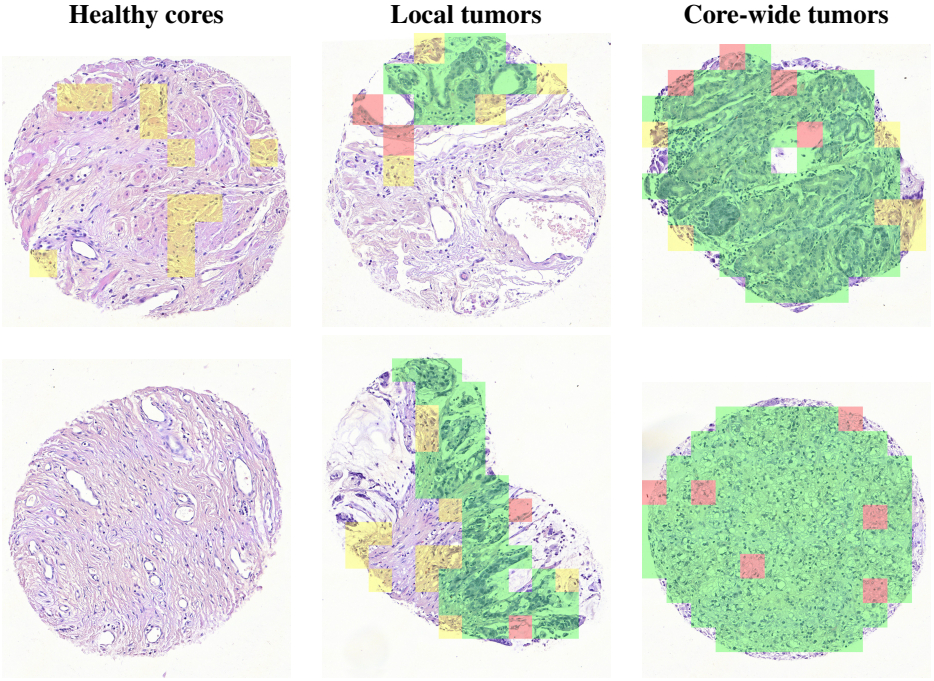
Biopsy imaging is a widely used cancer diagnosis technique in clinical pathology [22]. A sample is taken from the suspicious tissue, stained with hematoxylin & eosin, which dyes nuclei, stroma, lumen, and cytoplasm to different colours. Afterwards, the tissue is photographed under a microscope, and a pathologist examines the resultant image for diagnosis. In many cases, diagnosis of one patient requires careful scanning of several tissue slides of extensive sizes. Considerable time could be saved by an algorithm that finds the tumors and leads the pathologist to tumorous regions.

We evaluate DPMIL in the task of finding Barrett's cancer tumors in human esophagus tissue images from image-level supervision. Our data consists of 210 tissue core images (143 cancer and 67 healthy) taken from 97 Barrett's cancer patients. We treat tumor regions drawn by expert pathologists as ground truth. We split each tissue core (with average size of 2179x1970 pixels) into a grid of 200x200 pixel patches. We represent each patch by a 738-dimensional feature vector of SIFT descriptors, local binary patterns with 20×20-pixel cells, intensity histogram of 26 bins for each of the RGB channels, and the mean of the features described in [13] for cells lying in that patch. The data set includes 14303 instances, 53.4% of which are cancerous. We treat each image as a bag and each patch belonging to that image as an instance. A bag is labeled as positive if it includes tumor, and negative otherwise. Similarly to above, we reduce the data dimensionality to 30 by KPCA with an RBF kernel having a length scale of $\sqrt{30}$. We evaluate generalization performance by 4-fold cross-validation over bags. We repeated this procedure 5 times.

The patch-level diagnosis performance comparison of models is given in Table 2. Prediction performance of DPMIL lies in the middle of the chance level of 53.4% and the upper bound of 83.5% which is reached by patch-level

[2] Liu et al. [17] report 0.42 AUC-PR for KI-SVM and 0.41 AUC-PR for mi-SVM in Table 5.

Figure 3: Patch prediction results on sample tissue core images. **Green:** correctly detected cancer (true positive), **Red:** Missed detection of cancer (false negative), **Yellow:** False cancer alarm (false positive). **Rest:** True negative.



training of an SVM with RBF kernel. DPMIL clearly outperforms existing models both in prediction accuracy and F1 score (harmonic mean of precision and recall). Figure 3 shows prediction results of DPMIL on six sample tissue cores (bags) with different proportions of tumor. DPMIL produces few false positives for the healthy tissues (left-most column), detects local tumors with reasonable accuracy (middle columns), and produces few false negatives for tissue cores covered entirely by tumor (right-most column).

Figure 4 shows the mixture weights of the clusters for the class distributions averaged over data splits. The *healthy* class is dominated by a single cluster due to the relatively uniform structure of a healthy esophagus tissue. On the other hand, for the *cancer* class, the weights are more evenly distributed among five clusters. This result is consistent with the fact that the data set includes images from various grades of cancer. Each grade of cancer causes a different visual pattern in the tissue, resulting in a multi-modal distribution of tumor patches. As shown in Figure 5, clusters capture meaningful visual structures. Patches in the first row correspond to a stage of Barrett's cancer where cells form circular structures called *glands* which do not exist in a healthy esophagus tissue. The second row illustrates samples of cells with faded color, and in the third row the tissue is covered by an overly high population of poorly differentiated cells.

4.3 Learning rate and computational time

Weak supervision often emerges as a necessity for analyzing big data. Hence, computational efficiency of an MIL model is of key importance for feasibility for real-world scenarios. To this end, we provide an empirical analysis of the learning rate and the training time of DPMIL. As shown in Figure 6, the variational lower bound $\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{r})$ exhibits a sharp increase in the first few iterations, and saturates within 50 iterations.

Figure 6: Evolution of the variational lower bound $\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{r})$ throughout training iterations for the Barrett's cancer data set. DPMIL exhibits a steep learning curve and converges in less than 50 iterations.

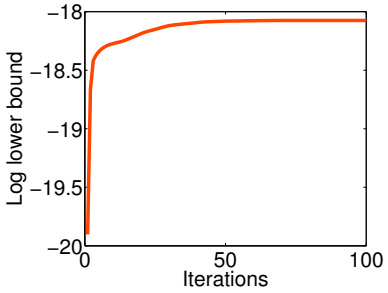


Table 3 shows the average training times of the models in comparison for one data split. Thanks to its Bayesian nonparametric nature, DPMIL does not require a cross-validation stage for model selection, unlike the other mod-

Figure 4: Cluster mixture coefficients for *cancer* ($y_b = +1$) and *healthy* ($y_b = -1$) in the Barrett's cancer data set. The healthy class distribution is dominated by a single mode unlike the cancer class distribution, supporting that a healthy tissue has a more even look than the cancer class which includes images belonging to various levels of cancer.

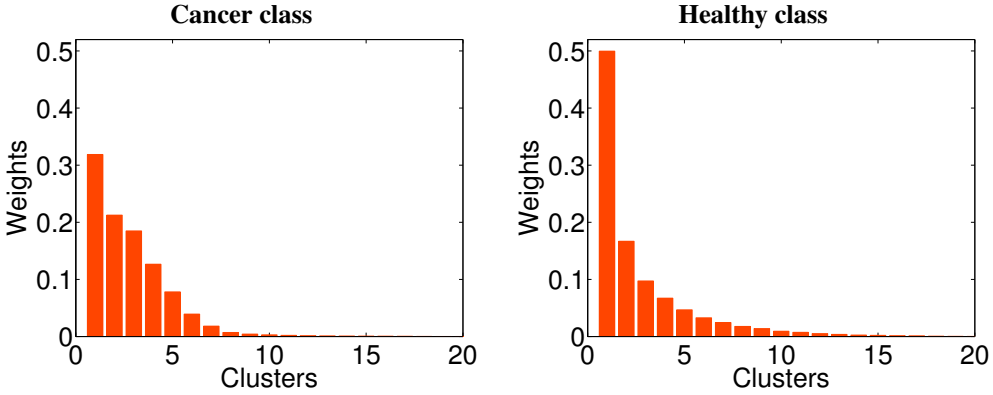
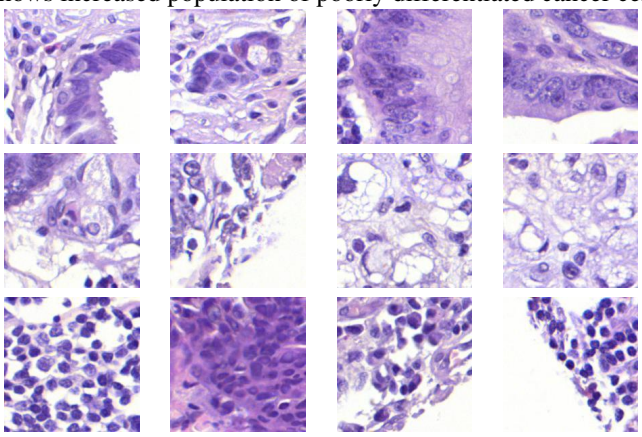


Figure 5: Sample patches from three different clusters (one in each row) of the *cancer* class. Each patch belongs to a different image. First cluster shows glandular formations of cancer cells, second cluster contains single cancer cells with faded color, and third cluster shows increased population of poorly differentiated cancer cells.



els. To avoid variability due to the desired level of detail in hyperparameter tuning (grid resolution and number of validation splits) which could lead to unfair comparison, we excluded the cross-validation time for the competing models. As a result of its steep learning rate, DPMIL provides reasonable training time, ranking as the most efficient model in text categorization and third in Barrett's cancer diagnosis.

Table 3: Training times (in seconds) of models in comparison for one data split. Thanks to the efficient variational inference procedure, DPMIL can be trained in reasonable time.

Model name	Text categorization	Barrett's cancer
DPMIL	**2.9**	44.7
KISVM-B	11.0	107.7
mi-SVM	12.2	126.6
KISVM-I	10.1	15.3
GPMIL	90.5	1491.7
MISVM	4.1	**10.8**

5 DISCUSSION

Multiple instance learning methods have long been developed and evaluated for bag label prediction. In this paper, we focus on the harder problem of instance level prediction from bag level training. We approach the problem from a semi-supervised learning perspective, and attempt to discover the unknown labels of positive bag instances by rich modeling of class distributions in a generative manner. We model these distributions by Gaussian mixture models with full covariance to handle complex multimodal cases. To avoid the model selection problem (i.e. predetermination of the number of data modes), we apply Dirichlet process priors over mixture coefficients.

As experimented in a large set of benchmark data sets and one cancer diagnosis application, our method clearly improves the state-of-the-art in instance classification from

weak labels. We attribute this improvement to the effectiveness of the *let the data speak* attitude in semi-supervised learning: The model discovers the unknown positive bag instance labels by assigning them to the class that explains the data generation process better (i.e. the class that increases the variational lower bound more). Of the other methods in our comparison, mi-SVM, VF, and KISVM are ignorant about the class distributions. The remaining methods are tailored for predicting bag, but not instance labels.

Generative modeling of data is commonly undesirable in standard pattern classification tasks, as a result of Vapnik's razor principle [3]. However, our results imply that generative data distribution modeling turns out to be an effective strategy when weak supervision is an additional source of uncertainty.

Modeling class distributions with mixture models brings enchanced interpretability as a by-product. Analysis of inferred clusters may provide additional information, or may support further modeling decisions. Even though we restrict our analysis to binary classification for illustrative purposes, extension of our method to multiclass cases is simply a matter of increasing the number of Gaussian mixture models from two to a desired number of classes.

Appendix 1: Variational update equations and predictive density

Variational update equations of the approximate posterior q correspond to those of the Gaussian mixture model as described in [3] where the Dirichlet prior on mixture weights are replaced by a Dirichlet process prior and instances are assigned to the appropriate distribution by indicator functions $\mathbf{1}(\cdot)$.

For $q(v_{lk}) = Beta(\gamma^1_{lk}, \gamma^2_{lk})$,

$$\gamma^1_{lk} = 1 + \sum_{b=1}^{B}\sum_{n=1}^{N_b} q(z_{lbn} = k)\mathbf{1}(r_{bn} = l),$$

$$\gamma^2_{lk} = \alpha + \sum_{b=1}^{B}\sum_{n=1}^{N_b} q(z_{lbn} > k)\mathbf{1}(r_{bn} = l).$$

For $q(z_{lbn} = k) = Mult(\tau^1_{lbn}, \cdots, \tau^K_{lbn})$,

$$\begin{aligned}\tau^k_{lbn} \leftarrow \Bigg(&\Psi(\gamma^1_{lk}) - \Psi(\gamma^1_{lk} + \gamma^2_{lk}) + \sum_{j=1}^{k=1}\left(\Psi(\gamma^2_{lk}) - \Psi\left(\gamma^1_{lk} + \gamma^2_{lk}\right)\right) \\ &+ \sum_{i=1}^{D}\Psi\left(\frac{\nu_{lk} + 1 - i}{2}\right) + D\log(2) + \log|\mathbf{W}_{lk}| - \frac{D}{2}\log(2\pi) \\ &- \frac{D}{2}\beta^{-1}_{lk} - \frac{1}{2}\nu_{lk}(\mathbf{x}_{bn} - \mathbf{m}_{lk})^T\mathbf{W}_{lk}(\mathbf{x}_{bn} - \mathbf{m}_{lk})\Bigg)\mathbf{1}(r_{bn} = l).\end{aligned}$$

[3] **Vapnik's razor principle:** When solving a (learning) problem of interest, do not solve a more complex problem as an intermediate step.

For $q(\boldsymbol{\mu}_{lk}, \boldsymbol{\Lambda}_{lk}) = \mathcal{N}(\boldsymbol{\mu}_{lk}|\mathbf{m}_{lk}, (\beta_{lk}\boldsymbol{\Lambda}^{-1}_{lk}))\mathcal{W}(\boldsymbol{\Lambda}_{lk}|\mathbf{W}_{lk}, \nu_{lk})$, where

$$\begin{aligned}\beta_{lk} &= \beta_0 + N_{lk}, \\ \mathbf{m}_{lk} &= \beta^{-1}_{lk}(\beta_0\mathbf{m}_0 + N_{lk}\bar{\mathbf{x}}_{lk}), \\ \mathbf{W}^{-1}_{lk} &= \mathbf{W}^{-1}_0 + N_{lk}\mathbf{S}_{lk} + \frac{\beta_0 N_{lk}}{\beta_0 + N_{lk}}(\bar{\mathbf{x}}_{lk} - \mathbf{m}_0)(\bar{\mathbf{x}}_{lk} - \mathbf{m}_0)^T, \\ \nu_{lk} &= \nu_0 + N_{lk} + 1.\end{aligned}$$

Here,

$$N_{lk} = \sum_{b}^{B}\sum_{n=1}^{N_b}\mathbf{1}(r_{bn} = l)q(z_{lbn} = k),$$

$$\bar{\mathbf{x}}_{lk} = \frac{1}{N_{lk}}\sum_{b}^{B}\sum_{n=1}^{N_b}\mathbf{1}(r_{bn} = l)q(z_{lbn} = k)\mathbf{x}_{bn},$$

$$\mathbf{S}_{lk} = \frac{1}{N_{lk}}\sum_{b=1}^{B}\sum_{n=1}^{N_b}\mathbf{1}(r_{bn} = l)q(z_{lbn} = k)(\mathbf{x}_{bn} - \bar{\mathbf{x}}_{lk})(\mathbf{x}_{bn} - \bar{\mathbf{x}}_{lk})^T.$$

For an inferred configuration $\hat{\mathbf{r}}$, the predictive density of DPMIL is identical to that of a standard Gaussian mixture model as given in [3]

$$\begin{aligned}p(\mathbf{x}^*_{bn}|\mathbf{X}, \mathbf{y}, \hat{\mathbf{r}}, y^*_{bn} = l) &= \int q(\boldsymbol{\theta}_l|\mathbf{X}, \mathbf{y}, \hat{\mathbf{r}})p(\mathbf{x}^*_{bn}|\boldsymbol{\theta}_l)d\boldsymbol{\theta}_l, \\ &= \frac{1}{\hat{\pi}_l}\sum_{k=1}^{K}\pi_{lk}St\left(\mathbf{x}^*_{bn}\middle|\mathbf{m}_k, \frac{(\nu_k + 1 - D)\beta_k}{1 + \beta_k}\mathbf{W}_k, \nu_k + 1 - D\right),\end{aligned}$$

where $\hat{\pi}_{lk} = \sum_{k=1}^{K}\pi_l$ and $St(\cdot|\cdot, \cdot, \cdot)$ is the Student's t density function.

References

[1] T. Adel, B. Smith, R. Urner, D. Stashuk, and D.J. Lizotte. Generative multiple-instance learning models for quantitative electromyography. In *UAI*, 2013.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2003.

[3] C.M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[4] D.M. Blei and M.I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] O. Chapelle, B. Schölkopf, A. Zien, et al. *Semi-supervised learning*. MIT Press, Cambridge, 2006.

[7] T.G. Dietterich, R.H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997.

[8] J.R. Foulds and P. Smyth. Multi-instance mixture models and semi-supervised learning. In *SIAM Int'l Conf. Data Mining*, 2011.

[9] G. Fu, X. Nan, H. Liu, R. Patel, P. Daga, Y. Chen, D. Wilkins, and R. Doerksen. Implementation of multiple-instance learning in drug activity prediction. *BMC Bioinformatics*, 13(Suppl 15):S3, 2012.

[10] T. Gärtner, P.A. Flach, A. Kowalczyk, and A.J. Smola. Multi-instance kernels. In *ICML*, 2002.

[11] H. Hajimirsadeghi, J. Li, G. Mori, M. Zaki, and T. Sayed. Multiple instance learning by discriminative training of markov networks. In *UAI*, 2013.

[12] A. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *NIPS*, 2002.

[13] M. Kandemir, A. Feuchtinger, A. Walch, and F.A. Hamprecht. Digital Pathology: Multiple instance learning can detect Barrett's cancer. In *ISBI*, 2014.

[14] M. Kim and F. De La Torre. Gaussian process multiple instance learning. In *ICML*, 2010.

[15] G. Krummenacher, C.S. Ong, and J. Buhmann. Ellipsoidal multiple instance learning. In *ICML*, 2013.

[16] Y.-F. Li, J.T. Kwok, I.W. Tsang, and Z-H. Zhou. A convex method for locating regions of interest with multi-instance learning. In *Machine learning and knowledge discovery in databases*, pages 15–30. Springer, 2009.

[17] G. Liu, J. Wu, and Z.-H. Zhou. Key instance detection in multi-instance learning. *Journal of Machine Learning Research-Proceedings Track*, 25:253–268, 2012.

[18] F.A. Minhas and A. Ben-Hur. Multiple instance learning of Calmodulin binding sites. *Bioinformatics*, 28(18):i416–i422, 2012.

[19] G. Quellec, M. Lamard, M.D. Abràmoff, E. Decencière, B. Lay, A. Erginay, B. Cochener, and G. Cazuguel. A multiple-instance learning framework for diabetic retinopathy screening. *Medical Image Analysis*, 2012.

[20] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML*, 2007.

[21] V.C. Raykar, B. Krishnapuram, J. Bi, M. Dundar, and R.B. Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *ICML*, 2008.

[22] R. Rubin and D.S. Strayer. *Rubin's pathology: clinicopathologic foundations of medicine*. Lippincott Williams & Wilkins, 2008.

[23] P. Sharma, C. Huang, and R. Nevatia. Unsupervised incremental learning for improved object detection in a video. In *CVPR*, 2012.

[24] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Journal of Machine Learning Research*, 2000.

[25] P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. *NIPS*, 2006.

[26] J. Wang and J.D. Zucker. Solving multiple-instance problem: A lazy learning approach. *ICML*, 2000.

[27] Q. Wang, Y. Yuan, P. Yan, and X. Li. Saliency detection by multiple-instance learning. *IEEE Trans. on Systems, Man and Cybernetics B*, 43(2):660 – 672, 2013.

[28] Y. Xu, J. Zhang, E.-C. Chang, M. Lai, and Z. Tu. Context-constrained multiple instance learning for histopathology image segmentation. *Lecture Notes in Computer Science*, 7512:623–630, 2012.

[29] Y. Xu, J.Y. Zhu, E. Chang, and Z. Tu. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *CVPR*, 2012.

[30] Q. Zhang, S.A. Goldman, et al. EM-DD: An improved multiple-instance learning technique. *NIPS*, 14, 2001.

[31] Z.H. Zhou, Y.Y. Sun, and Y.F. Li. Multi-instance learning by treating instances as non-iid samples. In *ICML*, 2009.

[32] Z.H. Zhou, X.B. Xue, and Y. Jiang. Locating regions of interest in cbir with multi-instance learning techniques. In *AI 2005: Advances in Artificial Intelligence*, pages 92–101. Springer, 2005.