Journal of Machine Learning Research 11 (2010) 1297-1322 Submitted 9/09; Revised 2/10; Published 4/10

Learning From Crowds

Vikas C. Raykar VIKAS.RAYKAR@SIEMENS.COM
Shipeng Yu SHIPENG.YU@SIEMENS.COM
CAD and Knowledge Solutions (IKM CKS)
Siemens Healthcare
Malvern, PA 19355 USA

Linda H. Zhao LZHAO@WHARTON.UPENN.EDU
Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104 USA

Gerardo Hermosillo Valadez GERARDO.HERMOSILLOVALADEZ@SIEMENS.COM
Charles Florin CHARLES.FLORIN@SIEMENS.COM
Luca Bogoni LUCA.BOGONI@SIEMENS.COM
CAD and Knowledge Solutions (IKM CKS)
Siemens Healthcare
Malvern, PA 19355 USA

Linda Moy LINDA.MOY@NYUMC.ORG
Department of Radiology
New York University School of Medicine
New York, NY 10016 USA

Editor: David Blei

Abstract

For many supervised learning tasks it may be infeasible (or very expensive) to obtain objective and reliable labels. Instead, we can collect subjective (possibly noisy) labels from multiple experts or annotators. In practice, there is a substantial amount of disagreement among the annotators, and hence it is of great practical interest to address conventional supervised learning problems in this scenario. In this paper we describe a probabilistic approach for supervised learning when we have multiple annotators providing (possibly noisy) labels but no absolute gold standard. The proposed algorithm evaluates the different experts and also gives an estimate of the actual hidden labels. Experimental results indicate that the proposed method is superior to the commonly used majority voting baseline.

Keywords: multiple annotators, multiple experts, multiple teachers, crowdsourcing

1. Supervised Learning From Multiple Annotators/Experts

A typical supervised learning scenario consists of a training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ containing N instances, where $\boldsymbol{x}_i \in \mathcal{X}$ is an instance (typically a d-dimensional feature vector) and $y_i \in \mathcal{Y}$ is the corresponding known label. The task is to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ which generalizes well on unseen data. Specifically for binary classification the supervision is from the set $\mathcal{Y} = \{0,1\}$, for multi-class classification $\mathcal{Y} = \{1,\ldots,K\}$, for ordinal regression $\mathcal{Y} = \{1,\ldots,K\}$ (with an ordering $1 < \ldots < K$), and $\mathcal{Y} = \mathbb{R}$ for regression.

However, for many real life tasks, it may not be possible, or may be too expensive (or tedious) to acquire the actual label y_i for training—which we refer to as the *gold standard* or the *objective ground truth*. Instead, we may have multiple (possibly noisy) labels $y_i^1, \ldots, y_i^R$ provided by R different experts or annotators. In practice, there is a substantial amount of disagreement among the experts, and hence it is of great practical interest to address conventional supervised learning algorithms in this scenario.

Our motivation for this work comes from the area of computer-aided diagnosis[1] (CAD), where the task is to build a classifier to predict whether a suspicious region on a medical image (like a X-ray, CT scan, or MRI) is malignant (cancerous) or benign. In order to train such a classifier, a set of images is collected from hospitals. The actual gold standard (whether it is cancer or not) can only be obtained from a biopsy of the tissue. Since it is an expensive, invasive, and potentially dangerous process, often CAD systems are built from labels assigned by *multiple radiologists* who identify the locations of malignant lesions. Each radiologist visually examines the medical images and provides a *subjective* (possibly noisy) version of the gold standard.[2] The radiologist also annotates various descriptors of the potentially malignant lesion, like the size (a regression problem), shape (a multi-class classification problem), and also degree of malignancy (an ordinal regression problem). The radiologists come from a diverse pool including luminaries, experts, residents, and novices. Very often there is lot of disagreement among the annotations.

For a lot of tasks the labels provided by the annotators are inherently *subjective* and there will be substantial variation among different annotators. The domain of text classification offers such a scenario. In this context the task is to predict the category for a token of text. The labels for training are assigned by human annotators who read the text and attribute their subjective category. With the advent of crowdsourcing (Howe, 2008) services like Amazon's Mechanical Turk,[3] Games with a Purpose,[4] and reCAPTCHA[5] it is quite inexpensive to acquire labels from a large number of annotators (possibly thousands) in a short time (Sheng et al., 2008; Snow et al., 2008; Sorokin and Forsyth, 2008). Websites such as Galaxy Zoo[6] allow the public to label astronomical images over the internet. In situations like these, the performance of different annotators can vary widely (some may even be malicious), and without the actual gold standard, it may not be possible to evaluate the annotators.

In this work, we provide principled probabilistic solutions to the following questions:

1. How to adapt conventional supervised learning algorithms when we have multiple annotators providing subjective labels but no objective gold standard?

2. How to evaluate systems when we do not have absolute gold-standard?

3. A closely related problem—particularly relevant when there are a large number of annotators—is to estimate how reliable/trustworthy is each annotator.

1. See Fung et al. (2009) for an overview of the data mining issues in this area.
2. Sometimes even a biopsy cannot confirm whether it is cancer or not and hence all we can hope to get is subjective ground truth.
3. Mechanical Turk found at `https://www.mturk.com`.
4. Games with a Purpose found at `http://www.gwap.com`.
5. reCAPTCHA found at `http://recaptcha.net/`.
6. Galaxy Zoo found at `http://galaxyzoo.org`.

1.1 The Problem With Majority Voting

When we have multiple labels a commonly used strategy is to use the labels on which the majority of them agree (or average for regression problem) as an estimate of the actual gold standard. For binary classification problems this amounts to using the majority label,[7] that is,

$$\hat{y}_i = \begin{cases} 1 & \text{if } (1/R)\sum_{j=1}^{R} y_i^j > 0.5 \\ 0 & \text{if } (1/R)\sum_{j=1}^{R} y_i^j < 0.5 \end{cases},$$

as an *estimate of the hidden true label* and use this estimate to learn and evaluate classifiers/annotators. Another strategy is that of considering every pair (instance, label) provided by each expert as a separate example. Note that this amounts to using a soft probabilistic estimate of the actual ground truth to learn the classifier, that is,

$$\Pr[y_i = 1 | y_i^1, \ldots, y_i^R] = (1/R)\sum_{j=1}^{R} y_i^j.$$

Majority voting assumes all experts are equally good. However, for example, if there is only one true expert and the majority are novices, and if novices give the same incorrect label to a specific instance, then the majority voting method would favor the novices since they are in a majority. One could address this problem by introducing a weight capturing how good each expert is. But how would one measure the performance of an expert when there is no gold standard available?

1.2 Proposed Approach and Organization

To address the apparent chicken-and-egg problem, we present a maximum-likelihood estimator that *jointly* learns the classifier/regressor, the annotator accuracy, and the actual true label. For ease of exposition we start with binary classification problem in § 2. The performance of each annotator is measured in terms of the sensitivity and specificity with respect to the unknown gold standard (§ 2.1). The proposed algorithm automatically discovers the best experts and assigns a higher weight to them. In order to incorporate prior knowledge about each annotator, we impose a beta prior on the sensitivity and specificity and derive the maximum-a-posteriori estimate (§ 2.6). The final estimation is performed by an Expectation Maximization (EM) algorithm that iteratively establishes a particular gold standard, measures the performance of the experts given that gold standard, and refines the gold standard based on the performance measures. While the proposed approach is described using logistic regression as the base classifier (§ 2.2), it is quite general, and can be used with any black-box classifier (§ 2.7), and can also handle missing labels (that is, each expert is not required to label all the instances). Furthermore, we extend the proposed algorithm to handle categorical (§ 3), ordinal (§ 4), and regression problems (§ 5). In § 6 section we extensively validate our approach using both simulated data and real data from different domains.

1.3 Related Work and Novel Contributions

We first summarize the novel contributions of this work in context of other related work in this emerging new area. There has been a long line of work in the biostatistics and epidemiology literature on latent variable models where the task is to get an estimate of the observer error rates based

7. When there is no clear majority among the multiple experts (that is, $\hat{y}_i = 0.5$) in CAD domain the final decision is often made by an adjudicator or a super-expert. When there is no adjudicator a fair coin toss is used.

on the results from multiple diagnostic tests without a gold standard (see Dawid and Skene, 1979, Hui and Walter, 1980, Hui and Zhou, 1998, Albert and Dodd, 2004 and references therein). In the machine learning community Smyth et al. (1995) first addressed the same problem in the context of labeling volcanoes in satellite images of Venus. We differ from this previous body of work in the following aspects:

1. Unlike Dawid and Skene (1979) and Smyth et al. (1995) which just focused on estimating the ground truth from multiple noisy labels, we specifically address the issue of *learning a classifier*. Estimating the ground truth and the annotator/classifer performance is a byproduct of our proposed algorithm.

2. In order to learn a classifier Smyth (1995) proposed to first estimate the ground truth (without using the features) and then use the probabilistic ground truth to learn a classifier. In contrast, our proposed algorithm *learns the classifier and the ground truth jointly*. Our experiments (§ 6.1.1) show that the classifier learnt and ground truth obtained by the proposed algorithm is superior to that obtained by other procedures which first estimates the ground truth and then learns the classifier.

3. Our solution is more general and can be easily extended to categorical(§ 3), ordinal(§ 4), and continuous data(§ 5). It can also be used in conjunction with any supervised learning algorithm. A preliminary version of this paper (Raykar et al., 2009) mainly discussed the binary classification problem.

4. Our proposed algorithm is also Bayesian—we impose a prior on the experts. The priors can potential capture the skill of different annotators. In this paper we refrain from doing a full Bayesian inference and use the mode of the posterior as a point estimate. A recent complete Bayesian generalization of these kind of models has been developed by Carpenter (2008).

5. The EM approach used in this paper is similar to that proposed by Jin and Ghahramani (2003). However their motivation is somewhat different. In their setting, each training example is annotated with a set of possible labels, only one of which is correct.

There has been recent interest in the natural language processing (Sheng et al., 2008; Snow et al., 2008) and computer vision (Sorokin and Forsyth, 2008) communities where they use Amazon's Mechanical Turk to collect annotations from many people. They show that it can be potentially as good as that provided by an expert. Sheng et al. (2008) analyzed when it is worthwhile to acquire new labels for some of the training examples. There is also some theoretical work (see Lugosi, 1992 and Dekel and Shamir, 2009a) dealing with multiple experts. Recently Dekel and Shamir (2009b) presented an algorithm which does not resort to repeated labeling, that is, each example does not have to be labeled by multiple teachers. Donmez et al. (2009) address the issue of active learning in this scenario—How to jointly learn the accuracy of labeling sources and obtain the most informative labels for the active learning task? There has also been some work in the medical imaging community (Warfield et al., 2004; Cholleti et al., 2008).

2. Binary Classification

We first describe our proposed noise model for the annotators. The performance of each annotator is measured in terms of the sensitivity and specificity with respect to the unknown gold standard.

2.1 A Two-coin Model for Annotators

Let $y^j \in \{0,1\}$ be the label assigned to the instance $\boldsymbol{x}$ by the j^{th} annotator/expert. Let y be the actual (unobserved) label for this instance. Each annotator provides a version of this hidden true label based on two biased coins. If the true label is one, she flips a coin with bias α^j (*sensitivity*). If the true label is zero, she flips a coin with bias β^j (*specificity*). In each case, if she gets heads she keeps the original label, otherwise she flips the label.

If the true label is one, the sensitivity (true positive rate) for the j^{th} annotator is defined as the probability that she labels it as one.

$$\alpha^j := \Pr[y^j = 1 | y = 1]. \tag{1}$$

On the other hand, if the true label is zero, the specificity (1−false positive rate) is defined as the probability that she labels it as zero.

$$\beta^j := \Pr[y^j = 0 | y = 0]. \tag{2}$$

The assumption introduced is that α^j and β^j do not depend on the instance $\boldsymbol{x}$. For example, in the CAD domain, this means that the radiologist's performance is consistent across different sub-groups of data.[8]

2.2 Classification Model

While the proposed method can be used for any classifier, for ease of exposition, we consider the family of linear discriminating functions: $\mathcal{F} = \{f_{\boldsymbol{w}}\}$, where for any $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^d$, $f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$. The final classifier can be written in the following form: $\hat{y} = 1$ if $\boldsymbol{w}^\top \boldsymbol{x} \geq \gamma$ and 0 otherwise. The threshold γ determines the operating point of the classifier. The Receiver Operating Characteristic (ROC) curve is obtained as γ is swept from $-\infty$ to ∞. The probability for the positive class is modeled as a *logistic sigmoid* acting on $f_{\boldsymbol{w}}$, that is,

$$\Pr[y = 1 | \boldsymbol{x}, \boldsymbol{w}] = \sigma(\boldsymbol{w}^\top \boldsymbol{x}),$$

where the logistic sigmoid function is defined as $\sigma(z) = 1/(1+e^{-z})$. This classification model is known as *logistic regression*.

2.3 Estimation/Learning Problem

Given the training data $\mathcal{D}$ consisting of N instances with annotations from R annotators, that is, $\mathcal{D} = \{\boldsymbol{x}_i, y_i^1, \ldots, y_i^R\}_{i=1}^N$, the task is to estimate the weight vector $\boldsymbol{w}$ and also the sensitivity $\boldsymbol{\alpha} = [\alpha^1, \ldots, \alpha^R]$ and the specificity $\boldsymbol{\beta} = [\beta^1, \ldots, \beta^R]$ of the R annotators. It is also of interest to get an estimate of the unknown gold standard $y_1, \ldots, y_N$.

2.4 Maximum Likelihood Estimator

Assuming the training instances are independently sampled, the likelihood function of the parameters $\theta = \{\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ given the observations $\mathcal{D}$ can be factored as

$$\Pr[\mathcal{D} | \theta] = \prod_{i=1}^{N} \Pr[y_i^1, \ldots, y_i^R | \boldsymbol{x}_i, \theta].$$

8. While this is a reasonable assumption, it is not entirely true. It is known that some radiologists are good at detecting certain kinds of malignant lesions based on their training and experience.

Conditioning on the true label y_i, and also using the assumption y_i^j is conditionally independent (of everything else) given α^j, β^j and y_i, the likelihood can be decomposed as

$$\begin{aligned}\Pr[\mathcal{D}|\theta] &= \prod_{i=1}^{N}\big\{\Pr[y_i^1,\ldots,y_i^R|y_i=1,\boldsymbol{\alpha}]\Pr[y_i=1|\boldsymbol{x_i},\boldsymbol{w}] \\ &+ \Pr[y_i^1,\ldots,y_i^R|y_i=0,\boldsymbol{\beta}]\Pr[y_i=0|\boldsymbol{x_i},\boldsymbol{w}]\big\}.\end{aligned}$$

Given the true label y_i, we assume that $y_i^1,\ldots,y_i^R$ are independent, that is, the annotators make their decisions independently.[9] Hence,

$$\Pr[y_i^1,\ldots,y_i^R|y_i=1,\boldsymbol{\alpha}] = \prod_{j=1}^{R}\Pr[y_i^j|y_i=1,\alpha^j] = \prod_{j=1}^{R}[\alpha^j]^{y_i^j}[1-\alpha^j]^{1-y_i^j}.$$

Similarly, we have

$$\Pr[y_i^1,\ldots,y_i^R|y_i=0,\boldsymbol{\beta}] = \prod_{j=1}^{R}[\beta^j]^{1-y_i^j}[1-\beta^j]^{y_i^j}.$$

Hence the likelihood can be written as

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N}\big[a_i p_i + b_i(1-p_i)\big],$$

where we have defined

$$\begin{aligned}p_i &:= \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i). \\ a_i &:= \prod_{j=1}^{R}[\alpha^j]^{y_i^j}[1-\alpha^j]^{1-y_i^j}. \\ b_i &:= \prod_{j=1}^{R}[\beta^j]^{1-y_i^j}[1-\beta^j]^{y_i^j}.\end{aligned}$$

The maximum-likelihood estimator is found by maximizing the log-likelihood, that is,

$$\hat{\theta}_{\mathrm{ML}} = \{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{w}}\} = \arg\max_{\theta}\{\ln\Pr[\mathcal{D}|\theta]\}.$$

2.5 The EM Algorithm

This maximization problem can be simplified a lot if we use the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm is an efficient iterative procedure to compute the maximum-likelihood solution in presence of missing/hidden data. We will use the unknown hidden true label y_i as the missing data. If we know the missing data $\boldsymbol{y} = [y_1,\ldots,y_N]$ then the complete likelihood can be written as

$$\ln\Pr[\mathcal{D},\boldsymbol{y}|\theta] = \sum_{i=1}^{N} y_i \ln p_i a_i + (1-y_i)\ln(1-p_i)b_i.$$

9. This assumption is not true in general and there is some correlations among the labels assigned by multiple annotators. For example in the CAD domain if the cancer is in advanced stage (which is very easy to detect) almost all the radiologists assign the same label.

Each iteration of the EM algorithm consists of two steps: an Expectation(E)-step and a Maximization(M)-step. The M-step involves maximization of a lower bound on the log-likelihood that is refined in each iteration by the E-step.

1. **E-step.** Given the observation $\mathcal{D}$ and the current estimate of the model parameters θ, the conditional expectation (which is a lower bound on the true likelihood) is computed as

$$\mathbb{E}\{\ln \Pr[\mathcal{D}, \boldsymbol{y}|\theta]\} = \sum_{i=1}^{N} \mu_i \ln p_i a_i + (1-\mu_i)\ln(1-p_i)b_i, \tag{3}$$

   where the expectation is with respect to $\Pr[\boldsymbol{y}|\mathcal{D},\theta]$, and $\mu_i = \Pr[y_i = 1|y_i^1,\ldots,y_i^R,\boldsymbol{x}_i,\theta]$. Using Bayes' theorem we can compute

$$\begin{aligned} \mu_i &\propto \Pr[y_i^1,\ldots,y_i^R|y_i=1,\theta]\cdot\Pr[y_i=1|\boldsymbol{x_i},\theta] \\ &= \frac{a_i p_i}{a_i p_i + b_i(1-p_i)}. \end{aligned}$$

2. **M-step.** Based on the current estimate μ_i and the observations $\mathcal{D}$, the model parameters θ are then estimated by maximizing the conditional expectation. By equating the gradient of (3) to zero we obtain the following estimates for the sensitivity and specificity:

$$\alpha^j = \frac{\sum_{i=1}^N \mu_i y_i^j}{\sum_{i=1}^N \mu_i}, \qquad \beta^j = \frac{\sum_{i=1}^N (1-\mu_i)(1-y_i^j)}{\sum_{i=1}^N (1-\mu_i)}.$$

   Due to the non-linearity of the sigmoid, we do not have a closed form solution for $\boldsymbol{w}$ and we have to use gradient ascent based optimization methods. We use the Newton-Raphson update given by $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \boldsymbol{H}^{-1}\mathbf{g}$, where $\boldsymbol{g}$ is the gradient vector, $\boldsymbol{H}$ is the Hessian matrix, and η is the step length. The gradient vector is given by

$$\boldsymbol{g}(w) = \sum_{i=1}^{N} \left[\mu_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\right] \boldsymbol{x}_i.$$

   The Hessian matrix is given by

$$\boldsymbol{H}(w) = -\sum_{i=1}^{N} \left[\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\right]\left[1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\right] \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

   Essentially, we are estimating *a logistic regression model with probabilistic labels* μ_i.

These two steps (the E- and the M-step) can be iterated till convergence. The log-likelihood increases monotonically after every iteration, which in practice implies convergence to a local maximum. The EM algorithm is only guaranteed to converge to a local maximum. In practice multiple restarts with different initializations can potentially mitigate the local maximum problem. In this paper we use majority voting $\mu_i = 1/R\sum_{j=1}^R y_i^j$ as the initialization for μ_i to start the EM-algorithm.

2.6 A Bayesian Approach

In some applications we may want to trust a particular expert more than the others. One way to achieve this is by imposing priors on the sensitivity and specificity of the experts. Since α_j and β_j represent the probability of a binary event, a natural choice of prior is the beta prior. The beta prior is also conjugate to the binomial distribution. For any $a > 0$, $b > 0$, and $\delta \in [0,1]$ the beta distribution is given by

$$\text{Beta}(\delta|a,b) = \frac{\delta^{a-1}(1-\delta)^{b-1}}{\text{B}(a,b)},$$

where $\text{B}(a,b) = \int_0^1 \delta^{a-1}(1-\delta)^{b-1} d\delta$ is the beta function.We assume a beta prior[10] for both the sensitivity and the specificity as

$$\begin{aligned} \Pr[\alpha_j|a_1^j,a_2^j] &= \text{Beta}(\alpha_j|a_1^j,a_2^j). \\ \Pr[\beta_j|b_1^j,b_2^j] &= \text{Beta}(\beta_j|b_1^j,b_2^j). \end{aligned}$$

For sake of completeness we also assume a zero mean Gaussian prior on the weights $\boldsymbol{w}$ with inverse covariance matrix $\boldsymbol{\Gamma}$, that is, $\Pr[\boldsymbol{w}] = \mathcal{N}(\boldsymbol{w}|0,\boldsymbol{\Gamma}^{-1})$. Assuming that $\{\alpha_j\}$, $\{\beta_j\}$, and $\boldsymbol{w}$ have independent priors, the maximum-a-posteriori (MAP) estimator is found by maximizing the log-posterior, that is,

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \{\ln \Pr[\mathcal{D}|\theta] + \ln \Pr[\theta]\}.$$

An EM algorithm can be derived in a similar fashion for MAP estimation by relying on the interpretation of Neal and Hinton (1998). The final algorithm is summarized below:

1. Initialize $\mu_i = (1/R)\sum_{j=1}^R y_i^j$ based on majority voting.

2. Given μ_i, estimate the sensitivity and specificity of each annotator/expert as follows.

$$\begin{aligned} \alpha^j &= \frac{a_1^j - 1 + \sum_{i=1}^N \mu_i y_i^j}{a_1^j + a_2^j - 2 + \sum_{i=1}^N \mu_i}. \\ \beta^j &= \frac{b_1^j - 1 + \sum_{i=1}^N (1-\mu_i)(1-y_i^j)}{b_1^j + b_2^j - 2 + \sum_{i=1}^N (1-\mu_i)}. \end{aligned} \tag{4}$$

   The Newton-Raphson update for optimizing $\boldsymbol{w}$ is given by $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \boldsymbol{H}^{-1}\mathbf{g}$, with step length η, gradient vector

$$\boldsymbol{g}(w) = \sum_{i=1}^N \left[\mu_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\right] \boldsymbol{x}_i - \boldsymbol{\Gamma}\boldsymbol{w},$$

   and Hessian matrix

$$\boldsymbol{H}(w) = -\sum_{i=1}^N \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\left[1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\right] \boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{\Gamma}.$$

10. It may be convenient to specify a prior in terms of the mean μ and variance σ^2. The mean and the variance for a beta prior are given by $\mu = a/(a+b)$ and $\sigma^2 = ab/((a+b)^2(a+b+1))$. Solving for a and b we get $a = (-\mu^3 + \mu^2 - \mu\sigma^2)/\sigma^2$ and $b = a(1-\mu)/\mu$.

3. Given the sensitivity and specificity of each annotator and the model parameters, update μ_i as

$$\mu_i = \frac{a_i p_i}{a_i p_i + b_i(1-p_i)}, \tag{5}$$

where

$$\begin{aligned} p_i &= \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i). \\ a_i &= \prod_{j=1}^{R} [\alpha^j]^{y_i^j} [1-\alpha^j]^{1-y_i^j}. \\ b_i &= \prod_{j=1}^{R} [\beta^j]^{1-y_i^j} [1-\beta^j]^{y_i^j}. \end{aligned} \tag{6}$$

Iterate (2) and (3) till convergence.

2.7 Discussions

1. **Estimate of the gold standard** The value of the posterior probability μ_i is a soft probabilistic estimate of the actual ground truth y_i, that is, $\mu_i = \Pr[y_i = 1 | y_i^1, \ldots, y_i^R, \boldsymbol{x}_i, \theta]$. The actual hidden label y_i can be estimated by applying a threshold on μ_i, that is, $y_i = 1$ if $\mu_i \geq \gamma$ and zero otherwise. We can use $\gamma = 0.5$ as the threshold. By varying γ we can change the misclassification costs and obtain a ground truth with large sensitivity or large specificity. Because of this in our experimental validation we can actually draw an ROC curve for the estimated ground truth.

2. **Log-odds of μ** A particularly revealing insight can be obtained in terms of the log-odds or the *logit* of the posterior probability μ_i. From (5) the logit of μ_i can be written as

$$\begin{aligned} \text{logit}(\mu_i) = \ln \frac{\mu_i}{1-\mu_i} &= \ln \frac{\Pr[y_i = 1 | y_i^1, \ldots, y_i^R, \boldsymbol{x}_i, \theta]}{\Pr[y_i = 0 | y_i^1, \ldots, y_i^R, \boldsymbol{x}_i, \theta]} \\ &= \boldsymbol{w}^\top \boldsymbol{x}_i + c + \sum_{j=1}^{R} y_i^j [\text{logit}(\alpha^j) + \text{logit}(\beta^j)]. \end{aligned}$$

where $c = \sum_{j=1}^{R} \log \frac{1-\alpha^j}{\beta^j}$ is a constant term which does not depend on i. This indicates that the estimated ground truth (in the logit form of the posterior probability) is a *weighted linear combination* of the labels from all the experts. The weight of each expert is the sum of the logit of the sensitivity and specificity.

3. **Using any other classifier** For ease of exposition we used logistic regression. However, the proposed algorithm can be used with any generalized linear model or in fact with any classifier that can be trained with soft probabilistic labels. In each step of the EM-algorithm, the classifier is trained with instances sampled from μ_i. This modification is easy for most probabilistic classifiers. For general black-box classifiers where we cannot tweak the training algorithm an alternate approach is to replicate the training examples according to the soft label. For example a probabilistic label $\mu_i = 0.8$ can be effectively simulated by adding 8 training examples with deterministic label 1 and 2 examples with label 0.

4. **Obtaining ground truth with no features** In some scenarios we may not have features $\boldsymbol{x}_i$ and we wish to obtain an estimate of the actual ground truth based only on the labels from multiple annotators. Here instead of learning a classifier we estimate p which is the prevalence of the positive class, that is, $p = \Pr[y_i = 1]$. We further assume a beta prior for the prevalence, that is, $\text{Beta}(p|p_1, p_2)$. The algorithm simplifies as follows.

   (a) Initialize $\mu_i = (1/R)\sum_{j=1}^{R} y_i^j$ based on majority voting.

   (b) Given μ_i, estimate the sensitivity and specificity of each annotator using (4). The prevalence of the positive class is estimated as follows.

$$p = \frac{p_1 - 1 + \sum_{i=1}^{N} \mu_i}{p_1 + p_2 - 2 + N}.$$

   (c) Given the sensitivity and specificity of each annotator and prevalence, refine μ_i as follows.

$$\mu_i = \frac{a_i p}{a_i p + b_i (1-p)}.$$

   Iterate (2) and (3) till convergence. This algorithm is similar to the one proposed by Dawid and Skene (1979) and Smyth et al. (1995).

5. **Handling missing labels** The proposed approach can easily handle missing labels, that is, when the labels from some experts are missing for some instances. Let R_i be the number of radiologists labeling the i^{th} instance, and let N_j be the number of instances labeled by the j^{th} radiologist. Then in the EM algorithm, we just need to replace N by N_j for estimating the sensitivity and specificity in (4), and replace R by R_i for updating μ_i in (6).

6. **Evaluating a classifier** We can use the probability scores μ_i directly to evaluate classifiers. If z_i are the labels obtained from any other classifier, then sensitivity and specificity can be estimated as

$$\alpha = \frac{\sum_{i=1}^{N} \mu_i z_i}{\sum_{i=1}^{N} \mu_i}, \quad \beta = \frac{\sum_{i=1}^{N} (1-\mu_i)(1-z_i)}{\sum_{i=1}^{N} (1-\mu_i)}.$$

7. **Posterior approximation** At the end of each EM iteration a crude approximation to the posterior is obtained as

$$\begin{aligned} \alpha_j &\sim \text{Beta}\left(\alpha_j | a_1^j + \sum_{i=1}^{N} \mu_i y_i^j, a_2^j + \sum_{i=1}^{N} \mu_i (1 - y_i^j)\right), \\ \beta_j &\sim \text{Beta}\left(\beta_j | b_1^j + \sum_{i=1}^{N} (1-\mu_i)(1-y_i^j), b_2^j + \sum_{i=1}^{N} (1-\mu_i) y_i^j\right). \end{aligned}$$

3. Multi-class Classification

In this section we describe how the proposed approach for binary classification can be extended to categorical data. Suppose there are $K \geq 2$ categories. An example for categorical data from the CAD domain is in LungCAD, where the radiologist needs to label whether a nodule (known to be precursors of cancer) is a solid, a part-solid, or a ground glass opacity—which are three

different kinds on nodules. We can extend the previous model and introduce a vector of multinomial parameters $\boldsymbol{\alpha}_c^j = (\alpha_{c1}^j, \ldots, \alpha_{cK}^j)$ for each annotator, where

$$\alpha_{ck}^j := \Pr[y^j = k | y = c]$$

and $\sum_{k=1}^K \alpha_{ck}^j = 1$. Here α_{ck}^j denotes the probability that the annotator j assigns class k to an instance given the true class is c. When $K = 2$, α_{11}^j and α_{00}^j are sensitivity and specificity, respectively. A similar EM algorithm can be derived. In the E-step, we estimate

$$\Pr[y_i = c | \mathcal{Y}, \boldsymbol{\alpha}] \propto \Pr[y_i = c | \boldsymbol{x}_i] \prod_{j=1}^R \prod_{k=1}^K (\alpha_{ck}^j)^{\delta(y_i^j, k)},$$

where $\delta(u, v) = 1$ if $u = v$ and 0 otherwise and in the M-step we learn a multi-class classifier and update the multinomial parameter as

$$\alpha_{ck}^j = \frac{\sum_{i=1}^N \Pr[y_i = c | \mathcal{Y}, \boldsymbol{\alpha}] \delta(y_i^j, k)}{\sum_{i=1}^N \Pr[y_i = c | \mathcal{Y}, \boldsymbol{\alpha}]}.$$

One can also assign a Dirichlet prior for the multinomial parameters, and this results in a smoothing term in the above updates in the MAP estimate.

4. Ordinal Regression

We now consider the situation where the outputs are categorical and have an ordering among the labels. In the CAD domain the radiologist often gives a score (for example, 1 to 5 from lowest to highest) to indicate how likely she thinks it is malignant. This is different from a multi-class setting in which we do not have any preference among the multiple class labels.

Let $y_i^j \in \{1, \ldots, K\}$ be the label assigned to the i^{th} instance by the j^{th} expert. Note that there is an ordering in the labels $1 < \ldots < K$. A simple approach is to convert the ordinal data into a series of binary data (Frank and Hall, 2001). Specifically the K class ordinal labels are transformed into $K - 1$ binary class labels as follows:

$$y_i^{jc} = \begin{cases} 1 & \text{if } y_i^j > c \\ 0 & \text{otherwise} \end{cases} \qquad c = 1, \ldots, K-1.$$

Applying the same procedure used for binary labels we can estimate $\Pr[y_i > c]$ for $c = 1, \ldots, K-1$. The probability of the actual class values can then be obtained as

$$\Pr[y_i = c] = \Pr[y_i > c - 1 \text{ and } y_i \leq c] = \Pr[y_i > c - 1] - \Pr[y_i > c].$$

The class with the maximum probability is assigned to the instance.

5. Regression

In this section we develop a similar algorithm to learn a regression function using annotations from multiple experts. In the CAD domain as a part of the annotation process a common task for a radiologist is to measure the diameter of a suspicious lesion.

5.1 Model for Annotators

Let $y_i^j \in \mathbb{R}$ be the continuous target value assigned to the i^{th} instance by the j^{th} annotator. Our model is that the annotator provides a noisy version of the actual true value y_i. For the j^{th} annotator we will assume a Gaussian noise model with mean y_i (the true unknown value) and inverse-variance (precision) τ^j, that is,

$$\Pr[y_i^j|y_i,\tau^j] = \mathcal{N}(y_i^j|y_i,1/\tau^j), \tag{7}$$

where the Gaussian distribution is defined as $\mathcal{N}(z|m,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp(-(z-m)^2/2\sigma^2)$. The unknown inverse-variance τ^j measures the accuracy of each annotator—the larger the value of τ^j the more accurate the annotator. We have assumed that τ^j does not depend on the instance $\boldsymbol{x}_i$. For example, in the CAD domain, this means that the radiologist's accuracy does not depend on the nodule she is measuring. While this a practical assumption, it is not entirely true. It is known that some nodules are harder to measure than others.

5.2 Linear Regression Model for Features

As before we consider the family of linear regression functions: $\mathcal{F} = \{f_{\boldsymbol{w}}\}$, where for any $\boldsymbol{x},\boldsymbol{w} \in \mathbb{R}^d$, $f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$. We assume that the actual target response y_i is given by the deterministic regression function $f_{\boldsymbol{w}}$ with additive Gaussian noise, that is,

$$y_i = \boldsymbol{w}^\top \boldsymbol{x}_i + \varepsilon,$$

where ε is a zero-mean Gaussian random variable with inverse-variance (precision) γ. Hence

$$\Pr[y_i|\boldsymbol{x}_i,\boldsymbol{w},\gamma] = \mathcal{N}(y_i|\boldsymbol{w}^\top \boldsymbol{x}_i,1/\gamma). \tag{8}$$

5.3 Combined Model

Combining both the annotator (7) and the regressor (8) model we have

$$\Pr[y_i^j|\boldsymbol{x}_i,\boldsymbol{w},\tau^j,\gamma] = \int \Pr[y_i^j|y_i,\tau^j]\Pr[y_i|\boldsymbol{x}_i,\boldsymbol{w},\gamma]dy_i = \mathcal{N}(y_i^j|\boldsymbol{w}^\top \boldsymbol{x}_i,1/\gamma+1/\tau^j).$$

Since the two precision terms (γ and τ_j) are grouped together they are not uniquely identifiable. Hence we will define a new precision term λ^j as

$$\frac{1}{\lambda^j} = \frac{1}{\gamma} + \frac{1}{\tau^j}.$$

So we have the following model

$$\Pr[y_i^j|\boldsymbol{x}_i,\boldsymbol{w},\lambda^j] = \mathcal{N}(y_i^j|\boldsymbol{w}^\top \boldsymbol{x}_i,1/\lambda^j). \tag{9}$$

5.4 Estimation/Learning Problem

Given the training data $\mathcal{D}$ consisting of N instances with annotations from R experts, that is, $\mathcal{D} = \{\boldsymbol{x}_i,y_i^1,\ldots,y_i^R\}_{i=1}^N$, the task is to estimate the weight vector $\boldsymbol{w}$ and the precision $\boldsymbol{\lambda} = [\lambda^1,\ldots,\lambda^R]$ of all the annotators.

5.5 Maximum-likelihood Estimator

Assuming the instances are independent the likelihood of the parameters $\theta = \{\boldsymbol{w}, \boldsymbol{\lambda}\}$ given the observations $\mathcal{D}$ can be factored as

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} \Pr[y_i^1, \ldots, y_i^R | \boldsymbol{x}_i, \theta].$$

Conditional on the instance $\boldsymbol{x}_i$ we assume that $y_i^1, \ldots, y_i^R$ are independent, that is, the annotators provide their responses independently. Hence from (9) the likelihood can be written as

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} \prod_{j=1}^{R} \mathcal{N}(y_i^j | \boldsymbol{w}^\top \boldsymbol{x}_i, 1/\lambda^j).$$

The maximum-likelihood estimator is found by maximizing the log-likelihood

$$\widehat{\theta}_{\text{ML}} = \{\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{w}}\} = \arg\max_{\theta} \{\ln \Pr[\mathcal{D}|\theta]\}.$$

By equating the gradient of the log-likelihood to zero we obtain the following update equations for the precision and the weight vector.

$$\frac{1}{\widehat{\lambda}^j} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^j - \widehat{\boldsymbol{w}}^\top \boldsymbol{x_i} \right)^2. \tag{10}$$

$$\widehat{\boldsymbol{w}} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i \left(\frac{\sum_{j=1}^{R} \widehat{\lambda}^j y_i^j}{\sum_{j=1}^{R} \widehat{\lambda}^j} \right). \tag{11}$$

As the parameters $\widehat{\boldsymbol{w}}$ and $\widehat{\boldsymbol{\lambda}}$ are coupled together we iterate these two steps till convergence.

5.6 Discussions

1. **Is this standard least-squares?** Define the design matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$ and the response vector for each annotator as $\boldsymbol{y}^j = [y_1^j, \ldots, y_N^j]^\top$. Using matrix notation Equation 11 can be written as

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{y}} \quad \text{where} \quad \widehat{\boldsymbol{y}} = \frac{\sum_{j=1}^{R} \widehat{\lambda}^j \boldsymbol{y}^j}{\sum_{j=1}^{R} \widehat{\lambda}^j}. \tag{12}$$

   Equation 12 is essentially the solution to a standard linear regression model, except that we are training a linear regression model with $\widehat{\boldsymbol{y}}$ as the ground truth, which is a precision weighted mean of the response vectors from all the annotators. The variance of each annotator is estimated using (10). The final algorithm iteratively establishes a particular gold standard ($\widehat{\boldsymbol{y}}$), measures the performance of the annotators and learns a regressor given that gold standard, and refines the gold standard based on the performance measures.

2. **Are we better than the best annotator?** If we assume $\widehat{\boldsymbol{\lambda}}$ is fixed (i.e., we ignore the variability and assume that it is well estimated) then $\hat{\boldsymbol{w}}$ is an unbiased estimator of $\boldsymbol{w}$ and the covariance matrix is given by

$$\text{Cov}(\hat{\boldsymbol{w}}) = \text{Cov}(\widehat{\boldsymbol{y}}) \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} = \frac{1}{\sum_{j=1}^{R} \widehat{\lambda}_j} \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1}.$$

Since $\sum_{j=1}^{R} \widehat{\lambda}_j > \max_j(\widehat{\lambda}_j)$ the proposed method has a lower variance than the regressor learnt with the best annotator (i.e., the one with the minimum variance).

3. **Are we better than the average?** For a fixed $\boldsymbol{X}$ the error in $\widehat{w}$ depends only on the variance of $\widehat{y}^j$. If we know the true λ^j then $\widehat{y}_i$ is the best linear unbiased estimator for y_i which minimizes the variance. To see this consider any linear estimator of the form $\widehat{y}_i = \sum_j a^j (y_i^j - b^j)$. The variance is given by $\mathrm{Var}[\widehat{y}_i] = \sum_j (a^j)^2/\lambda_j$. Since $\mathrm{E}[\widehat{y}_i] = y_i \sum_j a^j$, for the bias of this estimator to be zero we require that $\sum_j a^j = 1$. Solving the constrained minimization problem we see that $a_j = \lambda^j / \sum_j \lambda_j$ minimizes the variance.

4. **Obtaining a consensus without features** When no features are available the same algorithm can be simplified to get a consensus estimate of the actual ground truth and also evaluate the annotators. Essentially we have to iterate the following two updates till convergence

$$\widehat{y}_i = \frac{\sum_{j=1}^{R} \widehat{\lambda}^j y_i^j}{\sum_{j=1}^{R} \widehat{\lambda}^j} \quad \frac{1}{\widehat{\lambda}^j} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^j - \widehat{y}_i \right)^2.$$

6. Experimental Validation

We now experimentally validate the proposed algorithms on both simulated and real data.

6.1 Classification Experiments

We use two CAD and one text data set in our experiments. The CAD data sets include a digital mammography data set and a breast MRI data set, both of which are biopsy proven, that is, the gold standard is available. For the digital mammography data set we simulate the radiologists in order to validate our methods. The breast MRI data has annotations from four radiologists. We also report results on a Recognizing Textual Entailment data collected by Snow et al. (2008) using the Amazon's Mechanical Turk which has annotations from 164 annotators.

6.1.1 DIGITAL MAMMOGRAPHY WITH SIMULATED RADIOLOGISTS

Mammograms are used as a screening tool to detect early breast cancer. CAD systems search for abnormal areas (*lesions*) in a digitized mammographic image. These lesions generally indicate the presence of malignant cancer. The CAD system then highlights these areas on the images, alerting the radiologist to the need for a further diagnostic mammogram or a biopsy. In classification terms, given a set of descriptive morphological features for a region on a image, the task is to predict whether it is potentially malignant (1) or not (0). In order to train such a classifier, a set of mammograms is collected from hospitals. The ground truth (whether it is cancer or not) is obtained from biopsy. Since biopsy is an expensive, tedious, and an invasive process, very often CAD systems are built from labels collected from *multiple expert radiologists* who visually examine the mammograms and mark the lesion locations—this constitutes our ground truth (multiple labels) for learning.

In this experiment we use a proprietary biopsy-proven data set (Krishnapuram et al., 2008) containing 497 positive and 1618 negative examples. Each instance is described by a set of 27 morphological features. In order to validate our proposed algorithm, we simulate multiple radiologists according to the two-coin model described in § 2.1. Based on the labels from multiple radiologists,

we can simultaneously (1) learn a logistic-regression classifier, (2) estimate the sensitivity and specificity of each radiologist, and (3) estimate the golden ground truth. We compare the results with the classifier trained using the biopsy proved ground truth as well as the majority-voting baseline. For the first set of experiments we use 5 radiologists with sensitivity $\boldsymbol{\alpha} = [0.90\ 0.80\ 0.57\ 0.60\ 0.55]$ and specificity $\boldsymbol{\beta} = [0.95\ 0.85\ 0.62\ 0.65\ 0.58]$. This corresponds to a scenario where the first two radiologists are experts and the last three are novices. Figure 1 summarizes the results. We compare on three different aspects: (1) How good is the learnt classifier? (2) How well can we estimate the sensitivity and specificity of each radiologist? (3) How good is the estimated ground truth? The following observations can be made.

1. **Classifier performance** Figure 1(a) plots the ROC curve of the learnt classifier on the training set. The dotted (black) line is the ROC curve for the classifier learnt using the actual ground truth. The solid (red) line is the ROC curve for the proposed algorithm and the dashed (blue) line is for the classifier learnt using the majority-voting scheme. The classifier learnt using the proposed method is as good as the one learnt using the golden ground truth. The area under the ROC curve (AUC) for the proposed algorithm is around 3.5% greater than that learnt using the majority-voting scheme.

2. **Radiologist performance** The actual sensitivity and specificity of each radiologist is marked as a black $\times$ in Figure 1(b). The end of the solid red line shows the estimates of the sensitivity and specificity from the proposed method. We used a uniform prior on all the parameters. The ellipse plots the contour of one standard deviation as obtained from the beta posterior estimates. The end of the dashed blue line shows the estimate obtained from the majority-voting algorithm. We see that the proposed method is much closer to the actual values of sensitivity and specificity.

3. **Actual ground truth** Since the estimates of the actual ground truth are probabilistic scores, we can also plot the ROC curves of the estimated ground truth. From Figure 1(b) we can see that the ROC curve for the proposed method dominates the majority voting ROC curve. Furthermore, the area under the ROC curve (AUC) is around 3% higher. The estimate obtained by majority voting is closer to the novices since they form a majority (3/5). It does not have an idea of who is an expert and who is a novice. The proposed algorithm appropriately weights each radiologist based on their estimated sensitivity and specificity. The improvement obtained is quite large in Figure 2 which corresponds a situation where we have only one expert and 7 novices.

4. **Joint Estimation** To learn a classifier, Smyth et al. (1995) proposed to first estimate the golden ground truth and then use the probabilistic ground truth to learn a classifier. In contrast, our proposed algorithm learns the classifier and the ground truth *jointly* as a part of the EM algorithm. Figure 3 shows that the classifier and the ground truth learnt obtained by the proposed algorithm is superior than that obtained by other procedures which first estimates the ground truth and then learns the classifier.

6.1.2 BREAST MRI

In this example, each radiologist reviews the breast MRI data and assesses the malignancy of each lesion on a BIRADS scale of 1 to 5. The BIRADS scale is defined as follows: 1 Negative, 2 Benign,

Majority Voting	True 1	True 2	True 3	True 4	True 5
Estimated 1	x	0.0217	0	x	0.0000
Estimated 2	x	**0.5869**	0	x	0.1785
Estimated 3	x	0.2391	0	x	0.1071
Estimated 4	x	0.1521	1	x	0.2500
Estimated 5	x	0.0000	0	x	**0.4642**
EM algorithm	**True 1**	**True 2**	**True 3**	**True 4**	**True 5**
Estimated 1	x	0.0000	0	x	0.0000
Estimated 2	x	**0.6957**	0	x	0.1428
Estimated 3	x	0.1304	0	x	0.0000
Estimated 4	x	0.1739	1	x	0.3214
Estimated 5	x	0.0000	0	x	**0.5357**

Table 1: The confusion matrix for the estimate obtained using majority voting and the proposed EM algorithm. The x indicates that there was no such category in the true labels (the gold standard). The gold-standard is obtained by the biopsy which can confirm whether it is benign (BIRADS=2) or malignant (BIRADS=5).

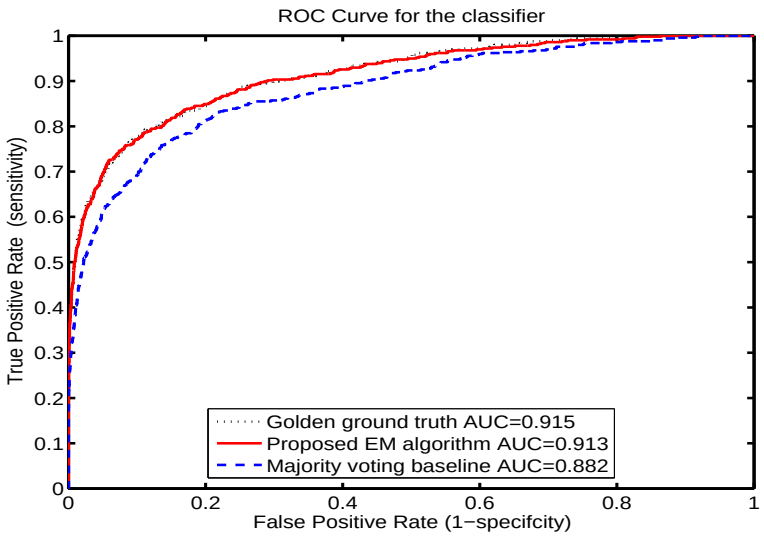
3 Probably Benign, 4 Suspicious abnormality, and 5 Highly suggestive of malignancy. Our data set comprises of 75 lesions with annotations from four radiologists, and the true labels from biopsy. Based on eight morphological features, we have to predict whether a lesion is malignant or not.

For the first experiment we reduce the BIRADS scale to a binary one: any lesion with a BIRADS > 3 is considered malignant and benign otherwise. The set included 28 malignant and 47 benign lesions. Figure 4 summarizes the results. We show the leave-one-out cross validated ROC for the classifier. The cross-validated AUC of the proposed method is approximately 6% better than the majority voting baseline.

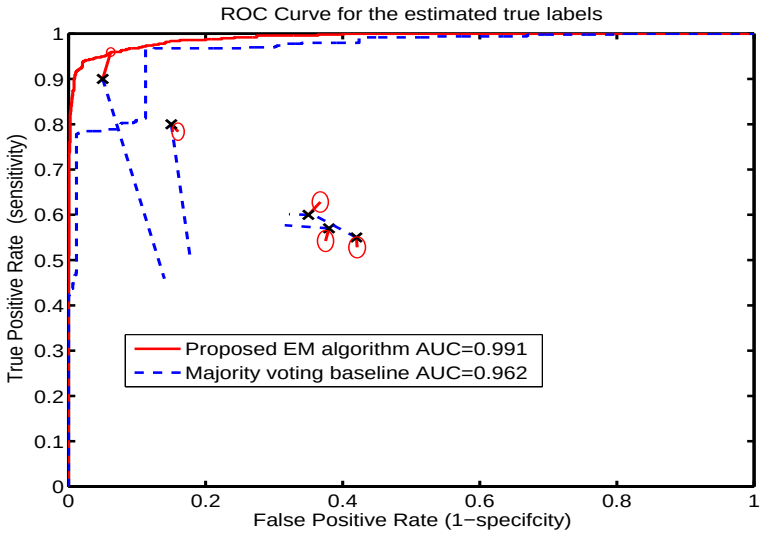
We also consider the BIRADS labels as a set of ordinal measurements since there is an ordering among the BIRADS label. The confusion matrix in Table 1 shows that the EM algorithm is significantly superior than the majority voting in estimating the true BIRADS.

6.1.3 RECOGNIZING TEXTUAL ENTAILMENT

Finally we report results on Recognizing Textual Entailment data collected by Snow et al. (2008) using the Amazon's Mechanical Turk. In this task, the annotator is presented with two sentences and given a choice of whether the second sentence can be inferred from the first. The data has 800 tasks and 164 distinct readers, with 10 annotations per task along with the golden ground truth. The majority of the entries (94 %) in the 800x164 matrix are missing. There is one annotator who has labeled all the tasks. We use this data set to obtain an estimate of the actual ground truth. Figure 5 plots the accuracy of the estimated ground truth as a function of the number of annotators. The proposed EM algorithm achieves a higher accuracy than majority voting. In other words to achieve a desired accuracy the proposed algorithm needs fewer annotators than the majority voting scheme.

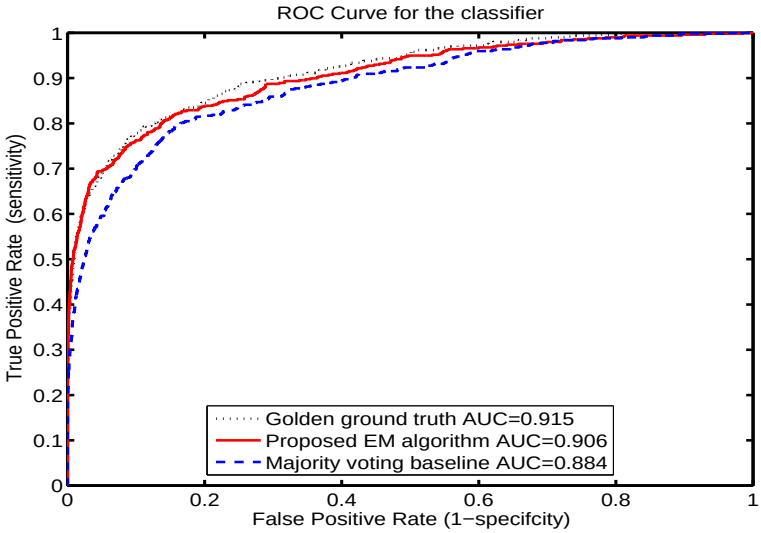


(a)

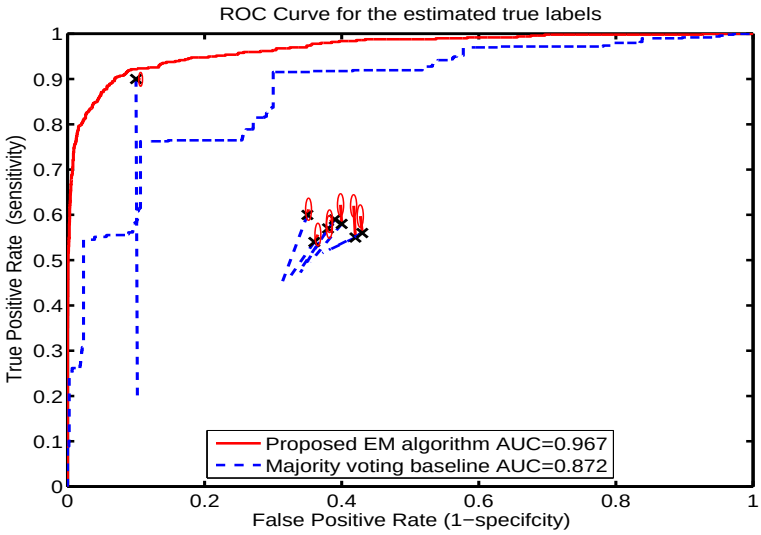


(b)

Figure 1: Results for the digital mammography data set with annotations from 5 simulated radiologists. (a) The ROC curve of the learnt classifier using the golden ground truth (dotted black line), the majority voting scheme (dashed blue line), and the proposed EM algorithm (solid red line). (b) The ROC curve for the estimated ground truth. The actual sensitivity and specificity of each of the radiologists is marked as a $\times$. The end of the dashed blue line shows the estimates of the sensitivity and specificity obtained from the majority voting algorithm. The end of the solid red line shows the estimates from the proposed method. The ellipse plots the contour of one standard deviation.

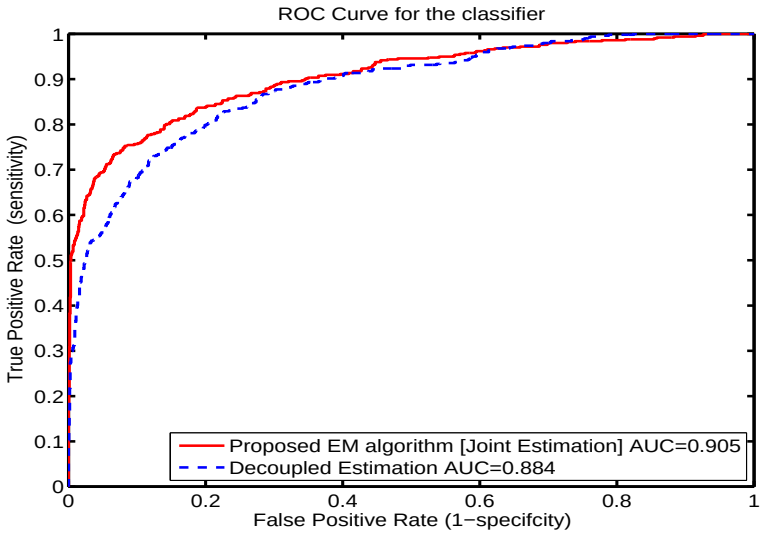


(a)

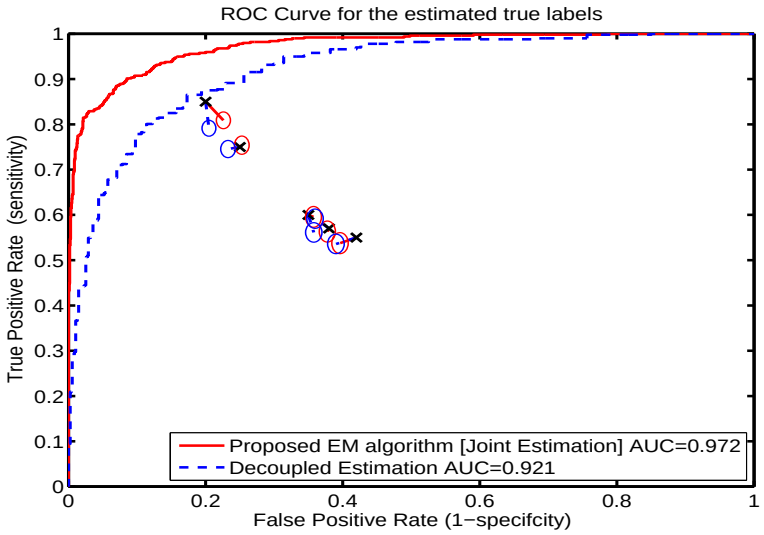


(b)

Figure 2: Same as Figure 1 except with 8 different radiologist annotations.

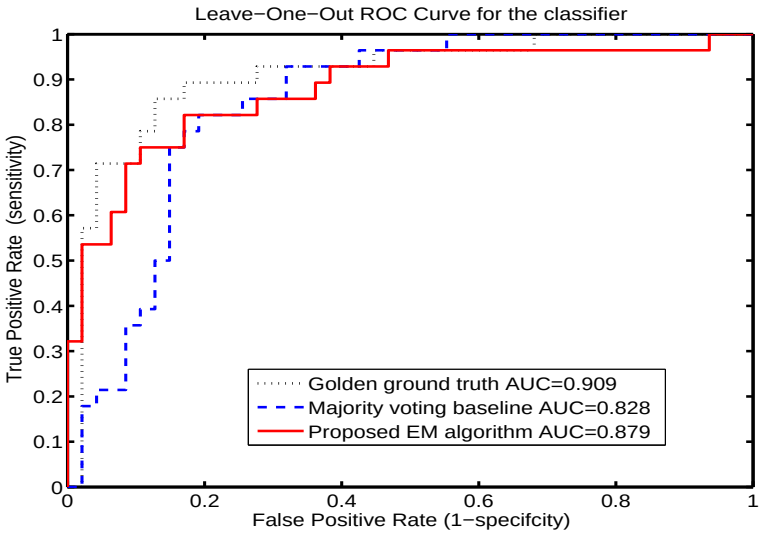


(a)

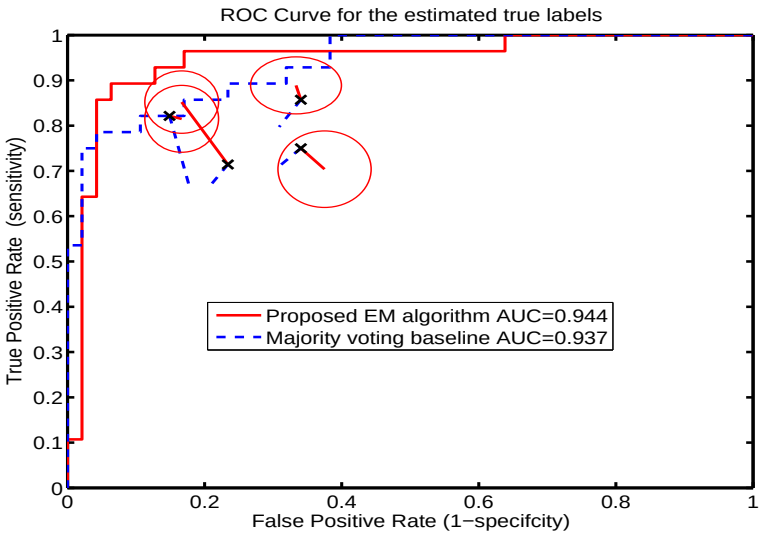


(b)

Figure 3: ROC curves comparing the proposed algorithm (solid red line) with the *Decoupled Estimation* procedure (dotted blue line), which refers to the algorithm where the ground truth is first estimated using just the labels from the five radiologists and then a logistic regression classifier is trained using the soft probabilistic labels. In contrast the proposed EM algorithm estimates the ground truth and learns the classifier simultaneously during the EM algorithm.



(a)



(b)

Figure 4: Breast MRI results. (a) The leave-one-out cross validated ROC. (b) ROC for the estimated ground truth.

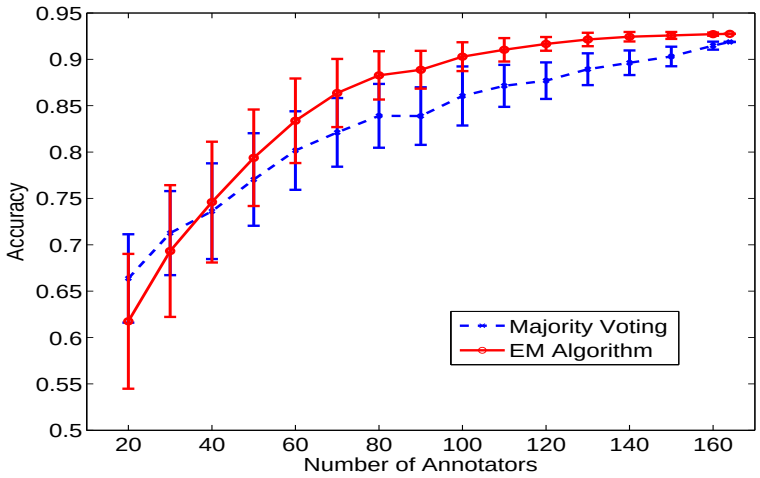


Figure 5: The mean and the one standard deviation error bars for the accuracy of the estimated ground truth for the Recognizing Textual Entailment task as a function of the number of annotators. The plot was generated by randomly sampling the annotators 100 times.

6.2 Regression Experiments

We first illustrate the algorithm on a toy dataset and then present a case study for automated polyp measurements.

6.2.1 ILLUSTRATION

Figure 6 illustrates the the proposed algorithm for regression on a one-dimensional toy data set with three annotators. The actual regression model (shown as a blue dotted line) is given by $y = 5x - 2$. We simulate 20 samples from three annotators with precisions 0.01, 0.1, and 1.0. The data are shown by the annotators's number. While we can fit a regression model using each annotators's response, we see that only the model for annotator three (with highest precision) is close to the true regression model. The green dashed line shows the model learnt using the average response from all the three annotators. The red line shows the model learnt by the proposed algorithm.

6.2.2 AUTOMATED POLYP MEASUREMENTS

Colorectal polyps are small colonic findings that may develop into cancer at a later stage. The diameter of the polyp is one of the key factors which decides the malignancy of a suspicious polyp. Hence accurate size estimation is crucial to decide the action to be taken on a polyp. We have developed various algorithms to segment a polyp. Multiple segmentation algorithms give rise to a set of features which are correlated with the diameter of the polyp. We want to learn a regression function which can predict the diameter of a polyp as a function of these features. In order to learn a regression function we collect our ground truth by asking many radiologists to manually measure the the diameter of the polyps from the three-dimensional images. In practice there is a lot of disagreement among the radiologists as to the actual size of the polyp.

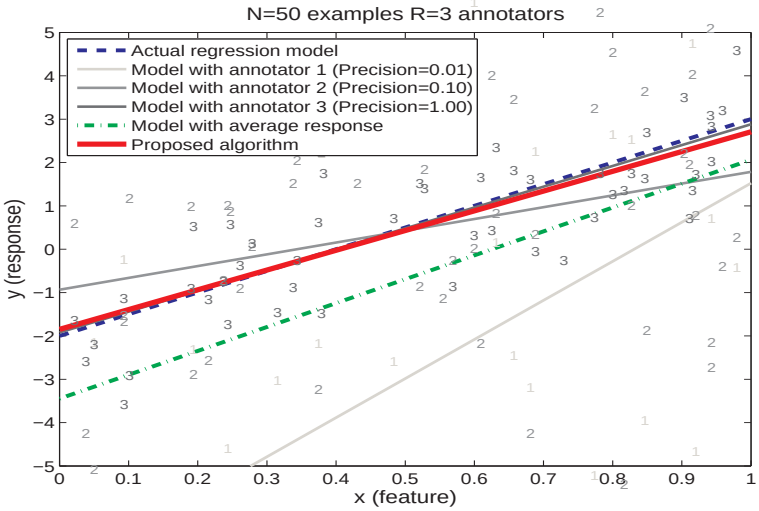
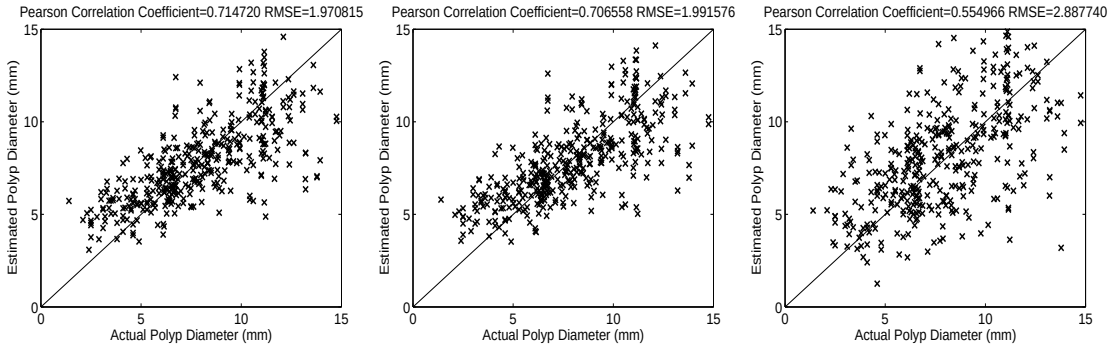


Figure 6: Illustration of the proposed algorithm on a one-dimensional toy data set. The actual regression model (shown as a blue dotted line) is given by $y = 5x - 2$. We simulate 50 samples from three annotators with precisions 0.01, 0.1, and 1.0. The data are shown by the annotators's number. While we can fit a regression model using each annotators's response, we see that only the model for annotator three (with highest precision) is close to the true regression model. The green dashed line shows the model learnt using the average response from all the three annotators. The red line shows the model learnt by the proposed algorithm.



(a) Gold standard model (b) Proposed model (c) Average model

Figure 7: Scatter plot of the actual polyp diameter vs the diameter predicted by the models learnt using (a) the actual gold standard, (b) the proposed algorithm with annotations from five radiologists, and (c) the average of the radiologist's annotations. (See § 6.2.2 for a description of the experimental setup.)

We use a proprietary data set containing 393 examples (which point to 285 distinct polyps—the segmentation algorithms generally return multiple marks on the same polyp.) along with the measured diameter (ranging from 2mm to 15mm) as our training set. Each example is described by a set of 60 morphological features which are correlated to the diameter of the polyp. In order to validate the feasibility of our proposed algorithm, we simulate five radiologists according to the noisy model described in § 5.1 with $\tau = [0.001\ 0.01\ 0.1\ 1\ 10]$. This corresponds to a situation where the first three radiologists are extremely noisy and the last two are quite accurate. Based on the measurements from multiple radiologists, we can simultaneously (1) learn a linear regressor and (2) estimate the precision of each radiologist. We compare the results with the classifier trained using the actual golden ground truth as well as the regressor learnt using the average of the radiologists measurements. The results are validated on an independent test set containing 397 examples (which point to 298 distinct polyps).

Figure 7 shows the scatter plot of the actual polyp diameter vs the diameter predicted by the three different models. We compare the performance based on the root mean squared error (RMSE) and also the Pearson's correlation coefficient. The regressor learnt using the proposed iterative algorithm (Figure 7(b)) is almost as good as the one learnt using the golden ground truth (Figure 7(a)). The correlation coefficient for the proposed algorithm is significantly larger than that learnt using the average of the radiologists response. The estimate obtained by averaging is closer to the novices since they form a majority (3/5). The proposed algorithm appropriately weights each radiologist based on their estimated precisions.

7. Conclusions and Future Work

In this paper we proposed a probabilistic framework for supervised learning with multiple annotators providing labels but no absolute gold standard. The proposed algorithm iteratively establishes a particular gold standard, measures the performance of the annotators given that gold standard, and then refines the gold standard based on the performance measures. We specifically discussed binary/categorical/ordinal classification and regression problems.

We made two key assumptions: (1) the performance of each annotator does not depend on the feature vector for a given instance and (2) conditional on the truth the experts are independent, that is, they make their errors independently. As we pointed out earlier these assumptions are not true in practice. The annotator performance depends on the instance he is labeling and there is some degree of correlation among the annotators. We briefly discuss some strategies to relax these two assumptions.

7.1 Instance Difficulty

One drawback of the current model is that it doesn't estimate difficulty of items. It is often observed that for the easy instances all the annotators agree on the labels—thus violating our conditional independence assumption. The difficulty of annotating an item can be captured by another latent variable γ_i for each instance—which modulates the annotators performance. Models for this have been developed in the area of item-response theory (Baker and Kim, 2004) and also in epidemiology (Uebersax and Grove, 1993)—see also Whitehill et al. (2009) for a recent paper in the machine learning community. While these models do not take into account the available features our pro-

posed model for sensitivity and specificity can be extended as follows (in place of (1) and (2)):

$$\alpha^j(\gamma_i) := \Pr[y_i^j = 1 | y_i = 1, \gamma_i] = \sigma(a_{j1} + b_{j1}\gamma_i).$$

$$\beta^j(\gamma_i) := \Pr[y_i^j = 0 | y_i = 0, \gamma_i] = \sigma(a_{j0} + b_{j0}\gamma_i).$$

Here the parameters a_{j1} and a_{j0} are related to the sensitivity and specificity of the j^{th} annotator, while the latent term γ_i captures the difficulty of the instance. The key assumption here is that the annotators are independent conditional on both y_i and γ_i. Various assumptions can be made on two parameters b_{j1} and b_{j0} to simplify these models further—for example we could set $b_{j1} = b_1$ and $b_{j0} = b_0$ for all the annotators.

7.2 Annotators Actually Look at the Data

In our model we made the assumption that the sensitivity α^j and the specificity β^j of the j^{th} annotator does not depend on the feature vector $\boldsymbol{x}_i$. For example, in the CAD domain, this meant that the radiologist's performance is consistent across different sub-groups of data—which is not entirely true. It is known that some radiologists are good at detecting certain kinds of malignant lesions based on their training and experience. We can extend the previous model such that the sensitivity and the specificity depends on the feature vector $\boldsymbol{x}_i$ explicitly as follows

$$\alpha^j(\gamma_i, \boldsymbol{x}_i) := \Pr[y_i^j = 1 | y_i = 1, \gamma_i, \boldsymbol{x}_i] = \sigma(a_{j1} + b_{j1}\gamma_i + \boldsymbol{w}_\alpha^{j\top} \boldsymbol{x}_i).$$

$$\alpha^j(\gamma_i, \boldsymbol{x}_i) := \Pr[y_i^j = 0 | y_i = 0, \gamma_i, \boldsymbol{x}_i] = \sigma(a_{j0} + b_{j0}\gamma_i + \boldsymbol{w}_\beta^{j\top} \boldsymbol{x}_i).$$

However this change increases the number of parameters to be learned.

References

P. S. Albert and L. E. Dodd. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60:427–435, 2004.

F. B. Baker and S. Kim. *Item Response Theory: Parameter Estimation Techniques*. CRC Press, 2 edition, 2004.

B. Carpenter. Multilevel bayesian models of categorical data annotation. Technical Report available at http://lingpipe-blog.com/lingpipe-white-papers/, 2008.

S. R. Cholleti, S. A. Goldman, A. Blum, D. G. Politte, and S. Don. Veritas: Combining expert opinions without labeled data. In *Proceedings of the 2008 20th IEEE international Conference on Tools with Artificial intelligence*, 2008.

A. P. Dawid and A. M. Skene. Maximum likeihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979.

O. Dekel and O. Shamir. Vox Populi: Collecting high-quality labels from a crowd. In *COLT 2009: Proceedings of the 22nd Annual Conference on Learning Theory*, 2009a.

O. Dekel and O. Shamir. Good learners for evil teachers. In *ICML 2009: Proceedings of the 26th International Conference on Machine Learning*, pages 233–240, 2009b.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.

P. Donmez, J. G. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *KDD 2009: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 259–268, 2009.

E. Frank and M. Hall. A simple approach to ordinal classification. *Lecture Notes in Computer Science*, pages 145–156, 2001.

G. Fung, B. Krishnapuram, J. Bi, M. Dundar, V. C. Raykar, S. Yu, R. Rosales, S. Krishnan, and R. B. Rao. Mining medical images. In *Fifteenth Annual SIGKDD International Conference on Knowledge Discovery and Data Mining: Third Workshop on Data Mining Case Studies and Practice Prize*, 2009.

J. Howe. *Crowd sourcing: Why the Power of the Crowd Is Driving the Future of Business*. 2008.

S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36:167–171, 1980.

S. L. Hui and X. H. Zhou. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research*, 7:354–370, 1998.

R. Jin and Z. Ghahramani. Learning with multiple labels. In *Advances in Neural Information Processing Systems 15*, pages 897–904. 2003.

B. Krishnapuram, J. Stoeckel, V. C. Raykar, R. B. Rao, P. Bamberger, E. Ratner, N. Merlet, I. Stainvas, M. Abramov, and A. Manevitch. Multiple-instance learning improves CAD detection of masses in digital mammography. In *IWDM 2008: Proceedings of the 9th international workshop on Digital Mammography*, pages 350–357. 2008.

G. Lugosi. Learning with an unreliable teacher. *Pattern Recognition*, 25(1):79–87, 1992.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

V. C. Raykar, S. Yu, L .H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *ICML 2009: Proceedings of the 26th International Conference on Machine Learning*, pages 889–896, 2009.

V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622, 2008.

P. Smyth. Learning with probabilistic supervision. In *Computational Learning Theory and Natural Learning Systems 3*, pages 163–182. MIT Press, 1995.

P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems 7*, pages 1085–1092. 1995.

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast - But is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *Proceedings of the First IEEE Workshop on Internet Vision at CVPR 08*, pages 1–8, 2008.

J. S. Uebersax and W. M. Grove. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, 49:823–835, 1993.

S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pages 2035–2043. 2009.