Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit

Vikas C. Raykar[1] VIKAS.RAYKAR@SIEMENS.COM
Shipeng Yu[1] SHIPENG.YU@SIEMENS.COM
Linda H. Zhao[2] LZHAO@WHARTON.UPENN.EDU
Anna Jerebko[1] ANNA.JEREBKO@SIEMENS.COM
Charles Florin[1] CHARLES.FLORIN@SIEMENS.COM
Gerardo Hermosillo Valadez[1] GERARDO.HERMOSILLOVALADEZ@SIEMENS.COM
Luca Bogoni[1] LUCA.BOGONI@SIEMENS.COM
Linda Moy[3] LINDA.MOY@NYUMC.ORG

[1]CAD and Knowledge Solutions (IKM CKS), Siemens Healthcare, Malvern, PA 19355 USA
[2]Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104 USA
[3]Department of Radiology, New York University School of Medicine, New York, NY 10016 USA

Abstract

We describe a probabilistic approach for supervised learning when we have multiple experts/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.

1. Introduction

A typical two-class supervised classification scenario consists of a training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ containing N instances, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is an instance (the d-dimensional feature vector) and $y_i \in \mathcal{Y} = \{0, 1\}$ is the corresponding known class label. The task is to learn a classification function $f : \mathbb{R}^d \rightarrow \mathcal{Y}$ which generalizes well on unseen data.

However, for many tasks, it may not be possible, or may be too expensive to acquire the actual label y_i (*gold standard*) for training. 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 might be a substantial amount of disagreement among the experts, and hence it is of great practical interest to determine the optimal way to learn a classifier.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

Our motivation for this work comes from the area of computer-aided diagnosis (CAD) (see § 7.1), where the task is to build a classifier to predict whether a suspicious region on a medical image is malignant 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 be obtained from biopsies, but since it is an expensive and an invasive 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.

The domain of text classification offers another 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 services like Amazon's Mechanical Turk, it is quite inexpensive to acquire labels from a large number of annotators in a short time (Sheng et al., 2008; Snow et al., 2008; Sorokin & Forsyth, 2008). In situations like these, the performance of different annotators can vary widely, and without the actual gold standard, it may not be possible to evaluate the annotators.

In this work, we provide some principled probabilistic solutions to the following question: "How do we *learn* and *evaluate* classifiers when we have multiple annotators providing labels but no absolute gold standard?" A closely related problem–particularly relevant when there are a large number of annotators– is to estimate how reliable/trustworthy is each annotator.

1.1. Majority Voting

For binary classification problems, a common strategy is to use the majority label, *i.e.*,

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

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, *i.e.*,

$$\Pr[y_i = 1|y_i^1, \ldots, y_i^R] = (1/R)\sum_{j=1}^{R} y_i^j. \qquad (2)$$

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

To address the apparent chicken-and-egg problem, we present a maximum-likelihood estimator (§ 4) that *jointly* learns the classifier, the annotator accuracy, and the actual true label. The performance of each annotator is measured in terms of the sensitivity and specificity with respect to the gold standard (§ 2). The proposed algorithm automatically discovers out 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 (§ 5). The final estimate is an 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 (§ 3), it is quite general, and can be used with any black-box classifier (§ 6), and can also handle missing labels (*i.e.*, each expert is not required to label all the instances). Furthermore, it can be extended to handle categorical, ordinal, and regression problems (§ 8). We extensively validate our approach using both simulated data and real data (§ 7).

2. 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.

$$\alpha^j := \Pr[y^j = 1|y = 1]. \qquad (3)$$

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.

$$\beta^j := \Pr[y^j = 0|y = 0]. \qquad (4)$$

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. [1]

3. 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 parameter* θ determines the operating point of the classifier. The ROC curve is obtained as γ is swept from $-\infty$ to ∞. The posterior probability for the positive class is modeled as a *logistic sigmoid* acting on $f_{\boldsymbol{w}}$, *i.e.*,

$$\Pr[y = 1|\boldsymbol{x}, \boldsymbol{w}] = \sigma(\boldsymbol{w}^\top \boldsymbol{x}), \qquad (5)$$

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

Given the training data $\mathcal{D}$ consisting of N instances with annotations from R experts, *i.e.*, $\mathcal{D} = \{\boldsymbol{x}_i, y_i^1, \ldots, y_i^R\}_{i=1}^N$, the task is to estimate the weight vector $\boldsymbol{w}$ and also the sensitivity $\boldsymbol{\alpha} = [\alpha^1, \ldots, \alpha^R]$ and the specificity $\boldsymbol{\beta} = [\beta^1, \ldots, \beta^R]$.

[1] 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.

4. Maximum likelihood estimator

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

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} \Pr[y_i^1, \ldots, y_i^R | \boldsymbol{x}_i, \theta]. \tag{6}$$

Conditioning on the true label y_i, and also using the assumption that α^j and β^j do not depend on the instance $\boldsymbol{x}_i$, the likelihood can be written as

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} \Big\{ \Pr[y_i^1, \ldots, y_i^R | y_i = 1, \boldsymbol{\alpha}] \cdot \Pr[y_i = 1 | \boldsymbol{x_i}, \boldsymbol{w}] + \Pr[y_i^1, \ldots, y_i^R | y_i = 0, \boldsymbol{\beta}] \cdot \Pr[y_i = 0 | \boldsymbol{x_i}, \boldsymbol{w}] \Big\}. \tag{7}$$

Given the true label y_i, we assume that $y_i^1, \ldots, y_i^R$ are independent, *i.e.*, the annotators make their decisions independently. Hence,

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

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

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} \big[a_i p_i + b_i (1 - p_i) \big], \tag{8}$$

where we define $p_i = \Pr[y_i = 1 | \boldsymbol{x_i}, \boldsymbol{w}] = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)$, and

$$a_i = \prod_{j=1}^{R} [\alpha^j]^{y_i^j} [1 - \alpha^j]^{1 - y_i^j}, \quad b_i = \prod_{j=1}^{R} [\beta^j]^{1 - y_i^j} [1 - \beta^j]^{y_i^j}.$$

The maximum-likelihood estimator is found by maximizing the log-likelihood, *i.e.*,

$$\hat{\theta}_{\mathrm{ML}} = \{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}}\} = \arg\max_{\theta} \{\ln \Pr[\mathcal{D}|\theta]\}. \tag{9}$$

4.1. 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. Let $\boldsymbol{y} = [y_1, \ldots, y_N]$, be the complete data log-likelihood can be written as

$$\ln \Pr[\mathcal{D}, \boldsymbol{y}|\theta] = \sum_{i=1}^{N} y_i \ln p_i a_i + (1 - y_i) \ln (1 - p_i) b_i. \tag{10}$$

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

$$\mathbb{E}\left\{\ln \Pr[\mathcal{D}, \boldsymbol{y}|\theta]\right\} = \sum_{i=1}^{N} \mu_i \ln p_i a_i + (1 - \mu_i) \ln (1 - p_i) b_i, \tag{11}$$

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

$$\begin{aligned} \mu_i &\propto \Pr[y_i^1, \ldots, y_i^R | y_i = 1, \theta] \cdot \Pr[y_i = 1 | \boldsymbol{x_i}, \theta] \\ &= \frac{a_i p_i}{a_i p_i + b_i (1 - p_i)}. \end{aligned} \tag{12}$$

(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 (11) 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 $\boldsymbol{g}(w) = \sum_{i=1}^{N} \big[\mu_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\big] \boldsymbol{x}_i$. The Hessian matrix is given by $\boldsymbol{H}(w) = -\sum_{i=1}^{N} \big[\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\big] \big[1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\big] \boldsymbol{x}_i \boldsymbol{x}_i^\top$. Essentially, we are solving *a regular logistic regression problem with probabilistic labels* μ_i.

These two steps (the E- and the M-step) can be iterated till convergence. 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.

5. A Bayesian approach

In some applications we may want to trust a particular expert more than the others. This can be done 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. For any $a > 0$, $b > 0$, and $\delta \in [0, 1]$ the beta distribution is given by

$$\text{Beta}(\delta|a, b) = \frac{\delta^{a-1}(1-\delta)^{b-1}}{\text{B}(a, b)}, \tag{13}$$

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 for both the sensitivity and the specificity as $\Pr[\alpha_j|a_1^j, a_2^j] = \text{Beta}(\alpha_j|a_1^j, a_2^j)$ and $\Pr[\beta_j|b_1^j, b_2^j] = \text{Beta}(\beta_j|b_1^j, b_2^j)$. For sake of completeness we also assume a zero mean Gaussian prior on the weights $\boldsymbol{w}$ with inverse covariance matrix $\boldsymbol{\Gamma}$, *i.e.*, $\Pr[\boldsymbol{w}] = \mathcal{N}(\boldsymbol{w}|0, \boldsymbol{\Gamma}^{-1})$.

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, *i.e.*, $\hat{\theta}_{\text{MAP}} = \arg\max_\theta \{\ln \Pr[\mathcal{D}|\theta] + \ln \Pr[\theta]\}$. An EM algorithm can be derived in a similar fashion for MAP estimation by relying on the interpretation of (Neal & Hinton, 1998).

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

(2) Given μ_i, estimate the sensitivity and specificity of each annotator as follows.

$$\begin{aligned} \alpha^j &= \frac{a_1^j - 1 + \sum_{i=1}^N \mu_i y_i^j}{a_1^j + a_2^j - 2 + \sum_{i=1}^N \mu_i}. \\ \beta^j &= \frac{b_1^j - 1 + \sum_{i=1}^N (1-\mu_i)(1-y_i^j)}{b_1^j + b_2^j - 2 + \sum_{i=1}^N (1-\mu_i)}. \end{aligned} \tag{14}$$

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

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

$$\mu_i = \frac{a_i p_i}{a_i p_i + b_i(1-p_i)}, \tag{15}$$

where $p_i = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)$, $a_i = \prod_{j=1}^R [\alpha^j]^{y_i^j}[1-\alpha^j]^{1-y_i^j}$, and $b_i = \prod_{j=1}^R [\beta^j]^{1-y_i^j}[1-\beta^j]^{y_i^j}$.

Iterate (2) and (3) till convergence.

6. Discussions

6.1. Obtaining actual ground truth

The value of the posterior probability μ_i is a soft probabilistic estimate of the actual ground truth y_i, *i.e.*, $\mu_i = \Pr[y_i = 1|y_i^1, \ldots, y_i^R, \boldsymbol{x}_i, \theta]$. The actual hidden label y_i can be estimated by applying a threshold on μ_i, *i.e.*, $y_i = 1$ if $\mu_i \geq \gamma$ and zero otherwise. We can use $\gamma = 0.5$ as the threshold. By varying γ we can change the miss-classification costs.

6.2. Insight of the proposed framework

A particularly revealing insight can be obtained in terms of the log-odds or the *logit* of the posterior probability μ_i. From (15) the logit of μ_i can be written as

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

where $b = \sum_{j=1}^R \log \frac{1-\alpha^j}{\beta^j}$ 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.

6.3. 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.

6.4. 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, *i.e.*, $p = \Pr[y_i = 1]$. We further assume a beta prior for the prevalence, *i.e.*, $\Pr[p|p_1, p_2] = \text{Beta}(p|p_1, p_2)$.

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

(2) Given μ_i, estimate the sensitivity and specificity of each annotator using (14). The prevalence of the positive class is estimated as follows.

$$p = \frac{p_1 - 1 + \sum_{i=1}^N \mu_i}{p_1 + p_2 - 2 + N}. \tag{16}$$

(3) Given the sensitivity and specificity of each anno-

tator and prevalence, refine μ_i as follows.

$$\mu_i = \frac{a_i p}{a_i p + b_i(1-p)}. \quad (17)$$

Iterate (2) and (3) till convergence. This algorithm is similar to the one proposed in (Dawid & Skeene, 1979; Smyth et al., 1995).

6.5. Handling missing labels

The proposed approach can easily handle missing labels. 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 (14), and replace R by R_i for updating μ_i in (15).

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

$$\alpha = \frac{\sum_{i=1}^{N} \mu_i z_i}{\sum_{i=1}^{N} \mu_i}, \quad \beta = \frac{\sum_{i=1}^{N} (1-\mu_i)(1-z_i)}{\sum_{i=1}^{N} (1-\mu_i)}. \quad (18)$$

7. Experiments

We use two CAD and one text dataset in our experiments. The CAD datasets include a digital mammography data and a Breast MRI data, both of which are biopsy proven, *i.e.*, the gold standard is available. For the digital mammography dataset 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.

7.1. Digital Mammography

Mammograms are used as a screening tool to detect early breast cancer. CAD systems search for abnormal areas (*lesions*) in a digitized mammographic image. In classification terms, given a set of descriptive morphological features for a region in 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. We use a proprietary biopsy-proven dataset 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 the multiple radiologists according to the two-coin model described in § 2. 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 α = [0.90 0.80 0.57 0.60 0.55] and specificity β = [0.95 0.85 0.62 0.65 0.58]. In this scenario the first two radiologists are experts and the last three are novices. The results are as follows:

(1) **Classifier performance** Figure 1(Left) plots the ROC of the classifier on the training set. The dotted (black) line is the ROC for the classifier learnt using the actual ground truth. The solid (red) line is the ROC for the proposed algorithm and the dashed (blue) line is for the majority-voting scheme. The classifier learnt using the proposed method is as good as the one learnt using the golden ground truth. The 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(Right). 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. [2] The end of the dashed red 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(Right) 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). The proposed algorithm appropriately weights each radiologist based on their estimated sensitivity and specificity.

[2] At the end of each EM iteration, a good approximation to the posterior distribution can be obtained as $\alpha_j \sim \text{Beta}\left(\alpha_j | a_1^j + \sum_{i=1}^{N} \mu_i y_i^j, a_2^j + \sum_{i=1}^{N} \mu_i (1-y_i^j)\right)$ and $\beta_j \sim \text{Beta}\left(\beta_j | b_1^j + \sum_{i=1}^{N} (1-\mu_i)(1-y_i^j), b_2^j + \sum_{i=1}^{N} (1-\mu_i) y_i^j\right)$.

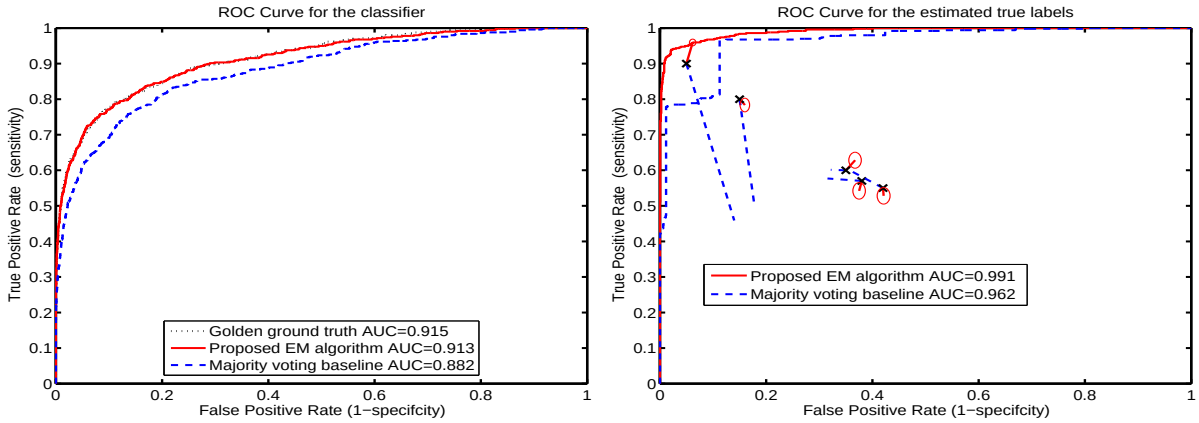

Figure 1. Results for the digital mammography dataset with annotations from 5 simulated radiologists. (Left) The ROC curve of the learnt classifier using the golden ground truth (dotted black line), the majority voting scheme (dashed blue line), and the proposed EM algorithm (solid red line). (Right) The ROC curve for the estimated ground truth. The actual sensitivity and specificity of each of the radiologist 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.

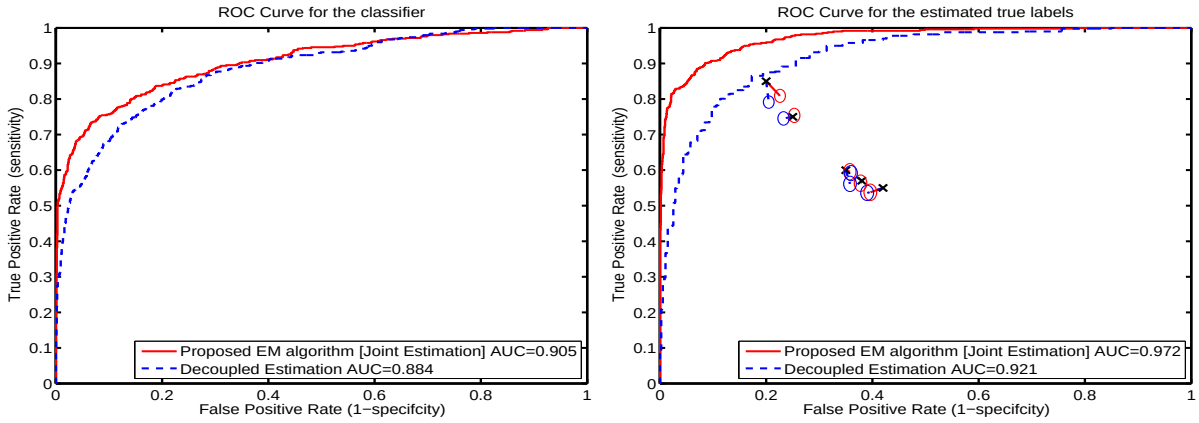

Figure 2. ROC curves comparing the proposed algorithm (solid red line) with 'Decoupled Estimation' procedure (dotted blue line), which refers to the algorithm where the ground truth is first estimated using just the lables from the five radiologists (§ 6.4) and then a logistic regression classifier is 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.

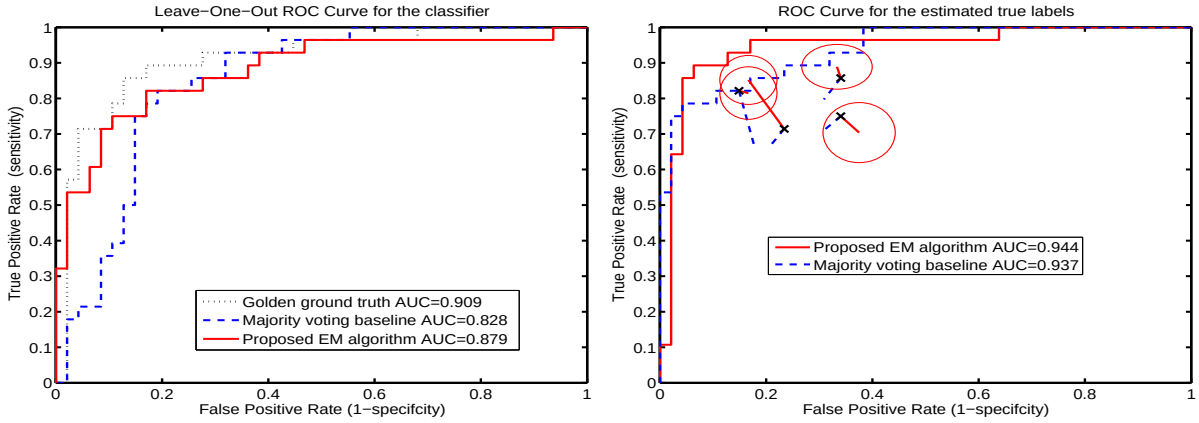

Figure 3. Breast MRI results. (Left) The leave-one-out cross validated ROC. (Right) ROC for the estimated ground truth.

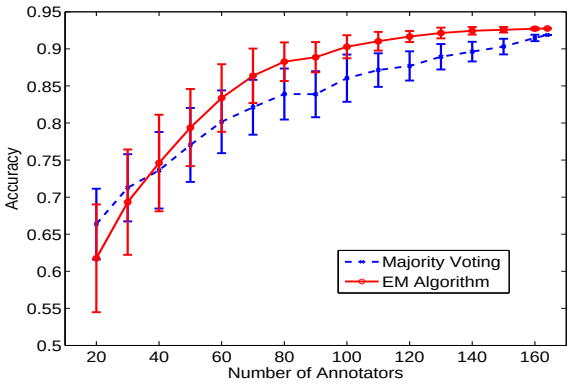

Figure 4. The mean and the one standard deviation error bars for the accuracy of the estimated ground truth for the Recognizing Textual Entailment task as a function of the number of annotators. The plot was generated by randomly sampling R annotators 100 times.

(4) **Joint Estimation** To learn a classifier, Smyth (1995) proposed to first estimate the golden ground truth (§ 6.4) 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 2 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.

7.2. 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. Our dataset comprises of 75 lesions with annotations from four radiologists, and the true labels from biopsy. Based on eight morphological features, we predict whether a lesion is malignant. 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 3 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.

7.3. 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. The majority of the entries (94 %) in the 800x164 matrix are missing. Figure 4 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.

8. Extensions

We briefly describe how the proposed approach can be extended to categorical, ordinal, and continuous data.

8.1. Categorical labels

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. We can extend the previous model and introduce a multinomial parameter $\boldsymbol{\alpha}_c^j = (\alpha_{c1}^j, \ldots, \alpha_{cK}^j)$ 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.

8.2. Ordinal labels

In some situations, the outputs have an ordering among the 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 & Hall, 2001). Specifically the K class ordinal labels are transformed into $K - 1$ binary class labels as follows: For $c = 1, \ldots, K-1$, $y_i^{jc} = 1$ if $y_i^j > c$ and 0 otherwise. 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 $\Pr[y_i = c] = \Pr[y_i > c - 1] - \Pr[y_i > c]$.

8.3. Continuous labels

As a part of the annotation process a task for a radiologist is also to measure the diameter of a nodule. This constitutes an example where the labels are real numbers. This situation can be handled as follows: Let $y_i^j \in \mathbb{R}$ be the target value assigned to the i^{th} instance by the j^{th} annotator. The annotator provides a noisy version of the actual value y_i. We will assume a Gaussian noise model with mean y_i and inverse-variance (precision) τ^j, *i.e.*, $\Pr[y_i^j | y_i, \tau^j] \sim \mathcal{N}(y_i^j | y_i, 1/\tau^j)$. The unknown precision τ^j is a measure of the accuracy

of each annotator. A similar EM algorithm can be derived– (1) Given y_i learn a regression function and estimate the precision for each annotator. (2) Given the precision for each annotator refine y_i.

9. Related work

There has been some work in the biostatistics community–see (Dawid & Skeene, 1979; Hui & Zhou, 1998) and references therein–on latent variable models where the task is to get an estimate of the error rates based on the results from multiple diagnostic tests without a gold standard. In the machine learning community (Smyth et al., 1995) first addressed the same problem in the context of labeling volcanoes. There has been recent interest in the natural language processing (Sheng et al., 2008; Snow et al., 2008) and computer vision (Sorokin & Forsyth, 2008) communities where they show that using annotations from many people can be potentially as good as that provided by an expert. There is also some theoretical work (see (Lugosi, 1992) and reference therein) dealing with multiple experts.

We differ from the previous body of work in the following aspects–(1) Unlike (Dawid & Skeene, 1979; Smyth et al., 1995) which just focused on estimating the ground truth, we specifically address the issue of *learning a classifier*. Estimating the ground truth and the expert/classifer performance is a byproduct of our proposed algorithm. (2) To learn a classifier (Smyth, 1995) propose to first estimate the 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. Our experiments (see Figure 2) 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, ordinal, and continuous data. It can also used in conjunction with any supervised learning algorithm.

10. Conclusions and future work

In this paper we proposed a Bayesian framework for supervised learning in the presence of 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. The proposed algorithm can handle binary/categorical/ordinal classification and regression. We made two key assumptions–(1) the experts performance does not depend on the feature vector and (2) the experts are independent. We are currently exploring strategies to relax these two assumptions.

References

Dawid, A. P., & Skeene, A. M. (1979). Maximum likeihood estimation of observed error-rates using the EM algorithm. *Applied Statistics*, *28*, 20–28.

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

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

Hui, S. L., & Zhou, X. H. (1998). Evaluation of diagnostic tests without a gold standard. *Statistical Methods in Medical Research*, *7*, 354–370.

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

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

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

Smyth, P. (1995). Learning with probabilistic supervision. In *Computational learning theory and natural learning systems 3*, 163–182. MIT Press.

Smyth, P., Fayyad, U., Burl, M., Perona, P., & Baldi, P. (1995). Inferring ground truth from subjective labelling of venus images. In *Advances in neural information processing systems 7*, 1085–1092.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 254–263).

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