
Modeling annotator expertise: Learning when everybody knows a bit of something

Yan Yan¹
Gerardo Hermosillo²
¹Northeastern Univ.
Boston, MA USA

Rómer Rosales²
Luca Bogoni²
²Siemens Healthcare
Malvern, PA USA

Glenn Fung²
Linda Moy⁴
³Univ. of British Columbia
Vancouver, BC Canada

Mark Schmidt³
Jennifer G. Dy¹
⁴New York Univ.
New York, NY USA

Abstract

Supervised learning from multiple labeling sources is an increasingly important problem in machine learning and data mining. This paper develops a probabilistic approach to this problem when annotators may be unreliable (labels are noisy), but also their expertise varies depending on the data they observe (annotators may have knowledge about different parts of the input space). That is, an annotator may not be consistently accurate (or inaccurate) across the task domain. The presented approach produces classification and annotator models that allow us to provide estimates of the true labels and annotator variable expertise. We provide an analysis of the proposed model under various scenarios and show experimentally that annotator expertise can indeed vary in real tasks and that the presented approach provides clear advantages over previously introduced multi-annotator methods, which only consider general annotator characteristics.

1 Introduction

The ease with which data can be shared, organized, and processed by a large number of entities using standard communication infrastructures (such as the Internet) is creating a number of interesting problems and opportunities for machine learning and data modeling in general. One of the main ramifications is that the knowledge from these different entities, in particular people, can now be easily collected and compounded

in a distributed fashion. There are numerous examples of this effect, the classical example being open source (*e.g.*, Linux), and more recently Wikipedia. However, combining the knowledge from different sources is far from being a solved problem. In this paper, we concentrate on efficiently utilizing the type of knowledge provided by different annotators (labelers).

Supervised learning traditionally relies on a domain expert playing the role of a *teacher* providing the necessary supervision. The most common case is that of an expert providing annotations that serve as data point labels in classification problems. The above *crowdsourcing* effect (Howe, 2008) motivates a natural shift from the traditional reliance on a single domain expert to several domain experts or even many more non-experts who contribute to a specific (learning) task. In supervised learning, more labeled data for training normally translate, under some assumptions, to higher test time accuracy. What do more labelers/annotators translate to and how can their knowledge be efficiently utilized? This paper tackles these problems in the context of learning from multiple annotators.

The availability of more annotators is not the only motivation for learning from multiple labelers. In many application areas, there are problems for which obtaining the ground-truth labels is simply impossible or very costly. For example, in cancer detection from medical images (*e.g.*, computer tomography, magnetic resonance imaging), an image region or volume associated to a body tissue can often be tested for the actual presence of cancer only by performing a biopsy; this is clearly a costly, risky procedure. In addition, many other annotation tasks are subjective by nature and thus there is no clear correct label. Almost any subjective opinion task falls in this category, such as the task of sentiment classification, product ratings from text, or lesion severity judgment from medical images.

In multi-labeler problems, building a classifier in the traditional single expert manner, without regard for the label source (annotator) properties may not be ef-

Appearing in Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

fective in general. The reasons for this include: some annotators may be more reliable than others, some may be malicious, some may be correlated with others, there may exist different prior knowledge about annotators, and in particular annotator effectiveness may vary depending on the data instance presented. We believe this last element is of great importance and has not been clearly considered in previous approaches.

1.1 Related Work

The problem of modeling data that has been processed by multiple annotators has been receiving increasing attention. However, similar problems have been studied for quite some time. For example, in clinical statistics, Dawid and Skeene (1979) studied the problem of error rate estimation given repeated but conflicting responses (labels) of patients to various medical questions. In this work, a point estimate of individual error rates was identified using latent variable models. Later, Spiegelhalter and Stovin (1983) used this model to quantify residual uncertainty of the label value.

We can divide the related work in multi-labeler classification in various sub-areas. One area of work consists on the estimation of error rates for the labelers independently from building a classifier. Both early works above (Dawid and Skeene, 1979; Spiegelhalter and Stovin, 1983) and others such as Hui and Zhou (1998), fall in this area, while more recently Snow et al. (2008) showed that employing multiple non-expert annotators can be as effective as employing one expert annotator when building a classifier.

Very recently, the interest has shifted towards more directly building classifiers from multi-labeler data. In this area, we can further subdivide the approaches into those attempting to use repeated labeling or prior knowledge about labeler similarities. Repeated labeling (Smyth et al., 1995; Donmez and Carbonell, 2008; Sheng et al., 2008) relies on the identification of what labels should be reacquired in order to improve classification performance or data quality. This form of active learning can be well suited when we can control assignments of data points to labelers. However, Dekel and Shamir (2009) provided arguments indicating that this approach *is wasteful* and negatively impacts the relative size of the training set. Approaches based on prior knowledge rely on the existence of some way to measure labeler relationships. These include the work of Crammer et al. (2008), where labeler similarities and their labels are used to identify what samples should be used to estimate classification models for each labeler, and Blitzer et al. (2007) where the multiple labels are obtained by labeling data drawn from multiple underlying domains (in the context of domain adaptation).

Application areas for multi-labeler learning vary widely. These include natural language processing (Snow et al., 2008), computer-aided diagnosis/radiology (Raykar et al., 2009; Spiegelhalter and Stovin, 1983), clinical data integration (Dawid and Skeene, 1979), and computer vision (Sorokin and Forsyth, 2008).

This paper differs from the related work in various axes. Unlike Dawid and Skeene (1979) and Spiegelhalter and Stovin (1983), we produce labeler error estimates and simultaneously build a classifier in a combined process. In contrast to Smyth et al. (1995) and Sheng et al. (2008), we do not assume that labels can be reacquired (the active learning setting). Also, we do not assume the existence of any prior information relating the different labelers or the domains from where the data is drawn, such as Crammer et al. (2008) or Blitzer et al. (2007). Like the approach presented by Raykar et al. (2009) and to some extent that by Jin and Ghahramani (2003), this paper estimates the error rates and the classifier simultaneously; however, unlike both approaches this paper models the error rates of the labelers as dependent on the data points.

A distinguishing factor in this paper is that, unlike previous approaches, it is not assumed that expert reliability or error rate is consistent across all the input data even for one task. This is a flawed assumption in many cases since annotator knowledge can fluctuate considerably depending on the input instance. In this paper, the classifiers are built so that they take into account that some labelers are better at labeling some types of points (compared with other data points).

2 Formulation

Given N data points $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^D$, each labeled by at most T labelers/annotators. We denote the label for the i -th data point given by annotator t as $y_i^{(t)} \in \mathcal{Y}$. The labels from individual labelers may not be correct. Let us denote the true (unknown) label for the i -th data point to be $z_i \in \mathcal{Z}$ (normally $\mathcal{Y} \equiv \mathcal{Z}$). For compactness, we set the matrices $X = [\mathbf{x}_1^T; \dots; \mathbf{x}_N^T] \in \mathbb{R}^{N \times D}$ and $Y = [y_1^{(1)}, \dots, y_1^{(T)}; \dots; y_N^{(1)}, \dots, y_N^{(T)}] \in \mathbb{R}^{N \times T}$, where $(\cdot)^T$ stands for matrix transpose. Given training data, X and Y , our goals are: to produce an estimate for the ground-truth $Z = [z_1, \dots, z_N]^T$, a classifier for predicting the label z for new instances \mathbf{x} , and a model of the annotators' expertise as a function of the input \mathbf{x} .

2.1 Probabilistic Model

Let us define the random variables $y^{(t)}$ over the space of labels \mathcal{Y} , provided by labeler t , for $t = \{1, \dots, T\}$.

Similarly, let us define the random variables $\mathbf{x} \in \mathcal{X}$ and $z \in \mathcal{Z}$ to represent input data points (observed) and unknown output respectively. We build our classifier by assuming a probabilistic model over random variables \mathbf{x} , y , and z with a graphical model as shown in Figure 1. The joint conditional distribution can be expressed as:

$$p(Y, Z|X) = \prod_i p(z_i|\mathbf{x}_i) \prod_t p(y_i^{(t)}|\mathbf{x}_i, z_i).$$

In this model, the annotation provided by labeler t depends both on the unknown true label z but also on the (normally) observed input \mathbf{x} . In other words, we do not assume that annotators are equally good (or bad) at labeling all the data, but it depends on what input they observe. As can be seen from the model, we make the assumption that the labelers $t = \{1, \dots, T\}$ are independent given the input and the true label. In order to further specify our model we need to define the form of the conditional probabilities. In this paper, we explored several variations. Let us consider each conditional distribution in turn.

$p(y_i^{(t)}|\mathbf{x}_i, z_i)$: Our simplest model assumes that each annotator t provides a noisy version of the true label z , $p(y_i^{(t)}|\mathbf{x}_i, z_i) = p(y_i^{(t)}|z_i) = (1 - \eta^{(t)})^{|y_i^{(t)} - z_i|} \eta^{(t)1 - |y_i^{(t)} - z_i|}$, with $\mathcal{Z} \equiv \mathcal{Y} = \{0, 1\}$. In this Bernoulli model, the parameter $\eta^{(t)}$ is the probability of labeler t to be correct (i.e., $y_i = z_i$). Another option we consider is the Gaussian model, where every labeler is expected to provide a distorted version of the true label z , $p(y_i^{(t)}|z_i) = \mathcal{N}(y_i^{(t)}; z_i, \sigma^{(t)})$. This Gaussian distribution associates a lower variance $\sigma^{(t)}$ to more *consistently* correct labelers compared to inconsistent labelers. Note that we employ a distribution for continuous random variables, which is more natural for regression rather than classification models (for y continuous). In these models, where we assume that $p(y_i^{(t)}|\mathbf{x}_i, z_i) = p(y_i^{(t)}|z_i)$, the additional independence assumptions mean that the graphical model is Markov-equivalent to the model $\mathbf{x} \rightarrow z \rightarrow \{y^{(t)}\}$. This is comparable to the models proposed by Raykar et al. (2009) and Jin and Ghahramani (2003), albeit with different parameterizations.

We use these models as a base for considering more general cases, where $p(y|\mathbf{x}, z) \neq p(y|z)$. In our experience with real applications, we noticed that the quality of labels by annotators is not only a function of their expert level, but also of the type of data presented to them as well. For example, radiologists will have difficulty providing quality labels on blurry images. Additionally, some labelers will be more affected by blurry images than others and moreover some labelers are more knowledgeable for some input types than

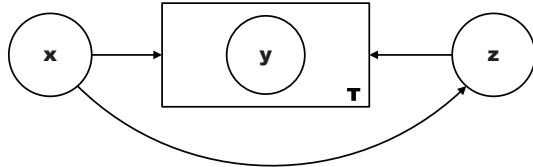


Figure 1: Graphical Model for \mathbf{x} , \mathbf{y} , and z .

others. In general, annotators will exhibit varying levels of expertise in different types of data. We believe this is particularly true for non-experts annotators.

In order to model this input dependent variability, we propose to replace our basic Gaussian model with the following:

$$p(y_i^{(t)}|\mathbf{x}_i, z_i) = \mathcal{N}(y_i^{(t)}; z_i, \sigma_t(\mathbf{x}_i)), \quad (1)$$

where the variance now depends on the input \mathbf{x} and is also specific to each annotator t .

Since the value of $y^{(t)}$ can only take the binary values 0/1, instead of allowing $\sigma_t(\mathbf{x})$ to be any value, we constrain it to be in the range between (0, 1] by setting $\sigma_t(\mathbf{x})$ as a logistic function of \mathbf{x}_i and t :

$$\sigma_t(\mathbf{x}) = (1 + \exp(-\mathbf{w}_t^T \mathbf{x}_i - \gamma_t))^{-1} \quad (2)$$

To make sure that $\sigma_t(\mathbf{x})$ does not go to zero, we added a small constant in our experiments. Similarly, we modify our Bernoulli model by setting $\eta_t(\mathbf{x})$ to be also now a function of both \mathbf{x}_i and t :

$$p(y_i^{(t)}|\mathbf{x}_i, z_i) = (1 - \eta_t(\mathbf{x}))^{|y_i^{(t)} - z_i|} \eta_t(\mathbf{x})^{1 - |y_i^{(t)} - z_i|} \quad (3)$$

And, we also set $\eta_t(\mathbf{x})$ to be a logistic function:

$$\eta_t(\mathbf{x}) = (1 + \exp(-\mathbf{w}_t^T \mathbf{x}_i - \gamma_t))^{-1} \quad (4)$$

$p(z_i|\mathbf{x}_i)$: One can set $p(z_i|\mathbf{x}_i)$ to be any distribution or in our case classifier $g : \mathcal{X} \rightarrow \mathcal{Z}$, which maps \mathbf{x} to z . In this paper we do not intend to demonstrate the advantages of different choices for $p(z_i|\mathbf{x}_i)$. For simplicity, we set $p(z_i|\mathbf{x}_i)$ to be the logistic regression model:

$$p(z_i = 1|\mathbf{x}_i) = (1 + \exp(-\alpha^T \mathbf{x}_i - \beta))^{-1}. \quad (5)$$

In the above case, the classification problem is assumed binary, but one can easily extend this to multiple classes, *e.g.*, using multiple logistic regression.

2.2 Maximum Likelihood Estimation

Given our model, we estimate the set of all parameters, $\theta = \{\alpha, \beta, \{\mathbf{w}_t\}, \{\gamma_t\}\}$, by maximizing the likelihood function. Equivalently

$$\arg \max_{\theta} \prod_t \prod_i p(y_i^{(t)}|\mathbf{x}_i; \theta), \quad (6)$$

which becomes the following problem after taking the logarithm and including the ground-truth variable z :

$$= \arg \max_{\theta} \sum_t \sum_i \log \sum_{z_i} p(y_i^{(t)}, z_i | \mathbf{x}_i; \theta) \quad (7)$$

Since we have missing variables z , a standard approach to solve our maximum likelihood problem is by employing the expectation maximization (EM) (Dempster et al., 1977) algorithm. We provide the specifics of our problem of interest below.

2.3 Algorithm

E-step: Compute $\tilde{p}(z_i) \triangleq p(z_i | \mathbf{x}_i, y_i)$.

$$\begin{aligned} \tilde{p}(z_i) &\propto p(z_i, y_i | \mathbf{x}_i) \\ &\stackrel{\text{i.d.}}{=} \prod_t p(y_i^{(t)} | \mathbf{x}_i, z_i) p(z_i | \mathbf{x}_i) \end{aligned} \quad (8)$$

M-step: Maximize $\sum_t \sum_i E_{\tilde{p}(z_i)} [\log p(y_i^{(t)}, z_i | \mathbf{x}_i)]$. The difficulty of this optimization depends on the specific form of the conditional probabilities. In the formulations that follow, we show the update equations for the more general case where $\sigma_t(\mathbf{x})$ and $\eta_t(\mathbf{x})$ are both functions of the data \mathbf{x}_i and labeler t . Since, there is no closed-form solution for maximizing $\sum_t \sum_i E_{\tilde{p}(z_i)} [\log p(y_i^{(t)}, z_i | \mathbf{x}_i)]$ with respect to the parameters, we apply the LBFGS quasi-Newton (Nocedal and Wright, 2003) method (that does not require second order information) to solve the following optimization problem:

$$\begin{aligned} \max_{\alpha, \beta, \{\gamma_t\}, \{\mathbf{w}_t\}} f_{\text{opt}}(\alpha, \beta, \{\gamma_t\}, \{\mathbf{w}_t\}) = \\ \max_{\alpha, \beta, \{\gamma_t\}, \{\mathbf{w}_t\}} \sum_{i,t} E_{\tilde{p}(z_i)} [\log p(y_i^{(t)} | \mathbf{x}_i, z_i) + \log p(z_i | \mathbf{x}_i)] \end{aligned}$$

For convenience, we provide the gradients with respect to the different parameters for the two candidate models (Gaussian or Bernoulli) here:

$$\begin{aligned} \frac{\partial f_{\text{opt}}}{\partial \alpha} &\propto \sum_i \frac{\Delta \tilde{p} \exp(-\alpha^T \mathbf{x} - \beta) \mathbf{x}}{(1 + \exp(-\alpha^T \mathbf{x} - \beta))^2} \\ \frac{\partial f_{\text{opt}}}{\partial \beta} &\propto \sum_i \frac{\Delta \tilde{p} \exp(-\alpha^T \mathbf{x} - \beta)}{(1 + \exp(-\alpha^T \mathbf{x} - \beta))^2}, \end{aligned}$$

where $\Delta \tilde{p} = \tilde{p}(z_i = 1) - \tilde{p}(z_i = 0)$. When a Gaussian model is applied for $p(y_i^{(t)} | \mathbf{x}_i, z_i)$:

$$\frac{\partial f_{\text{opt}}}{\partial \sigma_t(\mathbf{x})} = \frac{[y_i^{(t)2} - \tilde{p}(z_i = 1)(2y_i^{(t)} - 1)]}{\sigma_t^3(\mathbf{x})} - \frac{1}{\sigma_t(\mathbf{x})}$$

When a Bernoulli model is applied for $p(y_i^{(t)} | \mathbf{x}_i, z_i)$:

$$\frac{\partial f_{\text{opt}}}{\partial \eta_t(\mathbf{x})} = (-1)^{y_i^{(t)}} (\tilde{p}(z_i = 0) - \tilde{p}(z_i = 1))$$

$$\begin{aligned} \frac{\partial \eta_t(\mathbf{x})}{\partial \mathbf{w}_t} &= \frac{\partial \sigma_t(\mathbf{x})}{\partial \mathbf{w}_t} = \frac{\exp(-\mathbf{w}_t^T \mathbf{x}_i - \gamma_t) \mathbf{x}_i}{(1 + \exp(-\mathbf{w}_t^T \mathbf{x}_i - \gamma_t))^2} \\ &= \sigma_t(\mathbf{x})(1 - \sigma_t(\mathbf{x})) \mathbf{x}_i, \quad \text{for the Gaussian model (9)} \\ &= \eta_t(\mathbf{x})(1 - \eta_t(\mathbf{x})) \mathbf{x}_i, \quad \text{for the Bernoulli model (10)} \\ \frac{\partial \eta_t(\mathbf{x})}{\partial \gamma_t} &= \frac{\partial \sigma_t(\mathbf{x})}{\partial \gamma_t} = \frac{\exp(-\mathbf{w}_t^T \mathbf{x}_i - \gamma_t)}{(1 + \exp(-\mathbf{w}_t^T \mathbf{x}_i - \gamma_t))^2} \\ &= \sigma_t(\mathbf{x})(1 - \sigma_t(\mathbf{x})), \quad \text{for the Gaussian model (11)} \\ &= \eta_t(\mathbf{x})(1 - \eta_t(\mathbf{x})), \quad \text{for the Bernoulli model (12)} \end{aligned}$$

To learn the parameters $\alpha, \beta, \{\gamma_t\}, \{\mathbf{w}_t\}$, and obtain a distribution over the missing variables z_i , we iterate between the **E** and **M** steps until convergence. We summarize our method in Algorithm 1:

Algorithm 1 Probabilistic Multiple Labeler Algorithm

input: X, Y ; set: $\alpha = \mathbf{0}, \beta = 0$ and threshold ϵ
 initialize: $\alpha_{\text{new}}, \beta_{\text{new}}, \mathbf{w}_t$ and γ_t
while $\|\alpha - \alpha_{\text{new}}\|^2 + (\beta - \beta_{\text{new}})^2 \geq \epsilon$ **do**
 E-step: estimating $\tilde{p}(z)$ by using equation (8)
 M-step: updating $\alpha_{\text{new}}, \beta_{\text{new}}, \mathbf{w}_t$ and γ_t that maximize $\sum_t \sum_i E_{\tilde{p}(z_i)} [\log p(y_i^{(t)}, z_i | \mathbf{x}_i)]$ using the LBFGS quasi-Newton approximation to compute the step, with gradient equations (9-12).
end while
return $\alpha, \beta, \{\mathbf{w}_t\}, \{\gamma_t\}$

2.4 Classification

Once the parameters α, β have been estimated in the learning stage, a new data point x can be classified by simply letting $p(z = 1 | x) = (1 + \exp(-\alpha^T \mathbf{x} - \beta))^{-1}$, where $z = 1$ is the the class label of interest.

3 Analysis

In this section, we analyze the resulting classification model. In order to simplify the presentation, we use the set notation $\{y^{(t)}\}$ as a shorthand for $\{y^{(t)}\}_{t=1}^T \triangleq \{y^{(1)}, \dots, y^{(T)}\}$ and $\{y^{(t \setminus k)}\}$ as a shorthand for $\{y^{(t)}\}_{t=1, t \neq k}^T$.

3.1 Classification Model

It may be interesting to ask what the model is actually doing in order to estimate the ground truth from the information provided by all the labelers. One way to answer this question is by analyzing the posterior distribution $p(z | \{y^{(t)}\}, \mathbf{x})$, which is given by:

$$\begin{aligned} p(z | \{y^{(t)}\}, \mathbf{x}) &= p(\{y^{(t)}\} | z, \mathbf{x}) p(z | \mathbf{x}) / p(\{y^{(t)}\} | \mathbf{x}) \\ &= \frac{\prod_t p(y^{(t)} | z, \mathbf{x}) p(z | \mathbf{x})}{\sum_z \prod_t p(y^{(t)} | z, \mathbf{x}) p(z | \mathbf{x})}. \end{aligned} \quad (13)$$

If we consider the log-likelihood ratio $\text{LLR}(\{y^{(t)}\}, \mathbf{x}) = \log \frac{p(z=1|\{y^{(t)}\}, \mathbf{x})}{p(z=0|\{y^{(t)}\}, \mathbf{x})}$ for the Bernoulli case, we obtain:

$$\begin{aligned} \text{LLR} &= \text{logit}[p(z=1|\mathbf{x})] + \sum_t (-1)^{(1-y^{(t)})} \text{logit}[\eta_t(\mathbf{x})] \\ &= \alpha^T \mathbf{x} + \beta + \sum_t (-1)^{(1-y^{(t)})} \mathbf{w}_t^T \mathbf{x} + \gamma_t, \end{aligned} \quad (14)$$

where $\text{logit}(p) = \frac{p}{1-p}$. This provides the insight that the classification boundary depends on a linear combination of a score provided by the learned model with parameters (α, β) and the signed contributions from the T individual annotators. The annotator contributions are given by the annotator specific (linear) model of expertise, weighted positively or negatively depending on the label provided (1 or 0 respectively). Note that with a few notation changes this final form can be written as a logistic regression classifier as well.

For the Gaussian case, the ratio becomes:

$$\begin{aligned} \text{LLR} &= \text{logit}[p(z=1|\mathbf{x})] + \sum_t (-1)^{(1-y^{(t)})} \frac{1}{\sigma_t(\mathbf{x})} \\ &= \alpha^T \mathbf{x} + \beta + T^+ - T^- + \\ &\quad \sum_t (-1)^{(1-y^{(t)})} \exp(-\mathbf{w}_t^T \mathbf{x} - \gamma_t), \end{aligned} \quad (15)$$

where T^+ and T^- are the counts of positive and negative labels respectively. Similarly to the case above, the solution involves a linear combination of scores given by each labeler. In this case the score is calculated using the exponential function.

3.2 Missing Annotators

From Eq. 13 we can derive the posterior when not all the annotators provided a label for a data point by computing the appropriate marginal distributions. If annotator k was missing, one can show that the model provides a simple solution:

$$p(z|\{y^{t \setminus k}\}, \mathbf{x}) = \frac{\prod_{t \setminus k} p(y^{(t)}|z, \mathbf{x}) p(z|\mathbf{x})}{\sum_z \prod_{t \setminus k} p(y^{(t)}|z, \mathbf{x}) p(z|\mathbf{x})}, \quad (16)$$

which basically ignores the missing annotator. This implies the natural result that if all annotators are missing, we obtain Eq. 5.

3.3 Estimating the Ground-Truth without Observing Input Data (\mathbf{x})

The presented model provides an expression for estimating the ground-truth even purely from the observed annotations (when the input data has not been observed).

$$p(z|\{y^{(t)}\}) = \int \prod_t p(y^{(t)}|z, \mathbf{x}) p(z|\mathbf{x}) d\mathbf{x} \quad (17)$$

Since we do have a direct prior $p(x)$, we can rely on sampling. One proposal is to use the previously seen cases (training data) as a good sample for X . Let $\mathcal{X}_S = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_S\}$, a sample from the random variable X . We can use this sample to compute the posterior by:

$$p(z|\{y^{(t)}\}) \approx \frac{1}{S} \sum_{s=1}^S p(z|\mathbf{x}_s) \prod_t p(y^{(t)}|z, \mathbf{x}_s), \quad (18)$$

which can be done easily given a learned model.

3.4 Evaluating Annotators

If we knew the ground-truth (for a particular data point), we can straightforwardly evaluate the annotator accuracy. However, this is not the usual case. What if we do not have the ground-truth (it does not exist or is expensive to obtain)? The proposed approach provides a way to evaluate an annotator even without reliance on ground-truth. We can do this by evaluating the following conditional distribution:

$$\begin{aligned} p(y^{(k)}|\{y^{(t \setminus k)}\}, \mathbf{x}) &= \frac{p(\{y^{(t)}\}|\mathbf{x})}{p(\{y^{(t \setminus k)}\}|\mathbf{x})} \\ &= \frac{\sum_z p(\{y^{(t)}\}|z, \mathbf{x}) p(z|\mathbf{x})}{\sum_z p(\{y^{(t \setminus k)}\}|z, \mathbf{x}) p(z|\mathbf{x})} \end{aligned} \quad (19)$$

Note that if the ground-truth is given (along with the input data), the annotators are mutually independent and $p(y^{(k)}|\{y^{(t \setminus k)}\}, \mathbf{x}) = p(y^{(k)}|z, \mathbf{x})$, as expected.

4 Experiments

In this section, we used several simulated and real datasets to compare the performance of our proposed approach to other baseline and state-of-the-art methods. Our experiments were divided in three parts:

- (I) **Performance simulations on UCI data:** We tested our algorithm on four publicly available datasets from the UCI Machine Learning Repository (Asuncion and Newman, 2007): Ionosphere, Cleveland Heart, Glass, and Housing. Since there are no multiple annotations (labels) for these datasets, we artificially generated 5 simulated labelers with different *labeler expertise* and considered the provided labels as golden ground-truth.
- (II) **Modeling labeler's expertise on a heart motion abnormality detection problem:** In this case we perform experiments based on real cardiac data. This data is related to automatic assessment of heart wall motion abnormalities (Qazi et al., 2007). The purpose of this experiment is to measure how well our model learns the labeler's expertise based on the particular case characteristics (data point features).

(III) **Performance on the breast dataset:** Analogous to (I) but with a real dataset extracted for MR digital mammographies and used for classifying regions of interest in the breast into benign and malignant. The cases are labeled by three expert radiologists based on visual inspection of the images. The golden ground-truth was obtained by performing a biopsy in each case. This is quite a rare opportunity where ground-truth actually exists, in particular in the medical domain.

For our proposed multiple labelers method (**M.L.**), in our comparisons, we considered three different variations that depend on the modeling of $p(y|\mathbf{x}, z)$ as described in Sec. 3. **M.L.-Gaussian(x)** and **M.L.-Bernoulli(x)** will refer to the models that explicitly depend on \mathbf{x} . We will refer as **M.L.-Original** the original formulation that estimates a parameter σ per labeler in the spirit of Raykar et al. (2009) and Jin and Ghahramani (2003). For further comparisons, we also learn two additional logistic regression classifiers, one using the labelers majority vote as target labels for training (**Majority**), and the other one concatenates all the labelers information by repeating training data points as many times as needed to represent all the labelers (**Concatenation**).

For both Part I and II, we randomly divided the data into five equally sized folds (20% of the data each). For each dataset, we repeated the model training five times where we used four of the folds (80% of the data) for training and one fold for testing. For part III we used 40% for training and the remaining 60% for testing.

Part I: Performance simulations on UCI data

We performed experiments on four datasets from the UCI Irvine machine learning data repository (Asuncion and Newman, 2007): Ionosphere (351,34), Cleveland Heart (297,13), Glass (214,9), and Housing (506,13)(with (number of points, number of features) each). Since multiple labels for any of these UCI datasets are not available, we need to simulate several labelers with different *labeler expertise* or accuracy. In order to simulate the labelers, for each dataset, we proceeded as follows: first, we clustered the data into five subsets using k-means (Berkhin, 2002). Then, we assume that each one of the five simulated labelers $i = \{1, \dots, 5\}$ is an expert on cases belonging to cluster i , where their labeling coincides with the ground-truth; for the rest of the cases (cases belonging to the other four clusters), labeler i makes a mistake 35% of the times (we randomly switch labels for 35% of the points). Figures (2) and (3) show the ROC comparisons for different multi-labeler models and baseline logistic regression models for the four datasets.

The experimental results demonstrate the power of

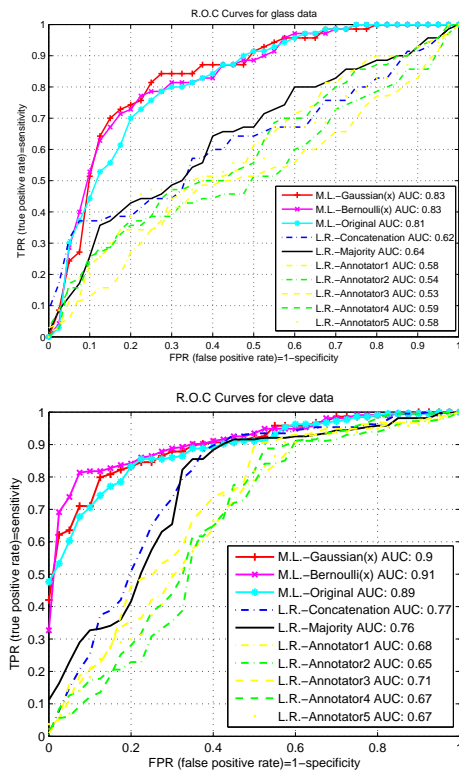


Figure 2: R.O.C. comparison of multi-labeler methods for the UCI datasets Glass and Cleveland.

our proposed approach, we can see that even when our labelers only have slightly better performance than random (around 60% AUC), our probabilistic models can achieve significantly better performance (around 90% AUC). Our models are successfully modeling who is a good labeler for different subsets of training data. Our approaches significantly outperform baseline methods where information from all the labelers is taken into account in a more naive way.

Part II: Modeling labeler’s expertise on the AWMA heart data

The Heart Motion Abnormality Detection data consists of 220 cases for which we have associated images all of which were generated using pharmacological stress. All the cases have been labeled at the heart wall segment level by a group of five trained cardiologists. According to standard protocol, there are 16 LV heart wall segments. Each of the segments were ranked from 1 to 5 according to its movement. For simplicity, we converted the labels to a binary (1 = normal, 2 to 5 = abnormal). For our experiments, we used 24 global and local image features for each node calculated from tracked contours.

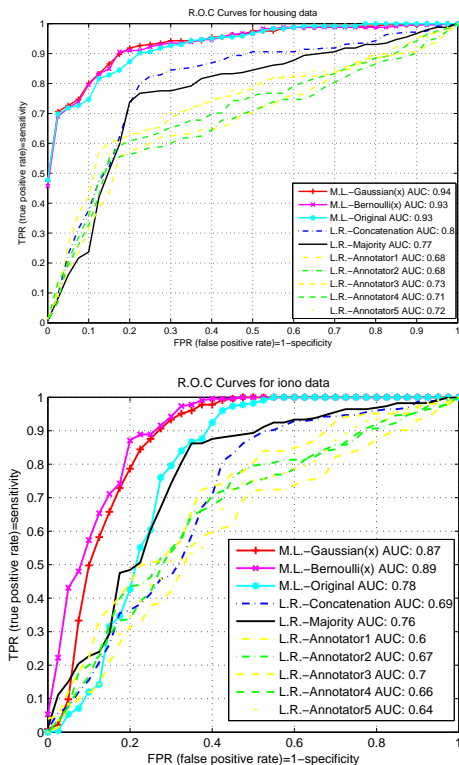


Figure 3: R.O.C. comparison of multi-labeler methods for UCI datasets Ionosphere and Housing.

Since we have 5 doctor labels but no golden ground-truth (biopsy), we will assume that the majority vote of the 5 doctors are a fair approximation to the true labels. For this experiment we proceeded as follows: after training, we used our model to pick the best labeler for each training data point. Then, we trained a simple logistic regression model using the suggested label. We compared our two proposed models (**M.L.-Gaussian(x)** and **M.L.-Bernoulli(x)**) against a baseline model where for each training data point the corresponding labeler is picked randomly among the five available labelers (**Random selection**). Figure 4 shows the corresponding ROCs for this experiment. Note that when using the annotator’s labels suggested by our model, a simple logistic regression method clearly outperforms a model trained using labels coming from a labeler picked at random among the five labels available from the annotators. This model has an interesting potential in a medical setting where annotating cases is expensive. The proposed model can rank experts by case and can help decide which annotator is more apt to label a given new case.

Part III: Performance on the breast dataset

CAD algorithms for mammography are designed to detect suspicious findings in a digitized mammographic

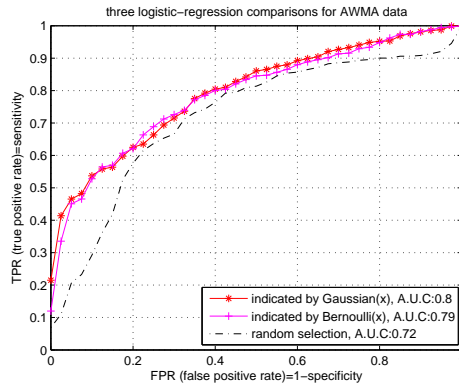


Figure 4: ROCs of the three logistic regression models for the cardiac data: **M.L.-Gaussian(x)**, **M.L.-Bernoulli(x)** and **Random selection**.

image with a high sensitivity. Given a set of descriptive morphological features for a region in an image, the task is to predict whether it is potentially malignant or not. We use a set of mammograms collected from hospitals that generate a biopsy-proven (which provides the golden ground-truth) dataset containing 28 positive and 47 negative examples. Each instance is described by a set of 8 morphological features and labeled independently by three doctors. Results are presented in Figure 5. Note that the results are similar to the ones obtained in part I. Our proposed M.L. methods again significantly outperform the baseline methods and each individual annotator even when using a reduced set of training data (only 40% in this case).

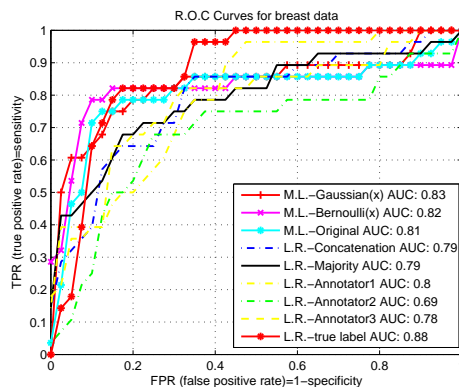


Figure 5: Results for the breast dataset

5 Conclusion

Traditionally, supervised learning relies on a single labeler playing the role of a teacher providing the necessary supervision. However, the increasing availability

of more annotators for certain domains, the difficulty of obtaining ground truth (such as in cancer detection in medical images), and/or the subjectivity of labeling (such as product ratings), lead to the growing importance of studying supervised learning when there are multiple annotators whose labels may be unreliable. A distinguishing factor in this paper, in contrast to previous approaches, is that we do not assume that the reliability of annotators is the same across all data. In many cases, annotator knowledge can fluctuate considerably depending on the specific input instance observed. This is the common case when everyone knows something about the problem domain, but everyone may know different aspects of the same problem (rarely does someone know everything). For example, radiologists specialized in heart images will be better at labeling lesions of the heart compared to radiologists with lung expertise, who on the other hand would label instances of lung diseases better.

In this paper, we developed a probabilistic model for learning a classifier from multiple annotators, where the reliability of the annotators vary on the annotator and the data that they observe. Our approach allows us to provide estimates for the true labels given new instances and also provide the expertise variability for each annotator across the domain task. Our experiments on benchmark and real cardiac and breast cancer data show that the expertise of annotators do vary across data and that our model provides better classification performance over various forms of data preprocessing (majority vote or concatenation of the labels provided by all the annotators), and more importantly improves the results over the model that ignores the effect of variable expertise across instances.

We have further provided an analysis of the proposed approach in terms of the resulting decision boundary properties. We showed how the model is suitable for handling missing annotators, for estimating the ground-truth, and for evaluating annotators when the ground-truth is not available. This was done in the context of statistical inference once the correct conditional distribution of interest is identified.

Acknowledgments

This work is supported by NSF IIS-0915910.

References

- A. Asuncion and D. Newman. UCI machine learning repository, 2007. URL <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- P. Berkhin. Survey of clustering data mining techniques (on-line), 2002. URL <http://www.ee.ucr.edu/~barth/EE242/clustering-survey.pdf>.
- J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Adv. Neural Information Processing Systems*, 2007.
- K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *J. of Machine Learning Research*, 9: 1757–1774, 2008.
- A. P. Dawid and A. M. Skeene. Maximum likelihood estimation of observed error-rates using the EM algorithm. *Applied Statistics*, 28:20–28, 1979.
- O. Dekel and O. Shamir. Good learners for evil teachers. In *Int. Conf. on Machine Learning*, 2009.
- A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data. *J. of the Royal Statistical Society (B)*, 39(1), 1977.
- P. Donmez and J. G. Carbonell. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Conf. on Information and Knowledge Management*, 2008.
- J. Howe. *Crowdsourcing: why the power of the crowd is driving the future of business*. Crown Business, 2008.
- S. L. Hui and X. H. Zhou. Evaluation of diagnostic tests without a gold standard. *Statistical Methods in Medical Research*, 7:354–370, 1998.
- R. Jin and Z. Ghahramani. Learning with multiple labels. In *Adv. Neural Information Processing Systems*, 2003.
- J. Nocedal and S. Wright. *Numerical Optimization (2nd ed.)*. Springer-Verlag, Berlin, New York, 2003.
- M. Qazi, G. Fung, S. Krishnan, R. Rosales, H. Steck, B. Rao, D. D. Poldermans, and D. Chandrasekaran. Automated heartwall motion abnormality detection from ultrasound images using Bayesian networks. In *Int. Joint Conf. on Artificial Intelligence*, 2007.
- V. C. Raykar, S. Yu, L. Zhao, A. Jerebko, C. Florin, G. Hermosillo-Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Int. Conf. on Machine Learning*, 2009.
- V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Knowledge Discovery and Data Mining (KDD)*, 2008.
- P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labeling of Venus images. In *Adv. Neural Information Processing Systems*, 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 *Conf. Empirical Methods on Natural Language Processing (EMNLP)*, 2008.
- A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *CVPR Workshop on Internet Vision*, 2008.
- D. J. Spiegelhalter and P. Stovin. An analysis of repeated biopsies following cardiac transplantation. *Stat. Med.*, 2 (1):33–40, Jan-Mar 1983.