Learning from multiple annotators: A Survey

Sharan Vaswani and Mohamed Osama Ahmed

Computer Science Department, University of British Columbia, Vancouver, BC, Canada
{sharanv,moahmed}@cs.ubc.ca

Abstract. For many supervised learning tasks, it is difficult or impossible to obtain ground truth labels for the set of training instances. Such applications include sentiment recognition, textual entailment where the labels are subjective, computer visions tasks like object detection, tracking and recognition which need large amounts of labeled data or medical applications such a tumor classification which requires a dangerous procedure such as biopsy to obtain the golden ground truth. All of these applications make use of multiple annotators to estimate the ground truth labels. There has been a substantial amount of research on how to combine ground truth estimates from multiple sources and use these estimates to train learning systems to make predictions on future instances. In this paper, we present a survey of the various methodologies that have been proposed in the literature to learn from multiple annotators. We briefly describe some of the approaches used and, in particular, we focus on the three methods proposed in [11], [19], and [12]. Moreover, we outline future research directions to better solve the problem.

1 Introduction

Supervised learning traditionally relies on a domain expert who 'teaches' the learning system with necessary supervision. In particular, the expert supervises the learning by providing the correct answers i.e. ground truth data (labels in the case of classification and real values in the case of regression) for a set of input instances called the training set. The supervised learning system once trained can then be used to predict labels/values for hitherto unseen input instances called the test set. Under some assumptions, the greater the size of the training set, the better the predictions on the test set. This motivates the need for multiple domain experts or even multiple non-experts providing labels for training the system. Indeed Snow et al. [16] show that multiple non-experts can perform as well as a single domain expert. Another motivation for using multiple annotators (annotator here refers to a person labeling a set of training instances) is when the ground truth data consists of subjective examples. This includes examples such as sentiment analysis, movie rating, and key phrase extraction. Such tasks warrant the need for multiple annotators for estimating the ground truth data. Another class of problems, where multiple annotators are useful, is scenarios where the true ground truth data (called the golden ground truth) is hard to obtain. Examples include the classification of a tumor as benign or malignant. The true ground truth data can be obtained only after a biopsy of the tissue which is an invasive, expensive and dangerous procedure and is not often done. To reduce the sparsity of the training set, multiple radiologists label the tumor as benign or malignant and their labels are combined to estimate the ground truth.

Effectively exploiting the ground truth data from various sources comes with its own set of problems. Each annotator has different accuracies in labeling the ground truth. For example, a domain expert will have a much higher accuracy compared to a non-expert annotating images on a crowd sourcing platform like Amazon Mechanical Turk. For training the system correctly, it becomes important to assign weights to the annotator estimates proportional to their accuracies in labeling the ground truth. There are other design decisions which need to be made. One can

estimate the ground truth from the annotator estimates first and then train the learning system using the appropriately weighted estimates as the training set [14] or estimate the ground truth and do the learning simultaneously [11]. Other important issues include the choice of the classifier for learning, the dependence of the annotator accuracies on input instances (some input instances might be relatively easier to label than others) and the dependence of the annotator accuracies on time.

In this survey, we compare state of the art methods to handle the issues described above and outline some directions of future work. This paper is structured as follows. Section 2 surveys the previous work on handling multiple annotators for the supervised learning setting. In section 3, we get into the precise mathematical details of three important methodologies. The survey is presented in an incremental chronological fashion. In particular, we start off with [14] and show how [11] improves it. We then present how [19] relaxes some of the assumptions of [11] leading to better results. Finally we present [12] which improves on [11] by formulating the problem in a different way. Section 4 presents some comparative results for each of the methodologies presented. In section 5 we conclude our survey and outline some directions for future work.

2 Overview

There has been much work on estimating ground truth labels from multiple annotators in the absence of gold ground truth data in the context of bio-statistics and epidemiology (see [6], [7], [3]). However, in the machine learning community, learning a classifier using estimates from multiple annotators was first done by Smyth et.al in [14,15] for labeling volcanoes in satellite images of Venus. However, this approach first estimated the ground truth from annotator estimates and used the probabilistic ground truths to learn a classifier. Some other approaches [1,2,13] exploit prior knowledge about the labeler similarities

In this paper, we survey approaches in the most general setting i.e. approaches which do not assume any prior knowledge about the labelers or relationships between them. Jin and Ghahramani [8] used the expectation maximization algorithm to estimate the ground truth and learn the classifier simultaneously. However they had a different use case where each training instance has a distribution over class labels. In 2009, Raykar et.al in [11] adapted this approach for learning from multiple annotators. They use the EM algorithm using the unknown ground truth labels as latent variables. In particular, they estimate the ground truth in the E step and use these estimates to train the classifier in the M step. They extend this idea for multiclass classification and regression. They also experiment with the Bayesian approach putting some prior on the weights and annotator accuracies to obtain the MAP estimate instead of the MLE. This work has been extended in different ways. For example, Raykar et.al assume the annotator accuracies are independent on the input instance at hand. This is not a very reasonable assumption and was relaxed in [19]. Other extensions include the use of Gaussian processes in [5] for jointly learning the ground truth and training the classifier. Yan et.al also explored an active learning approach for learning from multiple annotators in [18]. In [4], the authors consider the dependence of annotator accuracies on time. However modeling this dependence results only in a small improvement (around 1 - 2 %) in the results. Finally in 2013, Rodrigues et.al in [12] discussed certain disadvantages of modeling the problem using the unknown ground truth as latent variables. Instead, they use the also unknown annotator accuracies as latent variables and show the advantages of their proposed approach. In this survey, we focus on 3 important approaches described above namely the EM formulation proposed by Raykar et.al in [11]and its extensions in [19] and [12].

3 Methodology

In this section, we first describe the work done in [11]. The improvements presented in [19] and [12] are later discussed. We follow the notation used in [11]. Consider a data set of N data points $\left\{\mathcal{D}(x_i, y_i)\right\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ and and $y_i \in \mathcal{Y}$ are the feature vector and the actual label for the i^{th} instance, respectively. y_i is often referred to as the gold standard or the ground truth. Depending on $\mathcal{Y}$, there are 3 cases:

1. Binary classification: $\mathcal{Y} = \{0, 1\}$.
2. Multi-class classification: $\mathcal{Y} = \{1, .., K\}$.
3. Regression: $\mathcal{Y} = \mathbb{R}$.

In this paper, we focus on the classification problem. We consider the case of R different annotators or experts that provide the noisy labels $y_i^1, y_i^2, ..., y_i^R$ for the i^{th} instance.

In the following we discuss some of the different approaches for solving this problem

3.1 Majority voting

In the case of multiple labels, a commonly used strategy is to use the label that the majority agree on as an estimate of the true label. So the hidden true label $\hat{y}_i$ can be calculated as follows:

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

where the case of tie ($\hat{y}_i = 0.5$) can be broken by a super-expert or randomly [11].

3.2 Approach proposed by Rayker et.al[11]

The approach adopted in [11] assumes that the dependencies between x, y, and y^r can be modeled using the graphical model in Figure 1.

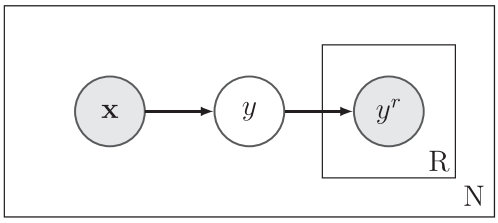

Fig. 1. Plate representation of general latent ground truth model [12]

.

This model assumes that the labels provided by the annotators y^r are independent of the input features given the true labels. This assumption may not be accurate, especially, in the case of easy instances where the annotators are less likely to make mistakes.

The algorithm proposed in [11] is based on a maximum-likelihood estimator that jointly learns the classifier, the annotator accuracy, and the actual true label. The performance of the r^{th} annotator is measured in terms of the sensitivity α^r and the specificity β^r with respect to the unknown gold standard. α and β are used to distinguish the true positive rate from the true

negative rate, respectively. Hence, $\alpha^r = \Pr[y^r = 1|y = 1]$ and $\beta^r = \Pr[y^r = 0|y = 0]$. In the case where these rates are symmetric, we use $\eta^r = \alpha^r = \beta^r$.

In order to incorporate prior knowledge about each annotator, the authors experiment with a Bayesian approach - they impose a prior on the sensitivity and specificity and derive the maximum-a-posteriori (MAP) estimate. However, we do not discuss this case here because of the spcae constraints. Moreover, we focus only on the classification case. However, the paper also discusses categorical, ordinal and regression problems.

Classification model: The paper uses the logistic regression model. This can be easily extended to other classification models. The classification model can be expressed as follows: $\hat{y} = \begin{cases} 1 & \text{if } w^T x \geq \gamma \\ 0 & \text{otherwise} \end{cases}$, where $\hat{y}$ is the estimated class label, w is the weight vector, x is the feature vector, and γ is the threshold that determines the operating point of the classifier. The receiver operating characteristics is determined by γ. For logistic regression, the model can be expressed as: $\Pr[y = 1|x, w] = \sigma(w^T x))$, where $\sigma(z) = 1/(1 + e^{-z})$.

Estimation problem: The problem can be formulated as follows. Given the training data $\mathcal{D} = \left\{(x_i, y_i^1, .., y_i^R)\right\}_{i=1}^N$ of N instances and R annotators, the goal is to estimate the sensitivity $\alpha = \left[\alpha^1, .., \alpha^R\right]$, the specificity $\beta = \left[\beta^1, .., \beta^R\right]$, and the ground truth $y_1, .., y_N$.

To solve this problem, [11] propose two approaches; the maximum likelihood estimator (MLE) and the MAP. We focus on the MLE approach.

Maximum likelihood estimator (MLE): With the assumption that the N training instances are sampled independently, the likelihood can be expressed by:

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} \Pr\left[y_i^1, .., y_i^R | x_i, \theta\right],$$

where $\theta = \{w, \alpha, \beta\}$ are the parameters.

Using the graphical model in Figure 1, the likelihood can be decomposed into:

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} \left\{\Pr\left[y_i^1, .., y_i^R | y_i = 1, \alpha\right] \times \Pr[y_i = 1|x_i, w] + \Pr\left[y_i^1, .., y_i^R | y_i = 0, \beta\right] \times \Pr[y_i = 0|x_i, w]\right\}.$$

Furthermore, the annotators are independent given the true label so:

$$\Pr\left[y_i^1, .., y_i^R | y_i = 1, \alpha\right] = \prod_{r=1}^{R} \Pr[y_i^r | y_i = 1, \alpha^r] = \prod_{r=1}^{R} \left[[\alpha^r]^{y_i^r} [1 - \alpha^r]^{1 - y_i^r}\right]$$

Similarly,

$$\Pr\left[y_i^1, .., y_i^R | y_i = 0, \beta\right] = \prod_{r=1}^{R} \Pr[y_i^r | y_i = 0, \beta^r] = \prod_{r=1}^{R} \left[[\beta^r]^{1 - y_i^r} [1 - \beta^r]^{y_i^r}\right].$$

EM algorithm: The parameter and label estimation is performed by an Expectation Maximization (EM) algorithm that iterates between estimating the ground truth and then finding the model parameters. The details of the algorithm are provided in [11]. The parameters are initialized using majority voting. The EM algorithm iterates between:

1. **E-step**: estimate the true labels $\mu_i = \Pr\left[y_i = 1 | y_i^1, .., y_i^R, x_i, \theta\right]$.
2. **M-step**: Using the estimates of the true labels, find the MLE parameters θ.

The detials of the derivation are provided in the paper.

3.3 Approach proposed by Yan et.al[19]

As explained in section 3.2, [11] assumed that the expert reliability is independent of the input instance. The work in [19] relaxes this assumption. It takes into account that some annotators are better at labeling some types of data points. Another difference is that [19] assumes that the performance of each annotator is symmetric for positive and negative examples so $\alpha = \beta = \eta$.

The graphical model that represents this approach is presented in Figure 2. It can be noticed in this figure that there is an edge from x to y^r which was not present in Figure 1. This edge models the dependence of the annotated labels on the input instance. This paper adopts two choices, namely the Bernoulli model and the Gaussian model for the probability distribution whereas it uses the logistic regression model as the classification model.

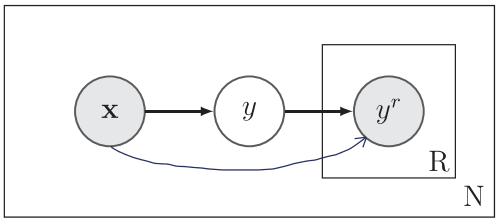

Fig. 2. Plate representation of the approach in [19]

.

Bernoulli model: This is similar to the model presented in section 3.2 except for the changes described above. It can be formulated as follows:

$$\Pr\left[y_i^{(r)}|x_i, y_i\right] = [1-\eta^r(x_i)]^{|y_i^r - y_i|}\,[\eta^r(x_i)]^{1-|y_i^r-y_i|},$$

where $\eta(x_i)$ is a logistic function given by: $\eta(x_i) = 1/(1+e^{-(u^r)^T x_i})$ and u^r is the weight vector for the r^{th} annotator.

Gaussian model: This model represents the dependence of the labels on the input using a Gaussian distribution as follows: $\Pr\left[y_i^{(r)}|x_i, y_i\right] = \mathcal{N}(y_i^{(r)}; y_i, \sigma(x_i))$, where $\sigma(x_i)$ represents the variance of the distribution given by: $\sigma(x_i) = 1/(1+e^{-(u^r)^T x_i})$ and u^r is the weight vector for the r^{th} annotator. This model can represent the performance of different annotators on different input instances. For example, large $\sigma(x_i)$ represents high uncertainty on this instance.

Maximum likelihood estimator (MLE): As in [11], the MLE was used to estimate the parameters $\theta = \{u_t, w_t\}$. This is done by maximizing the likelihood function which can be derived like in the previous subsection.

EM algorithm: This is similar to [11] which uses the EM approach to solve this problem. The difference in this paper is in the update rules which depends on the choice of the probability distributions. For the Gaussian and the Bernoulli models, there are no closed-form solution for this problems. So the authors used the LBFGS quasi-Newton method. A detailed derivation is provided in the paper.

3.4 Approach proposed by Rodrigues et.al[12]

In [12], the authors argue that the use of ground truth labels as latent variables is inefficient in cases where the number of classes is large. Since we need to marginalize the latent variables out of the likelihood, for applications like Part of Speech Tagging (POS), Named Entity Recognition (NER), the number of possible label sequences grows exponentially with the length of the input and hence it becomes intractable to marginalize over the output space. Another problem with using ground truth labels as latent variables is the explosion of the number of parameters which needs to be stored. For example for K classes, we need to store the probability $P(y_i^r|y_i)$ for each annotator r and for each class i. Thus we need to store $K \times K$ parameters for each annotator. The number of parameters become very large as K increases and this often leads to overfitting. To handle this problem, the authors propose a formulation which models the annotator accuracies as latent variables.

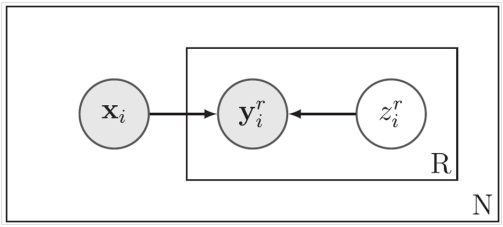

Fig. 3. Plate model for formulation considering annotator accuracies as latent variables

The authors introduce the variable z_i^r which in this case is a binary random variable whose value indicates whether the r^{th} annotator labeled the i^{th} instance correctly or not. Hence $z_i^r \sim Bernoulli\pi_r$, where π_r denotes the annotator accuracy. The plate model for this formulation is given in Figure 3. As explained before, x_i denotes the features for the i^{th} instance and y_i^r denotes the label assigned by annotator r to instance i. Similar to [11], the latent variable annotator accuracy in this case does not depend on the characteristics of the input instance being labeled (i.e. there is no arc from x_i to z_i^r. If for a particular instance i, an annotator r has $z_i^r = 0$, the label assigned to i by r is given by a random model $P_{rand}(y_i^r = k|x_i)$.

Maximum Likelihood Estimator (MLE): The likelihood function for this case is defined in equation 2

$$p(D, Z|\theta) = \prod_{i=1}^{N}\prod_{i=1}^{R} p(z_{i=1}^{R}|\pi^r)(p(y_i^r|x_i, z_i^r, \mathbf{w}) \tag{2}$$

Here $\theta = \{\ \pi,\ \mathbf{w}\ \}$ represent the model parameters. π is the vector of annotator accuracies whereas $\mathbf{w}$ represent the weights of the classifier. In this case too, logistic regression is used as the classifier. The likelihood can be further simplified to equation 3.

$$p(D, Z|\theta) = \prod_{i=1}^{N}\prod_{i=1}^{R} (\pi^r p_{LogReg}(y_i^r|x_i, \mathbf{w}) + (1 - \pi_r)p_{rand}(y_i^r|x_i)) \tag{3}$$

In this equation, $p_{LogReg}(y_i^r|x_i, \mathbf{w})$ is the output of the logistic regression classifier with weights $\mathbf{w}$. For simplicity, the authors use $p_{rand}(y_i^r|x_i) = \frac{1}{k}$ which corresponds to the intuition that if the annotator misclassifies a particular instance, the label assigned to it follows a uniform distribution.

EM algorithm: As in [11], the authors use Expectation Maximization to find the MLE for the likelihood function defined in equation 3. We skip the details of the EM derivation for this problem. The interested reader is advised to look into [12] for a detailed derivation.

4 Comparative experimental results

In this section, we present three experiments to compare the various approaches - simple majority voting baselines and the three approaches presented in [11,19,12].

4.1 Comparison 1: Between Raykar's approach and Majority Voting

This experiment uses a data set for digital mammography which is a screening tool that can detect breast cancer. The goal is to classify abnormal areas (lesions) into potentially malignant (1) or not (0), given a set of descriptive morphological features for a region on a image. The ground truth (whether the lesion is cancer or not) is obtained from biopsy. The ground truth was used to verify the results and to compare the conventional classifier with the proposed one. The dataset is available through [9]. It contains 497 positive and 1618 negative examples. Each instance is described by a set of 27 morphological features. The annotated labels were simulated according to α^r and β^r. This experiment consists of two trials as follows:

1. This trial simulated 5 radiologists with $\alpha = [0.9, 0.8, 0.57, 0.6, 0.55]$ and $\beta = [0.95, 0.85, 0.62, 0.65, 0.58]$.
2. This trial simulated 8 radiologists with only 1 expert (large value for α and β).

The paper compares the performance of the proposed algorithm with:

1. A classifier trained using the biopsy proved ground truth.
2. A classifier trained using majority-voting labels.

The results for the comparison are presented in Figures 4, 5, and 6. The comparison is based on the ROC curve. Figures 4 presents the results for the first trial. It shows that the method proposed in [11]is almost as good as that which uses the true ground truth labels and outperforms majority voting.

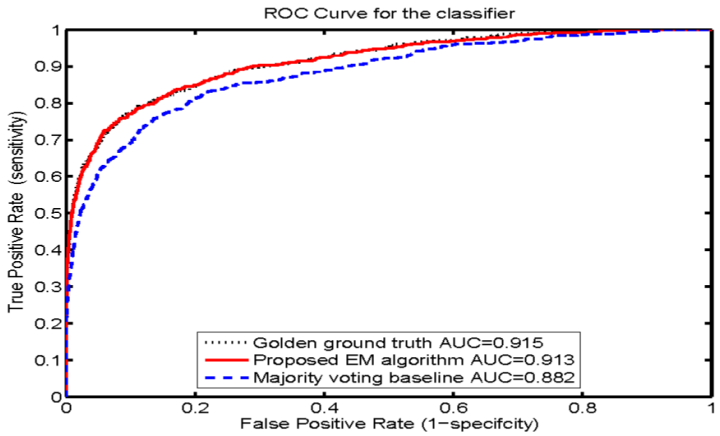

Fig. 4. Comparative ROC curves for digital mammography with 5 simulated radiologists

.

For the second trial, the results are presented in Figure 5. The figure shows the ROC curve for the estimated true labels. We see that the proposed method is better than the majority voting. This figure presents the estimate for specificity and sensitivity for each of the simulated radiologists. The estimates from the proposed methods are more accurate than the ones estimated using majority voting. This experiment shows the effectiveness of the proposed method, especially when there are only a few experts.

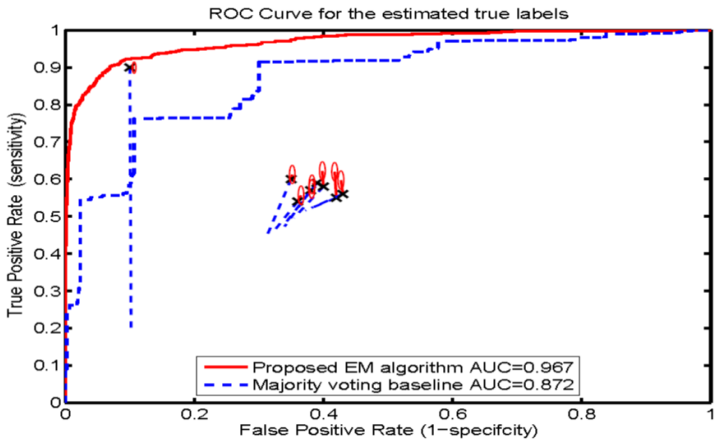

Fig. 5. Comparative ROC curves for digital mammography with 8 radiologists. The end of the dashed blue line shows the estimates of the sensitivity and specificity obtained from the majority voting algorithm. The end of the solid red line shows the estimates from the proposed method. The ellipse denotes one standard deviation

.

4.2 Comparison 2: Between Raykar's approach and Smythe's decoupled estimation

Figure 6 presents a comparison between the proposed method and the decoupled method proposed in [14] which first estimates the ground truth and then uses the probabilistic ground truth labels to train the classifier. This comparison uses the same digital mammography dataset with 8 simulated radiologists as described previously.

4.3 Comparison 3: Between Yan's and Raykar's approaches

This experiment was presented in [19]. It compares approaches used in [11] and [19]. As in section 4.1, the experiment uses a real data set for digital mammography. The data were labeled by three expert radiologists. Moreover, the ground truth was obtained through biopsy. The data contains 28 positive examples and 47 negative examples, and each instance is described by 8 morphological features. The data set was divided into training set (40 percent) and testing set (60 percent). Figure 7 presents the results for this experiment. The figure compares the following approaches:

1. M.L.-Original: The algorithm proposed in [11].
2. M.L.: The approach proposed in [19] with the following variations of the probability distributions:

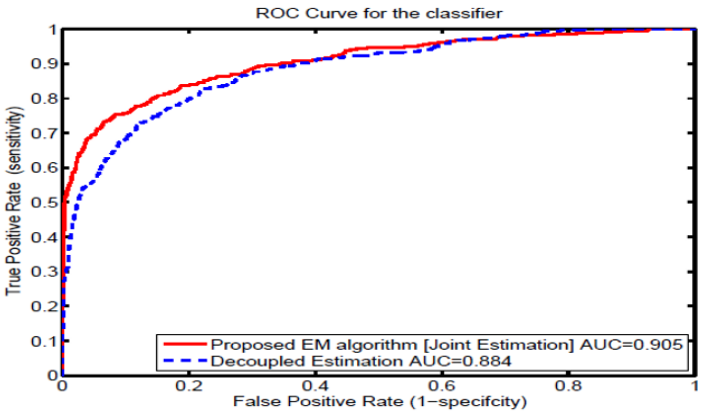

Fig. 6. Comparative ROC curves for [11] and the decoupled estimation in [14] for digital mammography

.

  - M.L.-Bernoulli(x): The approach described in Section 3.3
  - M.L.-Gaussian(x): The approach described in Section 3.3
3. L.R.-Majority: logistic regression classifier using the majority vote as the true label.
4. L.R.-Annotator$r$: logistic regression classifier using the labels from the r^{th} expert as the true label.
5. L.R.-true label: logistic regression classifier using the true label.
6. L.R.-Concatenation: This model was not clear from the paper.

The results shows that the algorithm proposed in [19] provides improvements over the one in [11]. This can be attributed to relaxation of the independence assumption in [11].

4.4 Comparison 4: Between Rodrigues's and Raykar's approaches

This subsection presents comparative results between the approach adopted in [11](Raykar) and [12] (MA - LR). Rodrigues et.al also compare against two baselines - against Soft Majority Voting (MVsoft) which trains the logistic regression classifier with the soft probabilistic labels obtained by the voting process and against Hard Majority Voting(MVhard) which trains the logistic regression classifier with deterministic labels obtained from majority voting.

The algorithms are compared on datasets from the UCI repository. Since these datasets contain only true labels for the training set, the authors simulate multiple annotators. To obtain the estimate for each annotator for each instance, the label for that instance is obtained by flipping the true label with a probability of $p(flip)$. This simulates an annotator with an average accuracy of $1 - p(flip)$. Since all the annotators do not label all instances, the authors introduce $p(label)$ which represents the fraction of instances in the training set labeled by the annotators. We show two of these simulated results - on the annealing dataset (798 instances, 38 features, 6 classes) and the ionosphere dataset (351 instances, 34 feature, 2 classes).

As can be seen from figures 8 and 9, the approach proposed by Rodrigues does as good as Raykar's approach when the number of classes is small (2 in the case of the Ionosphere dataset), whereas it dominates Raykar's approach as the number of classes increases to 6 in annealing dataset. This can be explained by the possible overfitting because of the large number of parameters in Raykar's approach. Rodrigues also provide comparative results on sentiment polarity using the dataset introduced in [10] and the music genre classification dataset introduced in [17]. The approach of Rodrigues et.al does better than Raykar's on such real world datasets as well.

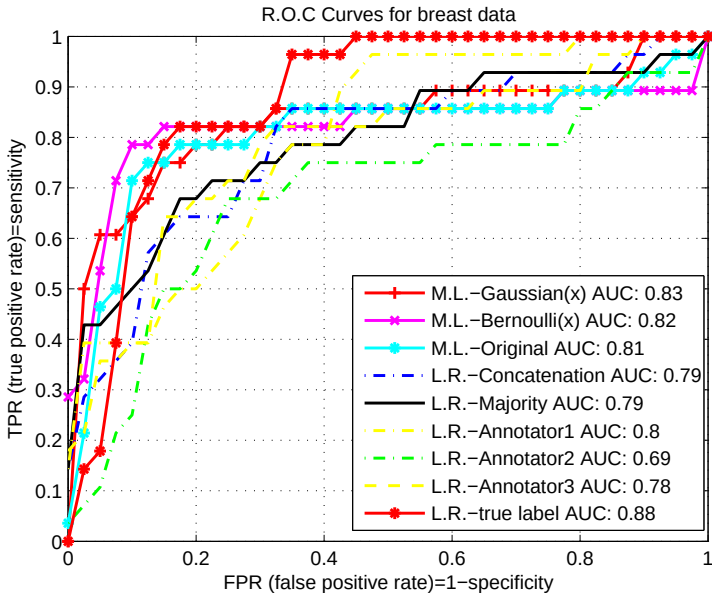

Fig. 7. Results for experiment 2 using breast cancer dataset [11]

.

5 Conclusion and Future research directions

In this paper, we presented some applications and scenarios for which learning from annotators becomes important. We present a survey of the field focusing particularly on supervised learning approaches which do not assume any prior knowledge on relation between annotators. These approaches are important in a variety of applications, such as NLP, computer vision, health care where the task can be crowd-sourced. We then presented the precise mathematical details of three of these approaches namely those proposed in [11], [19] and [12]. We compared their assumptions, mathematical formulations and results on certain benchmark datasets.

All previous literature assumes independence between the various annotators or assumes that this relation is known a priori. It will be interesting to model these relations and also learn these from the data. In fact Raykar et.al suggests a way of doing this. Rodrigues et.al proposes to extend their model using annotator accuracies as latent variables by modeling the dependence of these accuracies on the input instance as in [19]. We can integrate this model by also taking the dependence of these accuracies on time as in [4]. These extensions represent the most general framework for learning from multiple annotators. Another area which has not been explored is the dependence of the precision-recall on the classifier used. All the approaches described use logistic regression with the exception of [5] which uses Gaussian process regression. It will also be interesting to explore the Bayesian optimization in this context. Raykar et.al puts a prior on the weights and annotator accuracies and derive the maximum a-posteriori (MAP) estimate. The precision-recall using the Bayesian approach is much higher than the MLE estimate as was shown in experiments in [12]. It will be useful to explore the Bayesian approach for other models as well. We can also do the full Bayesian optimization rather than just the MAP estimate. This will give us appropriate confidence intervals for our estimate which may prove useful in sensitive applications such as medical diagnostics.

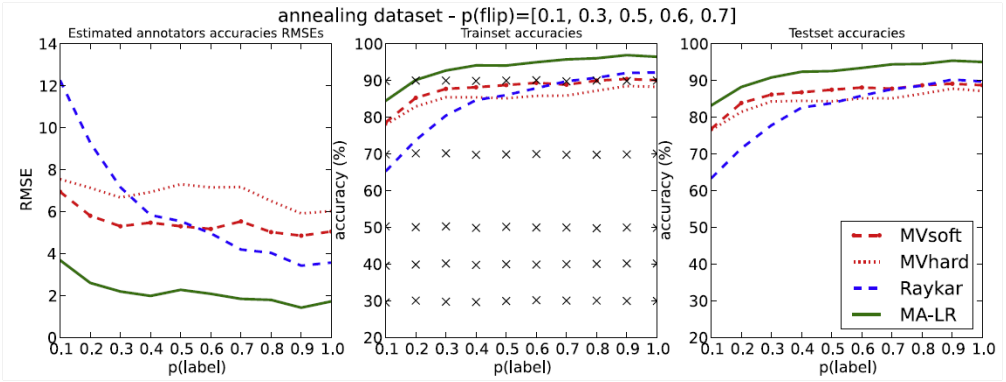

Fig. 8. 5 simulated annotators with $p(flip)$ = [0.1, 0.3, 0.5, 0.6, 0.7] Annealing Dataset (a) RMSE between the estimated and true annotator accuracies. (b) Training set accuracies (c) Test set accuracies

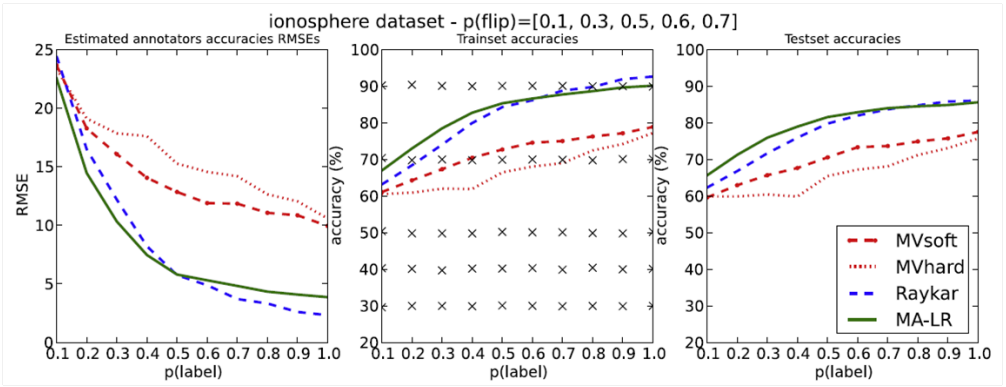

Fig. 9. 5 simulated annotators with $p(flip)$ = [0.1, 0.3, 0.5, 0.6, 0.7] Ionosphere Dataset (a) RMSE between the estimated and true annotator accuracies. (b) Training set accuracies (c) Test set accuracies

References

1. John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2007.
2. Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *The Journal of Machine Learning Research*, 9:1757–1774, 2008.
3. Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, pages 20–28, 1979.
4. Pinar Donmez and Jaime G Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 619–628. ACM, 2008.
5. Perry Groot, Adriana Birlutiu, and Tom Heskes. Learning from multiple annotators with gaussian processes. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 159–164. Springer, 2011.
6. Siu L Hui and Xiao H Zhou. Evaluation of diagnostic tests without gold standards. *Statistical methods in medical research*, 7(4):354–370, 1998.
7. Sui L Hui and Steven D Walter. Estimating the error rates of diagnostic tests. *Biometrics*, pages 167–171, 1980.

8. Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Advances in neural information processing systems*, pages 897–904, 2002.
9. Balaji Krishnapuram, Jonathan Stoeckel, Vikas Raykar, Bharat Rao, Philippe Bamberger, Eli Ratner, Nicolas Merlet, Inna Stainvas, Menahem Abramov, and Alexandra Manevitch. Multiple-instance learning improves cad detection of masses in digital mammography. In *Digital Mammography*, pages 350–357. Springer, 2008.
10. Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics, 2005.
11. Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *The Journal of Machine Learning Research*, 99:1297–1322, 2010.
12. Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Learning from multiple annotators: Distinguishing good from random labelers. *Pattern Recognition Letters*, 2013.
13. Victor S Sheng, Foster Provost, and Panagiotis 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. ACM, 2008.
14. Padhraic Smyth, U Fayyad, M Burl, P Perona, and P Baldi. Learning with probabilistic supervision. *Computational learning theory and natural learning systems*, 3:163–182, 1995.
15. Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems*, pages 1085–1092, 1995.
16. Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y 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. Association for Computational Linguistics, 2008.
17. George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.
18. Yan Yan, Glenn M Fung, Rómer Rosales, and Jennifer G Dy. Active learning from crowds. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1161–1168, 2011.
19. Yan Yan, Rómer Rosales, Glenn Fung, Mark W Schmidt, Gerardo H Valadez, Luca Bogoni, Linda Moy, and Jennifer G Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In *International Conference on Artificial Intelligence and Statistics*, pages 932–939, 2010.