Active Learning from Multiple Knowledge Sources

Yan Yan
ECE Northeastern Univ.
Boston, MA

Rómer Rosales
Yahoo! Labs
Santa Clara, CA

Glenn Fung
Siemens AG
Malvern, PA

Jennifer Dy
ECE Northeastern Univ.
Boston, MA

Abstract

The problem of learning by aggregating the opinions or knowledge of multiple sources does not fit the usual single-annotator learning scenario. In these problems, *ground-truth* may not exist and multiple annotators are available. In particular, active learning offers new challenges as, in addition to a data point, a knowledge source must also be optimally selected. This is of interest in a crowdsourcing setting as annotators may have varying expertise or be be adversarial; thus, the information they can offer will vary considerably. We propose an approach to address this situation by focusing on maximizing the information that an annotator label provides about the true (but unknown) label of data points.

1 Introduction

The traditional supervised learning scenario assumes that there is a single labeler/annotator (domain expert) that provides the necessary supervision. Such expert labels are considered the *ground-truth*. In settings involving automatic knowledge aggregation, such ground-truth labels may not be available and instead the information provided by multiple experts/non-experts (knowledge sources) need to be efficiently leveraged.

Machine learning approaches that address the multiple-annotator scenario in various settings have gained great interest recently [10, 17, 15, 6]. However, a consistent strategy for the active learning problem [7, 8] to a large extent is missing. In active learning, an algorithm is allowed to choose the samples from which it learns. In traditional active learning, an optimal sample is sought to be labeled by a knowledge source or process that is considered *unique* (*i.e.,* the varying level of expertise of multiple sources is mostly ignored). In contrast, in the crowdsourcing setting active learning additionally requires choosing from multiple knowledge sources. Since some annotators may be more reliable, some may be malicious, and their expertise may vary with the observed sample, clearly choosing an appropriate annotator is critical.

The presented approach focuses on maximizing the information that the chosen annotator label provides about the true (but unknown) label. Various ideas similar in spirit to the active learning scenario include: *repeated labeling* [14, 4, 13], the process of identifying labels that should be revised in order to improve classification performance, and more recently [9], a manner of learning where annotators are chosen randomly and then their responses corroborated using a separate model. The presented approach shares the motivation in [16] and to a lesser extent in [5, 1], but differs in the active learning criterion (Secs. 2-3).

2 Formulation

We consider a set of N data points $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ drawn independently from an input distribution. Let us denote $Y = \{y_i^{(t)}\}_{it}$ with $y_i^{(t)}$ the label for the i-th data point given by annotator t. The labels from

individual labelers might be incorrect, missing, or inconsistent with respect to each other. We introduce additional random variables $Z = \{z_1, \ldots, z_N\}$ to represent the *true* but usually unknown label for the corresponding data point. If we do not have access to *ground-truth*, all z_i are unobserved. We concentrate on this general case; however, in some problem instances partial ground-truth may exist. Some labels $y_i^{(t)}$ are observed, but in general it is expected that they are sparse and thus acquiring them optimally is of interest.

We let the annotation provided by labeler t to depend on the true (but unknown) label z and the input data point $\mathbf{x}$. As in [16], our motivation for this is that annotators may label certain data points with better accuracy than other data points and that this accuracy may depend on the properties of the data point itself. That is, their accuracy depend on the input being presented. In addition, labelers are assumed independent given the input data point and the true point label. A model consistent with the above can be given by

$$p(Y_O|X) = \prod_i \sum_{z_i} p(z_i|\mathbf{x}_i) \prod_{t|t\in\mathcal{T}_i} p(y_i^{(t)}|\mathbf{x}_i, z_i), \tag{1}$$

which represents the distribution for the observed labels $Y_O \subset Y$ conditioned on the input data, where $\mathcal{T}_i$ is the set of annotators that labeled the i-th data point. For binary classification, we employ a Bernoulli distribution to model annotator labels. However, we let its parameter(s) depend on the observed input:

$$p(y_i^{(t)}|\mathbf{x}_i, z_i) = (1-\eta_t(\mathbf{x}))^{|y_i^{(t)}-z_i|}\eta_t(\mathbf{x})^{1-|y_i^{(t)}-z_i|}; \text{with } \eta_t(\mathbf{x}) = (1+\exp(-\mathbf{w}_t^{\mathrm{T}}\mathbf{x}_i - \gamma_t))^{-1}. \tag{2}$$

Similarly, for the *true* label distribution $z|\mathbf{x}$ we let $p(z_i = 1|\mathbf{x}_i) = (1+\exp(-\alpha^T\mathbf{x}_i - \beta))^{-1}$.

For active learning we address the *pool-based* setting, where a number of unlabeled data points are simultaneously available for selection. For iteration τ we let the set $Y_U(\tau) \subset Y$ with $U = \{(k,s) \in \{1,...,N\}\times\{1,...T\}|y_k^{(s)} \text{ is unobserved}\}$ represent the labels that are unknown to the learning algorithm. As this is an iterative process, the set U varies across iterations. At each iteration τ, one tuple $(k^*, s^*) \in U(\tau)$ is chosen and the appropriate data point $\mathbf{x}_{k^*}$ is shown to labeler s^* for annotation. We use mutual information [2] as the criterion for optimally choosing the tuple $(k^*, s^*) \in U(\tau)$. Given this, the active learning problem is cast as follows:

$$[k^*, s^*] = \arg\max_{(k,s)\in U} I(z_k; [y_k^{(s)}, x_k]|X, Y_O), \tag{3}$$

where the information score is conditioned on having observed X and Y_O: the available data points and the labels provided by any annotator. We have assumed a given τ and thus removed U's dependency on it to ease notation. This maximization can be expressed in terms of the corresponding conditional entropies:

$$= \arg\max_{(k,s)\in U} \sum_{z_k, y_k^s} p(z_k|[y_k^s, x_k]; \theta)\log p(z_k|[y_k^s, x_k]; \theta) - \sum_{z_k} p(z_k|\theta)\log p(z_k|\theta). \tag{4}$$

In the above we have utilized a maximum likelihood estimate for the model parameters θ to calculate the information and therefore implying that all the information provided by the dataset is summarized by θ given the model structure. A MAP and Bayesian approach would follow in a similar manner.

From the model conditional independence assumptions, the first term can be written as $p(z_k|y_k^s, \mathbf{x}_k; \theta) = p(z_k|\mathbf{x}_k; \theta)p(y_k^s|\mathbf{x}_k, z_k; \theta)/\sum_{z_k} p(z_k|\mathbf{x}_k; \theta)p(y_k^s|\mathbf{x}_k, z_k; \theta)$

The second term can be estimated by observing $p(z_k|\theta) = \int p(z_k|\mathbf{x}_k; \theta)p(\mathbf{x}_k)$, since θ does not affect the prior $p(\mathbf{x}_k)$. An approximation $q(z_k) \approx p(z_k|\theta)$ can be obtained using X as a suitable sample from the prior distribution. Thus, we let: $q(z_k) = \frac{1}{N}\sum_{\mathbf{x}_k\in X} p(z_k|\mathbf{x}_k; \theta)$. Note that for this we have also made the standard assumption that the true distribution for z is consistent with the employed model.

Thus, the original optimization problem can then be solved by finding $[k^*, s^*]$:

$$[k^*, s^*] = \arg\max_{k,s} -\sum_{z_k} q(z_k)\log q(z_k) + \sum_{z_k, y_k^s} p(z_k|y_k^s, \mathbf{x}_k; \theta)\log p(z_k|y_k^s, \mathbf{x}_k; \theta). \tag{5}$$

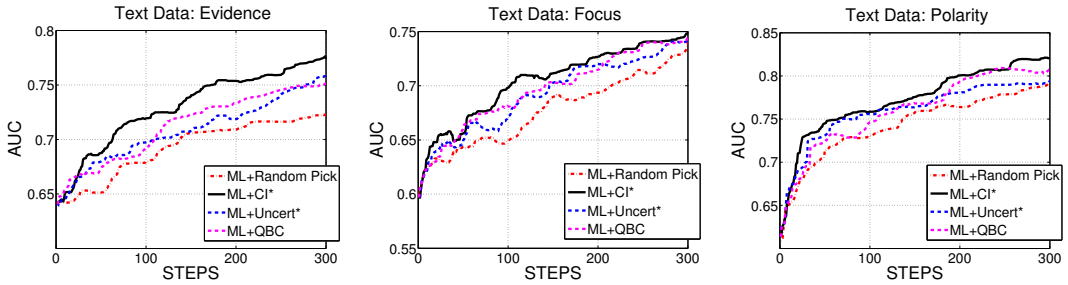

Figure 1: Accuracy and AUC for multi-labeler datasets as a function of number of active learning iterations.

This can be computed in $O(NT)$ once the appropriate distributions are known. The distributions for the two terms in the objective function require $O(|\mathcal{Z}||\mathcal{X}||\mathcal{X}|)$ and $O(|\mathcal{Z}||\mathcal{Y}||\mathcal{X}|)$ respectively for a given θ. For an efficient implementation, we calculate (for the first term) the entropy $H(z_k)$ at every iteration for each data point $\mathbf{x}_k$ that could be labeled. Likewise (for the second term), we calculate the appropriate entropy for each pair (k, s) still unlabeled. Note that after a data point is selected for labeling, this point may not necessarily be eliminated from the pool as it may be selected in the future for labeling by a different annotator.

For learning, we employ maximum likelihood to estimate $\theta = \{\alpha, \beta, \{\mathbf{w}_t\}, \{\gamma_t\}\}$ and use the Expectation Maximization algorithm [3] to maximize the conditional distribution for partially observed labels, Eq. 1. For inference, one can show that the task of inferring z for a new data point $\mathbf{x}$ is equivalent to calculating the conditional distribution $p(z|\mathbf{x})$ (given earlier).

3 Experiments

For the rest of the paper we will refer to our algorithm as *ML+CI**, where the prefix (ML) or **Multiple Labeler** refers to the nature of the classification algorithm, while the suffix (CI) or **conditional information** refers to the active learning strategy employed. The * symbol indicates that the algorithm selects the optimal labeler and point simultaneously. We consider several alternatives for selecting the samples and annotators:

1. *ML+QBC* (**Multi-Labeler utilizing Query by Committee**): utilizes the same multiple-labeler model as our approach but applies QBC[12] to select the optimal point to be labeled and any annotator (uniformly at random). Basically, without regard for annotator differences.

2. *ML+Uncert** (**Multi-Labeler utilizing Uncertainty**): utilizes the same multiple-labeler model as our approach but selects the sample and annotator based on the combination of uncertainty of label and annotator confidence given the data point $\eta(\mathbf{x})$ (see Eq. 2). This is the approach recently proposed in [16].

3. *ML+Random Pick*: utilizes the same multiple-labeler model but selects samples and annotators at random.

We test the different methods on scientific texts (PubMed and GeneWays corpus) prepared and made publicly available by [11]. This is a corpus of $10,000$ sentences, each annotated by 3 out of 8 available labelers. Each sentence was associated with several labels. We use the binarized *polarity*, *focus*, and *evidence* labels. We utilize a 1000-example subset where each sentence has been labeled by 5 annotators and a bag of words representation with 392 dictionary features. We randomly selected 300 samples as the initial training for the four competing methods, 300 points for active learning, and 400 points for testing. The 10-fold average test set AUC for the various (methods,tasks) pairs at each iteration is shown in Fig. 1.

The proposed *ML+CI** model achieved the best overall performance followed mostly by *ML+QBC*. This helps validate the information criterion utilized which allows for jointly optimizing for the sample-annotator

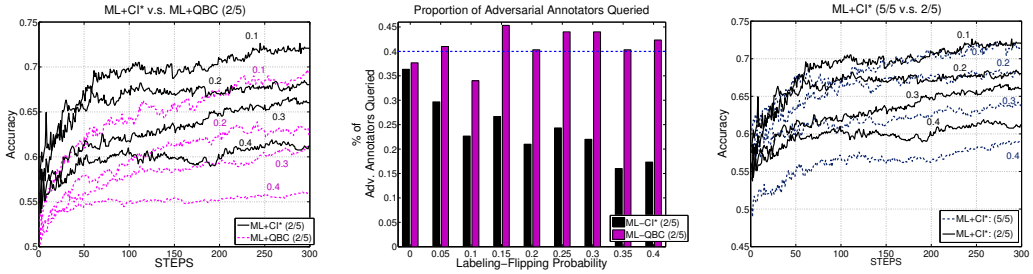

Figure 2: **Left:** Model accuracy (2/5 adversaries); **Center:** proportion of adversaries queried; **Right:** *ML+CI** accuracy (2/5 and 5/5 adversaries).

pair. *ML+QBC* achieves a reasonable performance because like our model, it selects the most informative sample; however, it assumes all annotators are *equally* good. *ML+Uncert** selects the most uncertain sample and the most confident annotator. However, choosing the most uncertain sample, may be suboptimal for improving classification performance, due to noise, outliers, or unimportant regions of interest.

We now investigate how adversarial labelers can hurt the performance of our approach, *ML+CI**. We conjecture that since our model selects annotators in each learning step, it can avoid or decrease the influence of these *bad* annotators. To simulate adversarial annotators, we randomly flip labels of points in the active learning pool with probability, p_ϵ. We performed the following experiments:

1. We compared the performance of *ML+CI** (for each active learning step) to *ML+QBC* as we vary $p_\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$ on two annotators. Larger p_ϵ leads to more aggressive adversaries. Due to limited space, we utilize *ML+QBC* as the comparative method because it had the second best performance in the previous set of experiments. To save space, we also show only the results on the *evidence* task. Similar results were obtained for the rest of the tasks. These are shown in Fig. 2(Left). These figures confirm that indeed *ML+CI** helps reduce the effects from bad annotators compared to *ML+QBC*.

2. In Fig. 2(Center), we report the proportion of adversarial annotators selected as we vary p_ϵ. This result indicates that our approach *ML+CI** is able to avoid malicious annotators better than *ML+QBC*.

3. In Fig. 2(Right), we show a comparison for our approach *ML+CI** when a) all five annotators are malicious and when b) only two annotators are malicious. As expected, *ML+CI** would perform worse as p_ϵ is increased and the drop in performance is less when there are fewer adversaries; however, the model maintains an acceptable performance that degrades slowly even when all labelers are not very accurate.

4 Conclusion

In this paper we developed and evaluated an approach for active learning in a multiple-annotator setting. In this new scenario, contrary to the traditional single labeler setting, an optimal sample- annotator pair must be determined. This adds an interesting and challenging dimension to active learning because some annotators may be more reliable than others, some may be malicious, and their expertise may vary with the observed sample. Thus, active learning is necessary as the information provided by some annotators is more valuable than that provided by others; moreover, the annotator value may depend on the specific unlabeled sample being considered. Our results show that the proposed approach outperforms several baseline and recently proposed methods in terms of both accuracy and area under the ROC curve (AUC). Similarly, our study comparing the resilience of these methods to malicious annotators reveals that our approach is robust in the sense that it largely avoids querying malicious annotators automatically. We believe this study on a new problem opens up interesting questions/directions for future research.

References

[1] A. Brew, D. Greene, and P. Cunningham. The interaction between supervised learning and crowdsourcing. In *NIPS Workshop on Comp. Social Science and the Wisdom of Crowds*, 2010.

[2] T. Cover, T. M., and J. A. Thomas. *Elements of information theory*. Wiley Interscience, New York, NY, USA, 1991.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data. *Journal of the Royal Statistical Society (B)*, 39(1), 1977.

[4] P. Donmez and J. G. Carbonell. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Conference on Information and Knowledge Management (CIKM)*, pages 619–628, 2008.

[5] P. Donmez, J. G. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Knowledge Discovery and Data Mining (KDD)*, 2009.

[6] G. Kasneci, J. V. Gael, D. Stern, and T. Graepel. CoBayes: Bayesian knowledge corroboration with assessors of unknown areas of expertise. In *Conference on Web Search and Data Mining*, pages 465–474, 2011.

[7] D. Lindley. On a measure of the information provided by an experiment. *Ann. Math. Stat*, 27:986–1005, 1956.

[8] D. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4:590–604, 1992.

[9] U. Paquet, J. Van Gael, D. Stern, G. Kasneci, R. Herbrich, and T. Graepel. Vuvuzelas and active learning for online classification. In *NIPS Workshop on Comp. Social Science and the Wisdom of Crowds*, 2010.

[10] 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 *International Conference on Machine Learning*, pages 889–896, 2009.

[11] A. Rzhetsky, H. Shatkay, and W. J. Wilbur. How to get the most out of your curation effort. *PLoS Computational Biology*, 5(5):e1000391, 2009.

[12] S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Fifth Workshop on Computational Learning Theory*, pages 287–94, 1992.

[13] 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)*, pages 614–622, 2008.

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

[15] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, 2011.

[16] Y. Yan, R. Rosales, G. Fung, and J. Dy. Active learning from crowds. In *International Conference on Machine Learning*, 2011.

[17] Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, and J. Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Int'l Conf. on Artificial Intelligence and Statistics*, pages 932–939, 2010.