Active Learning from Multiple Knowledge Sources

Y. Yan
Northeastern Univ.
Boston, MA USA

R. Rosales
Yahoo! Labs
Santa Clara, CA

G. Fung, F. Farooq, B. Rao
Siemens Healthcare
Malvern, PA

J. Dy
Northeastern Univ.
Boston, MA

Abstract

Some supervised learning tasks do not fit the usual single annotator scenario. In these problems, ground-truth may not exist and multiple annotators are generally available. A few approaches have been proposed to address this learning problem. In this setting active learning (AL), the problem of optimally selecting unlabeled samples for labeling, offers new challenges and has received little attention. In multiple annotator AL, it is not sufficient to select a sample for labeling since, in addition, an optimal annotator must also be selected. This setting is of great interest as annotators' expertise generally varies and could depend on the given sample itself; additionally, some annotators may be adversarial. Thus, clearly the information provided by some annotators should be more valuable than that provided by others and it could vary across data points. We propose an AL approach for this new scenario motivated by information theoretic principles. Specifically, we focus on maximizing the information that an annotator label provides about the true (but unknown) label of the data point. We develop this concept, propose an algorithm for active learning, and experimentally validate the proposed approach.

1 Introduction

The traditional supervised learning scenario assumes that there is a single teacher (or domain expert) that provides the necessary supervision. Such expert labels are then assumed to be the *ground-truth* utilized to build a learning model. In certain applications, such ground-truth labels may not be available and instead multiple experts/non-experts (annotators) provide the necessary supervision. For example, in medical image diagnosis, often radiologists disagree on a diagnosis unless a biopsy is made; this ground-truth can be impossible/expensive to collect. In addition, it is common for annotators to be more certain about some inputs. For example, radiologists are often experienced in certain medical conditions only; and thus, their opinion should be considered more valuable for a (usually unknown) subset of all the cases.

It is now increasingly easier to share and collect data from several sources; and consequently, possible to collect information (such as annotations) not just from one expert but by many experts and non-experts. This has fueled the phenomena such as *Crowdsourcing* [8] and, more concretely, large-scale collaboration tools such as Amazon Mechanical Turk (AMT). Other forms of multiple source supervision include opinions, reviews, product ratings, and many implicit forms of on-line user interaction.

Machine learning approaches that address the multiple-annotator scenario in various settings have gained great interest recently (*e.g.,* [14, 24, 22, 9]). However, a consistent strategy for the active learning problem [11, 12] has been missing to a large extent. In active learning, an algorithm is allowed to choose the data from which it learns. The most common setting is that where unlabeled data points are given and some of them must be chosen to be labeled by an oracle (*e.g.,* an expert). In the traditional active learning problem, an optimal sample is sought to be labeled by a unique annotator. In contrast, this paper addresses active learning from multiple annotators.

The new multiple annotator paradigm posits new challenges to the active learning algorithm – *not only do we need to select the optimal sample to label but also the optimal annotator to query.* Having multiple annotators adds an interesting dimension to active learning because some annotators may be more reliable than others, some may be malicious, and their expertise

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

may vary with the observed sample. Still, as labels have a cost, we would like to efficiently select points to be labeled to attain the *most gain* or achieve the best accuracy at a fixed cost.

There are two typical AL selection scenarios: *(a) pool-based active learning:* when examples can be chosen from an existing unlabeled point set and *(b) on-line or sequential active learning:* when a decision to label an example is made sequentially as each example becomes available. We focus on pool-based active learning.

Under certain assumptions active learning requires $O(\log(1/\epsilon))$ labeled examples to find a classification boundary providing ϵ error while passive learning requires $O(1/\epsilon)$ labeled examples [7]. Even though general theoretical guarantees on efficiency are available for a limited class of problems, empirical evidence suggests that active learning is efficient in common practical scenarios, where the objective is often that of maximizing accuracy given a budget.

Active learning (AL) methods can be divided into different categories depending on the strategy employed to select candidate examples. Active learning by *uncertainty sampling* [10, 2] selects the unlabeled data point whose label has highest uncertainty given the current model. *Query-by-Committee (QBC)* active learning [19, 7], selects data points that can optimally reduce the version space, a measure representing the volume of parameters consistent with the data. This results in the selection of points for which independently trained models disagree the most about their labels. *Expected error reduction* [16] aims to find an example that minimizes the expected generalization error (sometimes called risk) or reducing the expected total number of incorrect predictions. A related criterion, *expected model change* chooses the data point that when labeled maximizes the estimate of model change [18]. For some models this is equivalent to choosing a candidate unlabeled example that generates the objective function gradient of the largest magnitude. Some of the approaches above can also be tied to information theoretic criteria, in particular *QBC*.

In this paper we consider a more direct use of the mutual information criterion [3] and define the problem explicitly to address the multiple annotator situation, with the goal of selecting the *most informative* labels based on the annotator characteristics. Thus, annotator and data point selection are optimized simultaneously. Specifically, for the available data points, we focus on maximizing the information that the chosen annotator label provides about the true (but unknown) point label.

Various ideas similar in spirit to the active learning scenario include: *repeated labeling* [21, 5, 20], the process of identifying labels that should be revised in order to improve classification performance, and more recently [13], a manner of learning where annotators are chosen randomly and then their responses corroborated using a separate model.

The presented approach shares the motivation of [23] in the sense that both approaches address the multi-labeler active learning scenario, but they are different as [23] focuses on a form of uncertainty active learning. We compare with this approach in the experiments section. This paper is related to a lesser extent to [6, 1]. The approach in [6] is in the same class as [23] with respect to the use of the uncertainty sampling principle. However, [6] requires querying multiple labelers for every point, since the reward function depends on majority voting. The proposed approach does not require querying multiple labelers (this is likely wasteful) and does not need majority voting. Previous work suggests that majority voting is clearly sub-optimal ([15, 24, 22]). Finally, the work in [1] compares annotator selection based also on classifier uncertainty and (additionally) disagreement; however it focuses on the selection of data points and not on the selection of annotators.

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. In the setting addressed in this paper the labels from individual labelers might be incorrect, missing, or inconsistent with respect to each other. We introduce additional variables $Z = \{z_1, \ldots, z_N\}$ to represent the *true* but usually unknown label for the corresponding data point.

We let $\mathbf{x}_i$ and z_i for $i \in \{1, ...N\}$ be random variables in the input space $\mathcal{X}$ and output space $\mathcal{Z}$ respectively. Similarly, we let $y_i^{(t)}$ be random variables over the space of labels $\mathcal{Y}$, where $t \in \{1, ..., T\}$. If we do not have access to the *ground-truth*, all of the variables z_i are unobserved. We concentrate on this more general case; however, in some problem instances partial ground-truth may be available. 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.

2.1 A probabilistic model for multiple labelers

In modeling multiple annotators, we consider the annotation provided by labeler t to depend on the true (but usually unknown) label z and the input data point $\mathbf{x}$. Our motivation for this is that annotators may la-

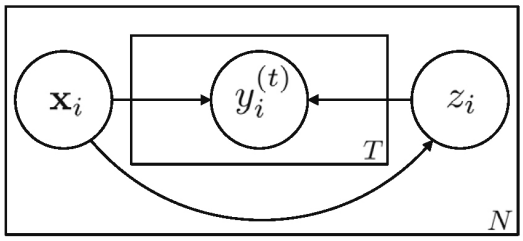

Figure 1: Graphical Model for $\mathbf{x}$, $\mathbf{y}$, and z respectively inputs, annotator-specific labels, and ground truth label (for simplicity $\alpha, \beta, \{\gamma_t\}$, and $\{\mathbf{w}_t\}$, with $t \in \{1, .., T\}$, are excluded)

bel 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.

These modeling considerations were proposed in [24] (exclusively for standard supervised learning) and are represented in the graphical model given in Figure 1. We will use this representation to develop and evaluate an active learning strategy.

Throughout this paper, we will be interested in the conditional distribution for the observed labels $Y_O \subset Y$ conditioned on the input data. One can show that this distribution is given by:

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

where we have used the conditional independence assumptions implied by the given graphical model and $\mathcal{T}_i$ is the set of annotators that provided a label for the i-th data point.

For binary classification, a Bernoulli conditional distribution is an appropriate choice for the conditional distribution of annotator labels:

$$p(y_i^{(t)}|\mathbf{x}_i, z_i) = (1-\eta_t(\mathbf{x}))^{|y_i^{(t)}-z_i|}\eta_t(\mathbf{x})^{1-|y_i^{(t)}-z_i|}, \qquad (2)$$

with $\eta_t(\mathbf{x})$ a logistic function of $\mathbf{x}_i$ and t:

$$\eta_t(\mathbf{x}) = (1+\exp(-\mathbf{w}_t^T\mathbf{x}_i - \gamma_t))^{-1}. \qquad (3)$$

Similarly, for the conditional distribution of the true label $z|\mathbf{x}$ we choose the standard logistic regression model:

$$p(z_i = 1|\mathbf{x}_i) \quad = (1+\exp\left(-\alpha^T\mathbf{x}_i - \beta\right))^{-1}. \qquad (4)$$

2.2 Active Learning

We utilize a *pool-based* active learning modality, where a number of data points available for labeling are known by the algorithm at a given moment in time. As some data points may not have been labeled by some annotators, we choose to represent the set of unobserved labels rather the set of unlabeled points. Therefore, for iteration τ we let the set $Y_U(\tau) \subset Y$ with $U = \{(k,s) \in \{1,...,N\} \times \{1,...T\} | y_k^{(s)} \text{ is unobserved}\}$ to represent the labels that are unknown to the learning algorithm.

As this is an iterative process, the set U could vary 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 t^* for annotation. Thus, after this, the label $y_{k^*}^{(s^*)}$ is no longer unobserved. Note that unlike the standard active learning problem, here both the data point and the labeler must be selected, rather than just a data point. In this iterative process, data points are chosen until a labeling budget has been depleted.

We consider the mutual information [3] as an appropriate criterion for choosing the tuple $(k^*, s^*) \in U(\tau)$. Given this, the active learning problem can be cast as follows:

$$(k^*, s^*) = \arg\max_{(k,s)\in U} I(z_k; [y_k^{(s)}, x_k]|X, Y_O), \qquad (5)$$

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 simplify the notation.

This maximization can be expressed in terms of the corresponding conditional entropies (H) as follows:

$$\begin{aligned}
&\max_{k,s} && H(z_k|X, Y_O) - H(z_k|[y_k^s, x_k]; X, Y_O) \\
&= \max_{k,s} && H(z_k|\theta) - H(z_k|[y_k^s, x_k]; \theta) \\
&= \max_{k,s} && \sum_{z_k, y_k^s} p(z_k|[y_k^s, x_k]; \theta) \log p(z_k|[y_k^s, x_k]; \theta) \\
& && \quad - \sum_{z_k} p(z_k|\theta) \log p(z_k|\theta), \qquad (6)
\end{aligned}$$

where a semicolon is employed to separate random variables from paramters for clarity (where required).

In the above we have utilized a maximum likelihood point estimate for the model parameters θ to simplify the calculation of the information score. We are implying that all the information provided by the dataset

is summarized in the model parameters θ given the proposed model structure. This could potentially be extended to incorporate a distribution over θ conditioned on the data X and Y_O in a MAP or Bayesian formulation.

The first term can be computed by using Bayes' rule:

$$p(z_k|y_k^s, \mathbf{x}_k; \theta) = \frac{p(z_k|\mathbf{x}_k;\theta)p(y_k^s|\mathbf{x}_k, z_k;\theta)}{\sum_{z_k} p(z_k|\mathbf{x}_k;\theta)p(y_k^s|\mathbf{x}_k, z_k;\theta)}. \quad (7)$$

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:

$$q(z_k) = \frac{1}{N} \sum_{\mathbf{x}_k \in X} p(z_k|\mathbf{x}_k;\theta). \quad (8)$$

Note that for this we have also made the standard assumption that the true distribution for z is consistent with the employed model.

Once these quantities have been calculated, the original optimization problem can be expressed in a simple manner:

$$\begin{aligned}(k^*, s^*) &= \arg\max_{k,s} -\sum_{z_k} q(z_k)\log q(z_k) \\ &+ \sum_{z_k, y_k^s} p(z_k|y_k^s, \mathbf{x}_k;\theta)\log p(z_k|y_k^s, \mathbf{x}_k;\theta) \quad (9)\end{aligned}$$

This can be computed in $O(NT)$ once the appropriate distributions in each term have been obtained. The required 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 and s that is still unlabeled. The first is a vector of size at most N, the second is a table of size at most $N \times T$. Note that after a data point is selected for labeling by an annotator, this point may not necessarily be eliminated from the pool as the same data point may be selected in the future for labeling by a different annotator.

2.3 Learning and Classification

We utilize the maximum likelihood learning criterion to estimate $\theta = \{\alpha, \beta, \{\mathbf{w}_t\}, \{\gamma_t\}\}$. The Expectation Maximization (EM) algorithm [4] is employed to maximize the conditional distribution for partially observed labels, Eqn. 1. The mathematical derivation is rather standard and is omitted due to space limitations. Note that the objective function is different from that utilized in [24] where the difference stems from the unavailability of all the annotations. One can show that the problem of inferring z for a new data point $\mathbf{x}$ not in the training set is equivalent to applying Eqn. 4 with input $\mathbf{x}$.

Algorithm 1 *ML+CI** Algorithm.

Inputs: data set X, available annotations Y_O, and missing annotations U.
Train multiple annotator model with preset training data.
while stopping condition not met **do**
  $O = U^C$ //C: set complement
  Find optimal label $(k^*, s^*) \in U$ to request which maximizes equation (9) conditioned on observations X and Y_O
  Request label $y_{k^*}^{(s^*)}$ for data point $\mathbf{x}_{k^*}$ from annotator s^*
  Update trained model by re-training with the new sample $[\mathbf{x}_{k^*}, y_{k^*}^{(s^*)}]$.
  $U \leftarrow U - \{(k^*, s^*)\}$
  $i = i + 1$.
end while
return

3 Experiments

In this section, we investigate how successfully our proposed algorithm uses conditional information for active learning. As we mentioned in our introduction, different annotators may have varying expertise and some annotators may even be (intentionally) malicious. In this new active learning setting, not only do we select a sample to label at each step but also the corresponding "best" annotator to label this sample. In this section, we also explore how robust our proposed algorithm is to malicious or adversarial annotators.

For the rest of the paper we will refer to our algorithm as *ML+CI**. The acronym before the plus sign (ML) means **Multiple Labeler** and refers to the nature of the classification algorithm, the acronym after the plus sign refers to the active learning strategy employed by the corresponding algorithm. In our case CI stands for **conditional information** and * indicates that our algorithm selects the "best" available labeler simultaneously.

In designing our algorithm, we found that there are several possible ways for selecting the samples and annotators. In this empirical study, we compare our ap-

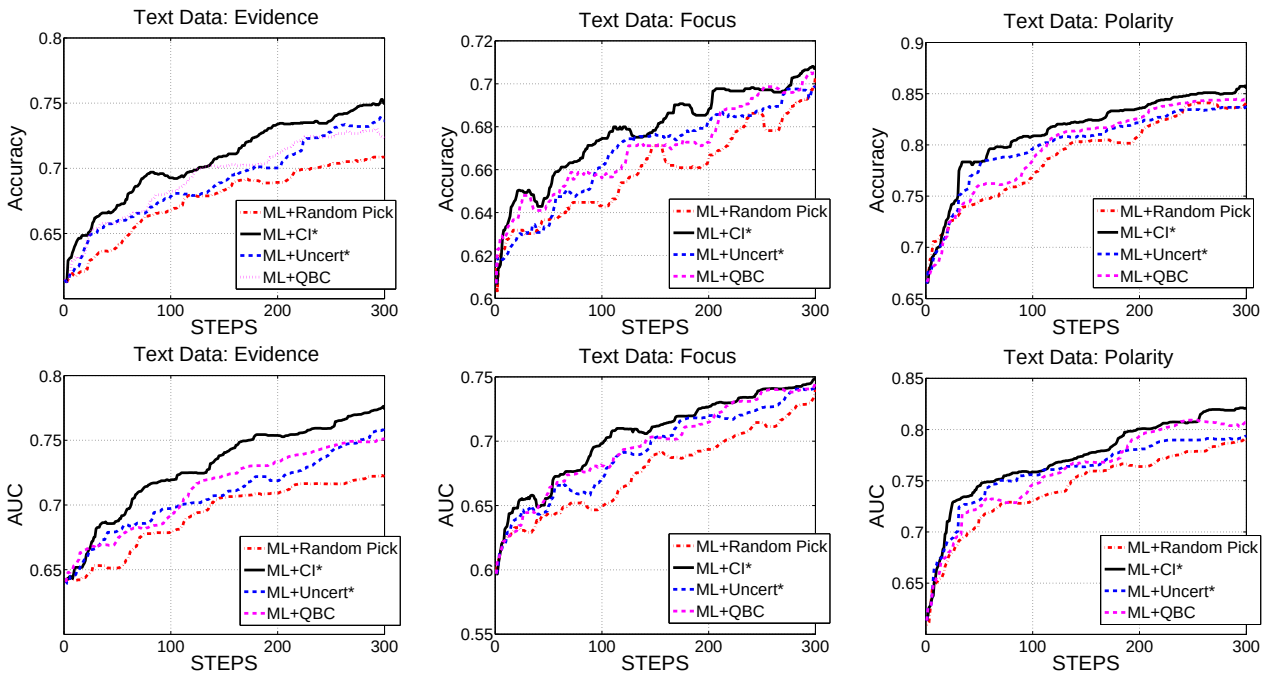

Figure 2: Accuracy and AUC for multi-labeler datasets as a function of number of active learning iterations.

proach to these possible alternative models/baselines (where we apply a naming convention, multi-labeler model + active learning model):

1. *ML+QBC* (**Multi-Labeler utilizing Query by Committee**): This method utilizes the same multiple-labeler model (Eqn. 4) as our approach but uses query by committee to perform active learning. We used $M = 5$ committee classifiers by randomly selecting 200 out of 300 points from our training data and an additional classifier trained on all training data as the final one for making predictions. In each step, it selects the sample for which the M classifiers disagree the most based on Kullback-Liebler divergence[3]. Then, it randomly picks an annotator to label the sample, without regard for annotator differences.

2. *ML+Uncert** (**Multi-Labeler utilizing Uncertainty**): This method utilizes the same multiple-labeler model (Eqn. 4) as our approach but selects the most uncertain sample. This means, it selects that sample that is closest to the boundary. Then, it queries the annotator that has the largest confidence based on η (see Eqn. 3). This is the same as the method in [23].

3. *ML+Random Pick*: This method utilizes the same multiple-labeler model as our approach but selects samples and annotators uniformly at random. This serves as a baseline approach.

3.1 Learning Performance

In this section we compare learning performance in terms of how efficient the same underlying model learns using the various competing active learning strategies described above. Note that all of the compared approaches use a multi-labeler model that allows for learning the annotator expertise. This was done so that the observed differences in performance can be attributed to differences in the active learning strategies compared.

3.1.1 Scientific text data

We test the different methods on scientific texts (PubMed and GeneWays corpus) prepared and made publicly available by [17]. It contains a corpus of $10,000$ sentences each that has been annotated by 3 out of 8 available labelers. For each sentence there are several available labels. Here, we use the *polarity*, *focus*, and *evidence* labels and binarize them into two classes. We utilize a 1000 examples subset where each sentence have been labeled by five annotators. For each available sentence, we calculated the frequency of occurrence of the most common words. After preprocessing and normalization of the occurrences, we ended up with 1000 samples and 292 features. We randomly selected 300 samples as the initial training for the four different competing methods mentioned above, 300 points for active learning sample selection, and the remaining 400 points to test the methods (i.e.,

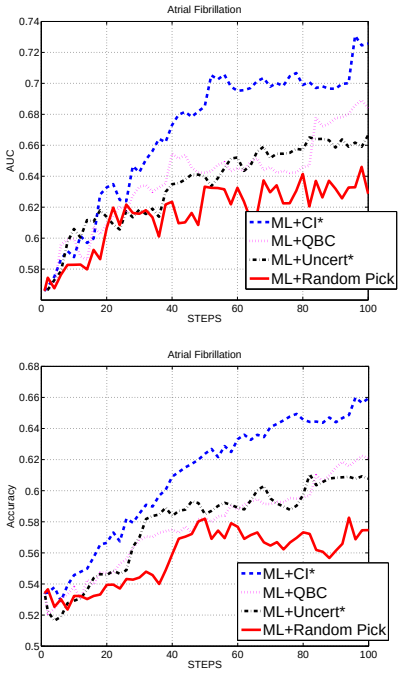

Figure 3: Accuracy and AUC for multi-labeler AF problem as a function of number of active learning iterations.

to measure the test accuracy and area under the receiver operating characteristic curve (AUC)) in each selection step. To test our active learning approach, we plot the test accuracies of the various methods after each active learning step. The average test set accuracy and AUC for the various (methods,tasks) pairs at each iteration is shown in Fig. 2. The classification problems are polarity, evidence and focus respectively.

As shown in the figures, our *ML+CI** model achieved the best overall performance; the second overall best performance is achieved by the *ML+QBC* model. *ML+CI** yields the best performance since it optimizes for both the sample and the annotator that allow the classifier to gain the most information simultaneously after every active learning step. *ML+QBC* achieves a relatively good performance because like our model, it selects the most informative sample; however, it randomly selects an annotator to query. In other words, 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, in some cases, may be suboptimal for improving classification performance, due to noise, outliers, or unimportant regions of interest.

3.1.2 Medical text data

We also tested the different methods on medical text data related to automatic detection of Atrial Fibrillation (cardiac arrhythmia of abnormal heart rhythm) from unstructured medical text. This is a representative example of a common and very relevant area in medical text analysis where the goal is to ascertain or infer that a piece of text (a sentence, passage, or document) refers to a particular, given topic or concept, in this case, atrial fibrillation (AF).

In this experiment we are using actual electronic medical records (EMR) from various medium/large-size hospitals. We designed this experiment to work at the passage level. A passage is a sequence of word/tokens extracted from a document. Thus, each training point represents a passage-based observation.

Our dataset consists of a set of 1058 passages from a medical database containing a variety of different medical records: discharge notes, visit notes, bills, etc. The passages have been annotated by an expert labeler (nurse abstractor) and four non-expert labelers. Each passage is labeled into one of two categories: whether the passage is relevant in determining (or providing clear evidence) that the patient has a history of AF or not. The text to be analyzed is represented based on a combination of the document metadata (document type, date, formatting information) and contextual information. For a passage of interest, the context is defined as the section the passage is in, the distribution of words in the passage, and the relationships between these words. When a document is analyzed, two main elements are identified: (1) document metadata and (2) the actual text in the document (the content). These are represented as a vector of real numbers.

After preprocessing, cleaning and normalization of the resulting representative vectors, we ended up with 998 samples and 323 features. We randomly selected 30 samples as the initial training for the four different competing methods mentioned above, 300 points for active learning sample selection (however, we stop after 100 samples have been selected), and the remaining points to test the methods. Like in the previous experiment, we test our active learning approach and plot the test accuracies of the various methods after each active learning step. The results, in Fig. 3, are consistent with the earlier results, and show a clearer difference between the proposed method and those compared, further helping validate the approach.

3.2 Adversarial Annotators

In this section, we investigate how adversarial labelers can hurt the performance of our approach, *ML+CI**.

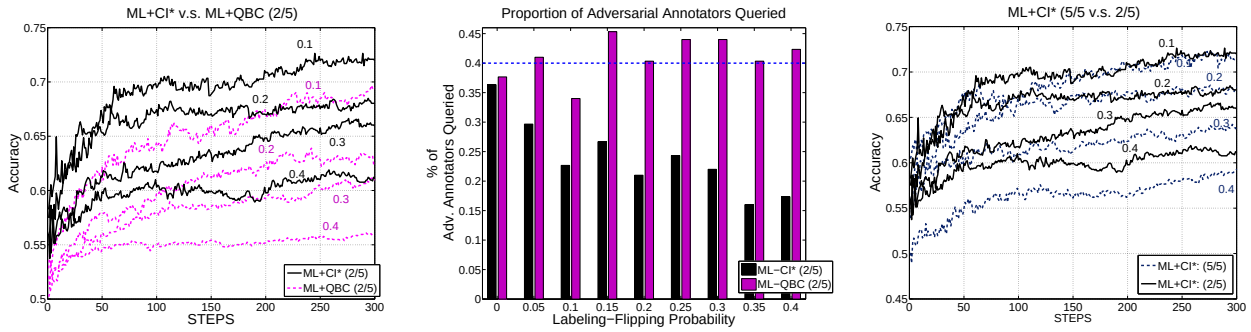

Figure 4: **Left:** Model accuracy (2/5 adversaries); **Center:** proportion of adversaries queried; **Right:** *ML+CI** accuracy (2/5 and 5/5 adversaries). Note that *ML+CI**(5/5) outperforms *ML+QBC*(2/5).

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:

- 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. This means that two of the annotators become adversarial with different degrees of "maliciousness" depending on the value of p_ϵ. 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 evidence. The results for the other labels provide similar conclusions. Results are shown in Figure 4(Left). These figures confirm that indeed *ML+CI** helps reduce the effects from bad annotators compared to *ML+QBC*.
- In Figure 4(Right), we provide a bar plot reporting how many times each method selects adversarial annotators as we vary the flipping probability p_ϵ. This results verify that our approach *ML+CI** is able to avoid malicious annotators better than *ML+QBC*.
- In Figure 4(Center), we show a comparison of performances of 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 the flipping probability 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 Conclusions

In this paper we have developed an approach for active learning in a multiple-annotator setting. This is one of the first attempts to formalize this active learning problem. In this new scenario, contrary to the traditional single labeler setting where only an optimal sample needs to be selected for labeling, an optimal (sample, annotator) pair must be determined. The chosen annotator is queried to label the selected sample. Having multiple annotators adds an interesting 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, the information provided by some annotators is more valuable than that provided by others; moreover, this may depend on the specific unlabeled sample being considered.

Our approach is based on maximizing the information that an annotator label provides about the true (but unknown) label of the data point. We validated our approach on real medical text data that have been labeled by multiple annotators. In this data, the annotators kept for learning are different from those used for test/evaluation. Our results show that the proposed approach outperforms baseline methods (a variant of our multiple-annotator active learning algorithm but one that selects the most uncertain sample, a query-by-committee approach, and random selection) in terms of both accuracy and area-under-the-curve. Similarly, our empirical study comparing the resilience of these methods to malicious annotators reveals that our approach is more robust compared to the competing methods. Moreover, our approach is able to largely avoid querying malicious annotators automatically. We believe that this study can motivate 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] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

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

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

[5] 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.

[6] 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.

[7] Y. Freund, S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 2-3:133–168, 1997.

[8] J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business, 2008.

[9] 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.

[10] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, pages 3–12, 1994.

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

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

[13] 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.

[14] 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.

[15] 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. Conference on Machine Learning (ICML)*, pages 889–896, 2009.

[16] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *18th International Conference on Machine Learning*, pages 444–448, 2001.

[17] 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.

[18] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *In Advances in Neural Information Processing Systems (NIPS*, pages 1289–1296. MIT Press, 2008.

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

[20] 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.

[21] 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.

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

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

[24] 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 *International Conference on Artificial Intelligence and Statistics*, pages 932–939, 2010.