Active Learning from Crowds

Yan Yan YAN.Y@HUSKY.NEU.EDU
ECE Department, Northeastern University, Boston, MA USA

Rómer Rosales ROMERR@YAHOO-INC.COM
Yahoo! Labs, Santa Clara, CA USA

Glenn Fung GLENN.FUNG@SIEMENS.COM
CAD and Knowledge Solutions, Siemens Healthcare, Malvern, PA USA

Jennifer G. Dy JDY@ECE.NEU.EDU
ECE Department, Northeastern University, Boston, MA USA

Abstract

Obtaining labels can be expensive or time-consuming, but unlabeled data is often abundant and easier to obtain. Most learning tasks can be made more efficient, in terms of labeling cost, by intelligently choosing specific unlabeled instances to be labeled by an oracle. The general problem of *optimally* choosing these instances is known as active learning. As it is usually set in the context of supervised learning, active learning relies on a single oracle playing the role of a teacher. We focus on the multiple annotator scenario where an oracle, who knows the ground truth, no longer exists; instead, multiple labelers, with varying expertise, are available for querying. This paradigm posits new challenges to the active learning scenario. We can now ask which data sample should be labeled next and which annotator should be queried to benefit our learning model the most. In this paper, we employ a probabilistic model for learning from multiple annotators that can also learn the annotator expertise even when their expertise may not be consistently accurate across the task domain. We then focus on providing a criterion and formulation that allows us to select both a sample and the annotator/s to query the labels from.

1. Introduction

Most research on supervised learning techniques rely on an often overlooked (still reasonable) assumption that a single domain expert can provide the required supervision; for example, ground-truth labels in classification problems. However, it is becoming more common for supervision to be available in many forms as data can be shared and processed by increasingly larger audiences. This makes it possible for not just one but many experts (and non-experts) to offer some form of supervision. A very clear example is that provided by *Crowdsourcing* (Howe, 2008) mechanisms such as Amazon Mechanical Turk (AMT), but it can be as implicit as many forms of on-line user interactions (*e.g.,* product ratings, opinions, user clicks, etc.).

This phenomena renders most supervised learning approaches sub-optimal and clearly motivates a necessary shift in machine learning towards models that are annotator aware. 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. Thus, this seems to indicate that in this new multi-labeler scenario, labels provided by different annotators should in general be treated differently.

The use of information from multiple annotators is motivated by a multitude of factors that can be summarized as follows: 1) It is difficult, and in some cases impossible, to collect a single *golden* ground-truth in some problem domains. For example, in the radiology field, specialists regularly disagree on the diagnosis for the same radiological image; thus, often requiring a biopsy which can be difficult or impossible to collect. 2) It is often the case that an annotator does not have the appropriate knowledge for annotating all the data, even for a particular domain. Some annotators will be accurate for certain situations, while some will have a highly variable accuracy, and some annotators will have specific biases. 3) In many instances, collecting annotations from multiple non-expert annotators can be less costly

Appearing in *Proceedings of the 28^{th} International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

than collecting annotations just from one expert. 4) Collaboration and knowledge sharing is becoming more common, and thus technology for combining multiple opinions (annotations) will become necessary.

Very recently, a few approaches that aim at addressing the challenges provided by this new setting have been proposed. These include (Yan et al., 2010; Raykar et al., 2009; Snow et al., 2008). The common topic in this family of work is how to properly utilize the labels provided by each annotator in a distinct and more optimal manner instead of treating all labels equally.

We address a new aspect of this problem motivated by the fact that while multiple annotators may be available, labels still have a cost. In fact, in many learning tasks the labeled data is limited in quantity or expensive to obtain, but the amount of unlabeled data is large or easy to obtain. If we want to efficiently label the unlabeled data to make the most gain (*e.g., learn the most at a given cost*), traditional supervised learning need only to efficiently identify the *most useful* data point to label given the information obtained so far. In this new multi-labeler scenario, an additional, interesting problem arises: How do we efficiently identify the *most useful* annotator given the information provided by the multitude of annotators?

Thus, the main goal of this paper is to address the key question: can we automatically choose the most appropriate annotator for a particular task so that learning can be sped up or be made more efficient in general (*e.g.,* less costly)? That is, we address the problem of active learning from multiple annotators.

2. Related Work

In the active learning scenario (Lindley, 1956; MacKay, 1992; Seung et al., 1992), unlabeled data are available and at each iteration an algorithm is able to choose an example for a user/oracle to label. There is normally a cost incurred for requesting each label. The objective is that of learning the appropriate concept with certain accuracy while incurring the lowest cost. An alternative problem is that of maximizing accuracy at a fixed cost. When examples can be chosen from an unlabeled data set this is normally referred to as *pool-based* active learning. In contrast, when a decision to label an example has to be made sequentially as each example becomes available, this is referred to as *online* active learning. In this paper we focus on pool-based active learning.

Active learning can drastically lower labeling costs. It has been shown that the number of data points needed for learning some functions can be reduced drastically (exponentially) if these points are chosen appropriately. For a class of noiseless, deterministic classification problems, active learning requires $O(\log(1/\epsilon))$ labels to find the classification boundary guaranteeing ϵ error while passive learning requires $O(1/\epsilon)$ examples (Freund et al., 1997). Even though strict error bounds like the above can be analytically obtained only for a limited class of problems, empirical evidence suggests that active learning can be efficient in more practical scenarios (Cohn et al., 1996; McCallum & Nigam, 1998).

One way to categorize active learning (AL) methods is by contrasting the underlying criteria that are optimized. Active learning by *uncertainty sampling*, such as (Lewis & Gale, 1994; Cohn et al., 1996), is the process of selecting the unlabeled data point whose label has highest uncertainty given the current model. The rationale behind this approach is the notion that by querying data points in the most uncertain areas, the model will efficiently improve its performance. A different criterion is provided in (Roy & McCallum, 2001) where the data point of choice is that which, when labeled, minimizes the estimate of expected (future) error. While this criterion attempts to directly optimize performance, an analytical expression for the expected error rarely can be obtained, and sampling needs to be employed. However, appropriately sampling from a distribution of interest is by itself a difficult task in practice. Moreover, this method requires retraining the model for every point that is considered for labeling; this can be computationally prohibitive.

Query-by-Committee (QBC) active learning (Seung et al., 1992; Freund et al., 1997), offers a different perspective: the data point that reduces the most the *size of the version space*, a measure representing the number or volume of parameters that are consistent with the data, is selected. An approximate solution is to choose that point for which a set of independently trained models disagrees the most (regarding its label). In QBC, the model only needs to be evaluated, not retrained, for every competing data point.

Despite the extensive work in active learning. We are not aware of any approaches that take into account the multi-labeler scenario. The closest approach to active learning in this area has been the use of repeated labeling (Smyth et al., 1995; Donmez & Carbonell, 2008; Sheng et al., 2008). This relies on the identification of what labels should be reacquired in order to improve classification performance or data quality. The idea of trying to model different levels of expertise among labelers has been addressed in (Raykar et al., 2009; Yan et al., 2010) and later in (Kasneci et al., 2011). However, none of these approaches addressed the active annotator selection based on individual annotator properties. More recently, the approach in (Paquet et al., 2010) proposed a form of learning where annotators are chosen randomly and then their responses corroborated using a separate model.

The active learning problem is challenging in the multi-labeler setting due that annotators in general provide different amounts of information for the learning model. This information content is dependent of the available data points. Therefore, an effective approach should simultaneously select the data point and annotators that provide the *most useful* information given what has been learned so far.

3. A Probabilistic Multi-Labeler Model

Let us consider N data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. We denote the label for the i-th data point given by annotator t as $y_i^{(t)}$. In general, the labels from individual labelers may not be correct, may be missing, and may not be consistent with each other. Let us denote the true (unknown) label for the i-th data point as z_i. In our formulation 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 $t \in \{1, ..., T\}$. In general, all the variables z_i and some of $y_i^{(t)}$ are not observed[1].

3.1. Model Definition

For compactness, we set X to be the collection of points $\mathbf{x}_1, ..., \mathbf{x}_N$ and Y to be the collection of associated labels that have been provided by all annotators (observed). Given training data, X and Y, the model we will utilize shall produce an estimate for the ground-truth, denoted by Z, a classifier function for predicting the label z for new instances $\mathbf{x}$, and a model of the annotators' expertise as a function of the input $\mathbf{x}$.

We train our classifier by assuming a probabilistic model over random variables $\mathbf{x}$, y, and z with a graphical model as shown in Figure 1, previously introduced in (Yan et al., 2010). Unlike this work, in our active learning scenario we are interested in cases for which not all labelers have provided a label for some data points; that is, Y is not fully observed. The conditional distribution is then given by:

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

where $\mathcal{T}_i^N$ is the set of annotators that provided a label for the i-th data point.

This model posits that the annotation provided by labeler t depends on the true (but unknown) label z and the input $\mathbf{x}$. This is both an interesting and realistic scenario as annotators may not necessarily have the knowledge to label all the data with equal accuracy. Instead their accuracy depends on what input they observe. We believe this is a valid assumption in particular for non-expert annotators (for an experimental validation see (Yan et al., 2010)). Note also that labelers are assumed independent given the input and the true label.

[1] In some applications a number of variables z_i may be observed, but in general we assume none is available.

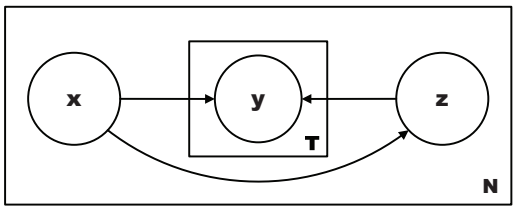

Figure 1. Graphical Model for $\mathbf{x}$, $\mathbf{y}$, and z.

In order to fully define our model, $p(y_i^{(t)}|\mathbf{x}_i, z_i)$ and $p(z_i|\mathbf{x}_i)$ need to be specified. Various options are possible depending on the problem domain. We use the definitions given in (Yan et al., 2010) for the labeler model. $p(y_i^{(t)}|\mathbf{x}_i, z_i)$. Two instances are considered:

A Gaussian model:

$$p(y_i^{(t)}|\mathbf{x}_i, z_i) \quad = \quad \mathcal{N}(y_i^{(t)}; z_i, \sigma_t(\mathbf{x}_i)), \qquad (2)$$

where the variance depends on the input $\mathbf{x}$ and is specific to each annotator t. For binary classification $\sigma_t(\mathbf{x})$ is a logistic function of $\mathbf{x}_i$ and t:

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

A Bernoulli model:

$$p(y_i^{(t)}|\mathbf{x}, z_i) = (1 - \eta_t(\mathbf{x}))^{|y_i^{(t)} - z_i|} \eta_t(\mathbf{x})^{1-|y_i^{(t)} - z_i|}, \quad (4)$$

where $\eta_t(\mathbf{x})$ is also a logistic function of the input and the labeler identity t.

The Gaussian model allows for assigning a lower variance to input regions where the labeler is more consistently correct relative to areas where there are inconsistencies. Similarly, the Bernoulli model assigns a higher probability of the labeler being correct to certain input areas relative to other areas. Ideally this differentiation can be learned from data.The distribution $p(z_i|\mathbf{x}_i)$ can take various forms. Since we are interested in classification, we use the logistic regression model:

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

A straightforward extension is to use multinomial logistic regression in case we are interested in multi-class classification or a Gaussian model for regression.

3.2. Learning

Given our model, we estimate the set of all parameters, $\theta = \{\alpha, \beta, \{\mathbf{w}_t\}, \{\gamma_t\}\}$, by employing the maximum like-

lihood criterion. A standard approach to solve our maximum likelihood problem, with missing variables z_i and various $y_i^{(t)}$, is to employ the expectation maximization (EM) (Dempster et al., 1977) algorithm. The relevant mathematical derivation, including the form of the E and M steps is given in (Yan et al., 2010).

Note that in the active learning scenario, not all the input data points will have complete labels (*i.e.,* a point may be labeled only by certain labelers during active learning), Eq. 1 is used instead of the joint distribution of the fully observed labels employed in (Yan et al., 2010)). The only subtle difference is that the product over data points does not include those terms associated to labels not provided by the respective labeler.

Once the parameters α, β have been estimated in the learning stage, a new data point $\mathbf{x}$ can be classified by utilizing Eq. 5. This is equivalent to inferring z for a new data point $\mathbf{x}$ (not labeled by any labeler) using the given graphical model.

4. Active Learning: Optimally Selecting New Training Points and Annotators

Given a trained model of multiple annotators as described in the previous section (Eq. 5), we want to simultaneously select the data point and labeler that will allow the model to learn efficiently (once this point/label is added to the training set). We have divided this problem into two goals:

a.) To pick a new training point to be labeled and added to the training set such that our model performance efficiently improves (usual active learning goal).

b.) To pick the appropriate labeler among the set of available labelers by choosing the one that will provide the most confident label for the new chosen training point.

In the next two subsections we propose two simple but effective strategies that can be combined to achieve these two goals in a optimal fashion.

4.1. Which new training point to pick?

One of the simplest and widely used strategies for querying new training samples in active learning scenarios is uncertainty sampling (Lewis & Gale, 1994). Under this strategy, an active learner queries samples for which the corresponding predictions are the most ambivalent, or least certain. Since we are considering a binary classification probabilistic model, under this simple strategy, we are interested in potential samples for which the probability of $p(z = 1|\mathbf{x})$ is close to $\frac{1}{2}$, in other words we want to query points that are solutions to the following simple optimization problem:

$$\min_{\mathbf{x}}(\frac{1}{2} - p(z|\mathbf{x}))^2$$

Using Eq. 5, we have:

$$\begin{aligned} &\text{argmin}_{\mathbf{x}}[\frac{1}{2} - p(z|\mathbf{x})]^2 \\ &= \text{argmax}_{\mathbf{x}} p(z=1|\mathbf{x})(1-p(z=1|\mathbf{x})) \\ &= \text{argmax}_{\mathbf{x}} \frac{\exp(-\boldsymbol{\alpha}'\mathbf{x}-\beta)}{(1+\exp(-\boldsymbol{\alpha}'\mathbf{x}-\beta))^2} \end{aligned} \quad (6)$$

For convenience, we denote $\tilde{\mathbf{x}} \triangleq [\mathbf{x}'1]'$ and $\tilde{\boldsymbol{\alpha}} \triangleq [\boldsymbol{\alpha}'\beta]'$ and simplify Eq. 6 as:

$$\max_{\mathbf{x}} f(x) = \max_{\mathbf{x}} \frac{\exp(-\tilde{\boldsymbol{\alpha}}'\tilde{\mathbf{x}})}{(1+\exp(-\tilde{\boldsymbol{\alpha}}'\tilde{\mathbf{x}}))^2} \quad (7)$$

The gradient of the above formulation is:

$$\nabla f(\mathbf{x}) = \frac{\exp(-\tilde{\boldsymbol{\alpha}}'\tilde{\mathbf{x}})(\exp(-\tilde{\boldsymbol{\alpha}}'\tilde{\mathbf{x}})-1)}{(1+\exp(-\tilde{\boldsymbol{\alpha}}'\tilde{\mathbf{x}}))^3}\tilde{\boldsymbol{\alpha}} \quad (8)$$

We can assume without loss of generality that $\tilde{\boldsymbol{\alpha}}$ is not zero. It is easy to show that $\tilde{\boldsymbol{\alpha}}$ is orthogonal to $\tilde{\mathbf{x}}$ ($\tilde{\mathbf{x}} \perp \tilde{\boldsymbol{\alpha}}$), or more specifically $\boldsymbol{\alpha}'\mathbf{x}+\beta = 0$ which basically corresponds to points on the classifier hyperplane. Since we know $\frac{\exp(-f)}{(1+\exp(-f))^2}$ is a concave function *w.r.t* f which takes the only maximum at $f = 0$, we have that $\boldsymbol{\alpha}'\mathbf{x}+\beta = 0$ defines a unique hyperplane that characterizes the space of samples $\mathbf{x}$ that interest us most.

4.2. Which expert to pick?

Uncertainty sampling does not provide a clear criterion to the problem of choosing an annotator. In our approach, among all the infinite set of points that reside on the aforementioned hyperplane, we choose one for which there exist a labeler that can provide a new label with (relative) maximal confidence. For a given $\mathbf{x}$, Eq. 3 provides information about how confident a labeler t would be in providing a label. So ideally we want a tuple $(\mathbf{x}^*, t^*)$ that solves the following optimization problem:

$$\min_{t,\mathbf{x}} \tilde{\sigma}(\mathbf{x}, t) \quad (9)$$

Where $\tilde{\sigma}(\mathbf{x}, t) = \sigma_t(\mathbf{x}) = (1+\exp(-\mathbf{w}_t^T\mathbf{x} - \gamma_t))^{-1}$ and $t \in \{1, \ldots, T\}$.

Unfortunately, this problem is very complex and computationally expensive to solve (non-convex, non-differentiable) which makes it not a feasible option in an active learning setting. However, using the fact that $f(x) = (1+\exp(-x))^{-1}$ (for $x \in \Re$) is a monotonically non-decreasing function, we can consider the following alternative bi-convex optimization problem:

$$\min_{\mathbf{x},\mathbf{p}} C(\boldsymbol{\alpha}'\mathbf{x}+\beta)^2 + \mathbf{p}'[\mathbf{w}_1,\ \mathbf{w}_2, \ldots,\ \mathbf{w}_T]'\mathbf{x} + \mathbf{p}'\boldsymbol{\gamma} \quad (10)$$

constrained to: $C \geq 0, \quad \mathbf{p} \geq \mathbf{0}, \quad \sum_t \mathbf{p} = 1$, where $\mathbf{p} \triangleq [p_1, p_2, \ldots, p_T]', \quad \boldsymbol{\gamma} \triangleq [\gamma_1\, \gamma_2, \ldots, \gamma_T]'$ and $C \geq 0$ is a trade-off between the two competing goals: most uncertain points and points labelers are confident on labelling. The components of $\mathbf{p} \succeq \mathbf{0}$ are automatically determined and can be thought of as terms that weigh annotator importances when minimizing variances.

This formulation defines a linearly constrained, bi-convex optimization problem for which solutions can be efficiently found using any Newton or Quasi-Newton optimization method. In our case we use a variation of the BFGS method for linearly constrained problems (Nocedal & Wright, 2003).

In general, when solving formulation (10) we obtain an optimal $\mathbf{x}^*$ to be labeled. This may be problematic in pool-based active learning (our experimental setting) as $\mathbf{x}^*$ may not be in our candidate pool. To address this problem, we search through our sample pool to find the most similar (closest in the Euclidean sense) sample to be added to our training set in the next iteration. We summarize this approach in Algorithm 1.

Algorithm 1 Multi-Labeler Active Learning Algorithm

Inputs: model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{w}_1, \ldots, \mathbf{w}_T$, $\boldsymbol{\gamma}$, C, and number of steps K;
$s = 1$;
while $s <= K$ **do**
  Use Eq. 10 to find the current best sample template $\mathbf{x}_{\mathtt{tem}}$;
  Find the nearest point $\mathbf{x}^*$ to $\mathbf{x}_{\mathtt{tem}}$;
  Use Eq. 3 to find the most reliable/confident annotator for $\mathbf{x}^*$ given the model learned up to this point;
  Re-train the model with new data point and label (update $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{w}_1, \ldots, \mathbf{w}_T$, $\boldsymbol{\gamma}$);
  $s = s + 1$;
end while
RETURN $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{w}_1, \ldots, \mathbf{w}_T$, $\boldsymbol{\gamma}$;

5. Experiments

In this section, we compare our active learning multiple labeler algorithm, *active learning+multi-labeler (blue)*, against baseline methods on text data from scientific sentences (Rzhetsky et al., 2009) with multiple annotations and on three UCI Machine Learning Repository (Newman et al., 1998) benchmark data (ionosphere, bupa and pima) with simulated multiple annotators. There are no existing active learning methods for the multiple annotator model; thus, we compare our method against the following baselines, testing different aspects of our model: (1) our active learning component for selecting the sample to train, but instead of selecting the annotators and learning from the multi-labeler approach, we use all annotators and simply use the majority vote to learn a logistic regression classifier, *active learning+majority vote (green)*; (2) we apply our multi-labeler algorithm to learn the classifier and select the annotator/s to label the new sample and apply random (from a uniform distribution) selection to sample instances for labeling, *random sample+multi-labeler (red)*; and, (3) random sampling to perform active learning and the majority vote of all annotators to learn a logistic regression classifier, *random sample+majority vote (magenta)*. We report the accuracies and area under the receiver operating characteristic curve (AUC) of the various approaches.

5.1. An Illustrative Example

We present a simple example to illustrate our approach. For this example, we used the galaxy dim data described in (Odewahn et al., 1992) which contains 4192 samples and 14 features for which binary labels are available. For simplicity and visualization, we picked the two features with the highest correlation with the labels for this experiment (Figure 2(l)). We assumed that there were labels available from 3 simulated annotators, by clustering the data into 3 parts (using k-means (Jain et al., 1999)) and assuming that each annotator is an expert on one single cluster (with labeling accuracy: 80%), but not familiar with the other two clusters (with labeling accuracy: 55%). Annotators expertise is represented in Figure 2(c).

We also divided the dataset into three subsets: 1000 data points for initial training; 2000 samples for active learning and 1192 for testing. We then tracked what exactly our model was selecting after 600 iterations, and in each iteration which annotators were selected to label the samples. As shown in Figure 2(r), we found that, the majority of the selected samples were close to the class boundary (as expected), and not surprisingly, the annotators labeling the selected boundary samples were in effect the ones who were confident with the boundary samples (annotators 2 and 3). Annotator 1 was never required to label samples since her/his confident cluster was far away from the boundary.

5.2. Text Data

Rzhetsky et al. (2009) prepared a publicly available corpus of $10,000$ sentences from scientific texts (PubMed and GeneWays corpus) each of which has been annotated by 3 out of 8 labelers. It contains sentences labeled based on different dimensions. Here, we use the *polarity*, *focus*, and *evidence* labels and binarize them into two classes. We utilize the 1000 data set from the second annotation cycle where each sentence is labeled by five annotators. After feature processing and normalization, we converted each sentence by recording the frequency (numbers of appear-

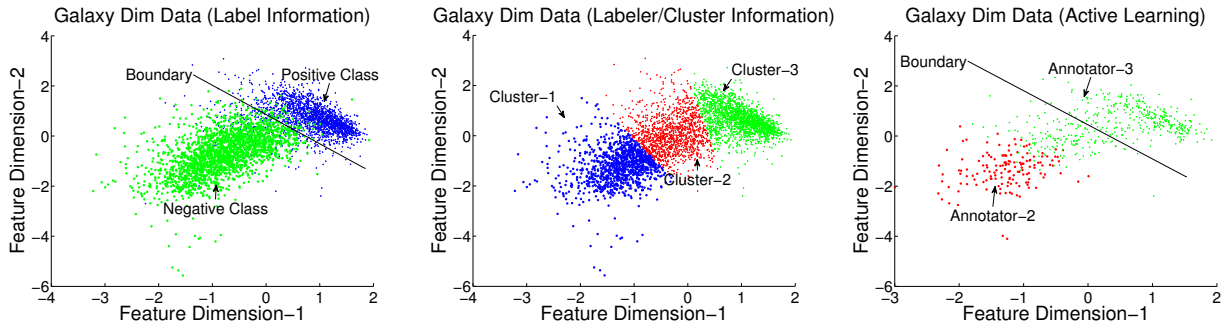

Figure 2. (**left**) Labels, (**center**) Areas of Labeler expertise and (**right**) annotator selection information for the simplified two-dimensional Galaxy Dim Data.

ance time) of the most common words in these data sets, and it ended up with a numerical feature matrix with 1000 samples and 292 column features. Then, 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., measure the test accuracy and AUC) in each selection step.

To test our active learning approach, we plot the test accuracies as the various methods learn each additional sample selected in every active learning step. Figures 3(l), (c) and (r) are the accuracy plots for label: polarity, evidence and focus respectively; while Figures 4(l), (c) and (r) are the AUC plots. As shown in all six figures, *active learning* combined with the probabilistic *multi-labeler* model (indicated as *active learning+multi-labeler*) maintained the best performance under both accuracy and AUC measures, the second best is our *multi-labeler* model but with *random sampling* in selecting samples to learn next. This makes sense because our multi-labeler could extract information available among the annotators and query the annotations from the most reliable annotator for incoming samples. The third best is *majority vote+active learning*, since the model for training simply accepts the majority vote from all annotators but without analyzing their annotation qualities, which makes it inappropriate for situations with difficult annotation tasks (*i.e.,* when experts' annotations may be varying and unreliable). Nevertheless, because of its active learning component, it is still better than the worst approach, *regression+random sample*.

5.3. UCI Benchmark Data

We also performed experiments on three sets of UCI (Asuncion & Newman, 2007) benchmark data: pima (351, 33), ionosphere (768, 8), and bupa Liver (345, 7), **(# samples, # features)**. Since multiple annotations 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. Then, we assume that each one of the five simulated labelers $i, i = 1 \dots 5$ 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 time (we randomly switch labels for 35% of the data samples).

We randomly divided the data into three sets: pima (100,100,151); ionosphere (200,200,368); bupa (100,100,145), where the items in the parenthesis stands for the number of samples in the initial model training, active learning set, and test set respectively. We repeat these random split five times for each of UCI data sets and measure the average and standard deviation of the accuracies and AUC at each active learning step.

In figures 5-6, we plot the average accuracies and area under the curve with their standard deviations on the four competing methods at each learning step. Again our *active learning+multi-labeler* dominated all of the other methods. In most cases, *random sample+multi-labeler* comes in second, except on the pima data (Figure 6. Here *random sampling* hurts its performance. The figures show that *active learning* does improve model performances compared to *random sampling* for both *multi-labeler* and *majority vote* regression classifiers. Interestingly, *active learning+majority vote* has performances close to *random sample+multi-labeler* in later steps for the ionosphere and bupa data (Figures 5-6) and even beat it in early steps for the pima data. Somehow the *active learning* and *multi-labeler* components help each other in improving classification performance. When one component is missing, depending on the data characteristics sometimes *active learning+majority vote* is better than *random sample+multi-labeler* or vice versa. Not surprisingly, *random sample+majority vote* which does not utilize multi-labeler information nor select the most informative sample

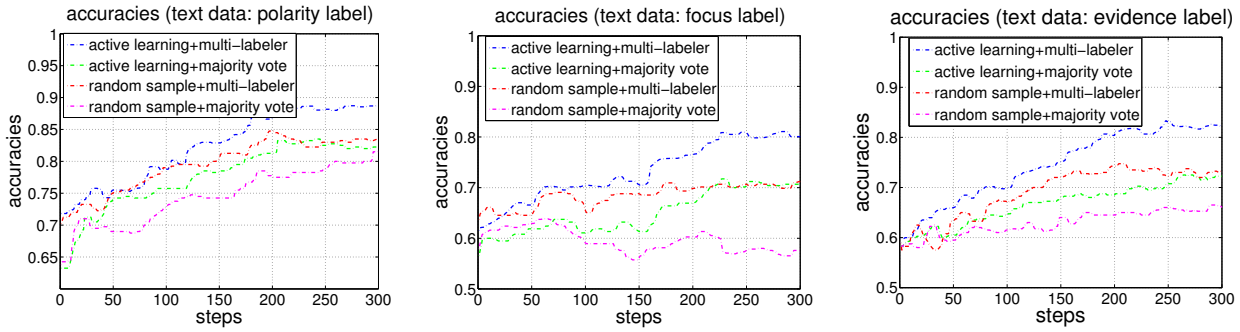

Figure 3. Accuracy comparisons on text data for the *polarity*, *focus* and the *evidence* labelings.

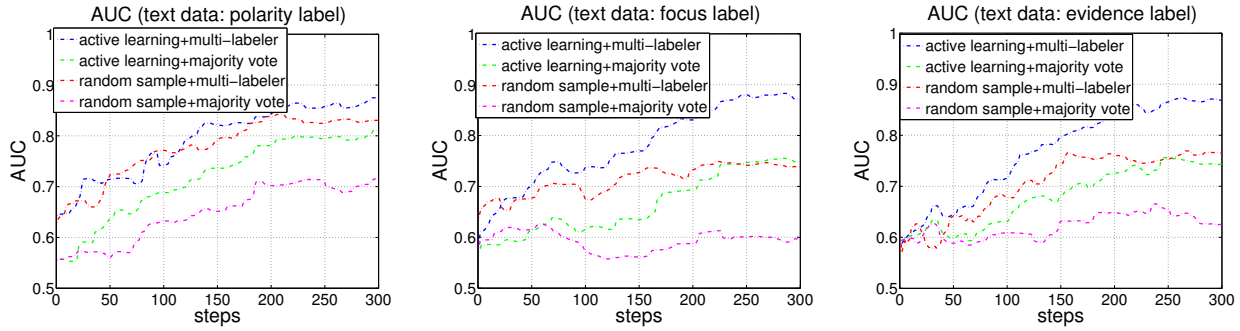

Figure 4. AUC comparisons on text data for the *polarity*, *focus* and the *evidence* labelings.

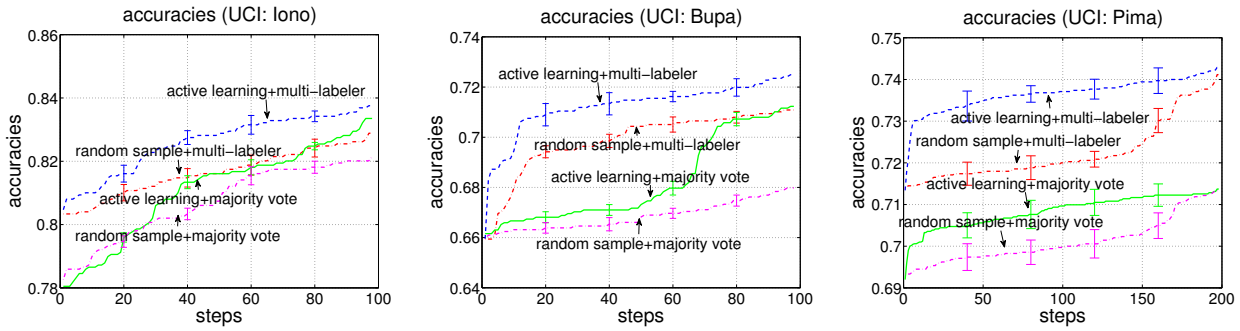

Figure 5. Accuracy comparisons on the Ionosphere, Bupa and Pima datasets.

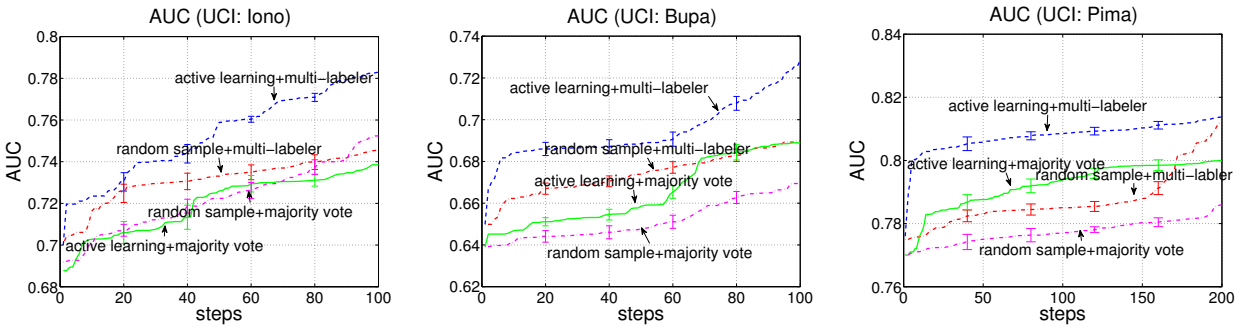

Figure 6. AUC comparisons on the Ionosphere, Bupa and Pima datasets.

lead to the worst performance.

6. Conclusions

Unlabeled data is most often abundant, but obtaining labels is expensive or time-consuming. Instead of simply labeling all the data or randomly selecting data to be labeled, the goal of active learning is to intelligently choose unlabelled instances to be labeled by an oracle to achieve higher accuracy with as few training labels as possible. In the multiple annotator paradigm, an oracle, who knows the ground truth, no longer exists; instead, multiple labelers, with varying expertise, are available for querying. This paradigm posits new challenges to the active learning scenario. We must ask which data sample should be labeled next and which annotator should we query to benefit our learning model the most. We are not aware of previous approaches to address this active learning problem in the presence of multiple annotators. In this paper, we employ a probabilistic multi-labeler model that allows for learning from multiple annotators, whose expertise across the data space may vary. We provide an optimization formulation that allows us to select the most uncertain sample and the annotator to query the labels from for active learning. Experiments on multiple annotator text data and on three UCI benchmark data show that our active learning approach together with taking advantage of information from multiple annotators clearly improves upon the learning rates (and performance) of baseline methods.

Acknowledgments: We thank NIH HHSN276201000029C for supporting this research.

References

Asuncion, A. and Newman, D.J. UCI machine learning repository, 2007.

Cohn, D., Ghahramani, Z., and Jordan, M. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

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

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

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

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

Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

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

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

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

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

McCallum, A. and Nigam, K. Employing EM in pool-based active learning for text classification. In *International Conference on Machine Learning*, pp. 350–358, 1998.

Newman, D.J., Hettich, S., Blake, C.L., and Merz, C.J. UCI repository of machine learning databases, 1998.

Nocedal, Jorge and Wright, Stephen. *Numerical Optimization (2nd ed.)*. Springer-Verlag, Berlin, New York, 2003.

Odewahn, S., Stockwell, E., Pennington, R., Hummphreys, R., and Zumach, W. Automated star/ galaxy discrimination with neural networks. *Astronomical J.*, 103(1):318–331, 1992.

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

Raykar, V. C., Yu, S., Zhao, L., Jerebko, A., Florin, C., Hermosillo-Valadez, G., Bogoni, L., and Moy, L. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *International Conference on Machine Learning*, pp. 889–896, 2009.

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

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

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

Sheng, V. S., Provost, F., and Ipeirotis, P. G. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Knowledge Discovery and Data Mining (KDD)*, pp. 614–622, 2008.

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

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pp. 254–263, 2008.

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