Proactive Learning with Multiple Class-Sensitive Labelers

Seungwhan Moon
Language Techonology Institute
Carnegie Mellon University
Pittsburgh, PA, 15213
Email: seungwhm@cs.cmu.edu

Jaime G. Carbonell
Language Techonology Institute
Carnegie Mellon University
Pittsburgh, PA, 15213
Email: jgc@cs.cmu.edu

***Abstract*—Proactive learning extends active learning by considering multiple labelers with different accuracies and costs, thus optimizing labeler selection as well as instance selection. In this paper, we propose a novel method to estimate labeler accuracy per class and to select labelers based on both cost and estimated accuracy, combined with an ensemble approach called *multi-class information density* (MCID) as a selection criterion. Our approach relaxes the common assumption found in past work that labeler accuracy is independent of class for multi-class learning, and by estimating the class-conditional accuracy better assigns instances to labelers. Results on several datasets with both real and simulated experts strongly demonstrate the efficacy of these methods.**

I. Introduction

The challenge in many machine learning or data mining tasks is that while unlabeled instances are abundant, acquiring their class labels often requires extensive human effort. The active learning paradigm addresses the challenge of insufficient labels by interactively optimizing the selection of queries [1], [2]. Several studies have shown that active learning reduces the sample complexity in a variety of applications, including network / graph analysis [3], text mining [4], [5], etc. However, active learning relies on tacit assumptions which prove limiting for real problems and applications. Primarily, active learning assumes the existence of a single omniscient labeling "oracle", whereas in real life it is more common to have multiple sources of annotations with different reliabilities or areas of expertise. In addition, active learning assumes that labeling different instances incurs uniform cost, regardless of the difficulty or the expected accuracy inherent in each annotation task. Some research has addressed cost-sensitive active learning, but only with respect to instances and features [6], [7].

Proactive learning has been proposed as a means to relax the unrealistic assumption of a single omniscient labeling oracle, permitting multiple labelers with different accuracies, different availabilities and different costs [8], [9], [10], [11]. This line of work has shown that proactive learning extends its reach to practical applications by combining estimation (learning) of the labeler accuracy with maximum utility of the labeler and instance selection. However, the prior work assumes that labeler accuracy is independent of labels in multi-class problems, calculating only an average accuracy across labels. In this paper, we address this limitation by explicitly estimating the dependency between label and labeler in order to optimize the assignment of new instances to labelers based on class priors or best-estimate of class membership.

We illustrate the novel contribution of the proposed method in the following example. Consider, for instance, the multi-class problem of a medical diagnosis of a patient with a disease that we know very little about (uncertainty in data): given multiple physicians who specialize in completely different areas (e.g. an oncologist, a cardiologist, or an internal medicine doctor), we need to assign one of the experts to diagnose the patient correctly (proactive learning selection). Querying an expert who has the best overall diagnosis performance across multiple domains (assuming uniform reliability across classes) may not give the most accurate results (analogous to the previous proactive learning methods [9]), because the chosen expert might lack knowledge in the specific disease of the patient. If we know that the patient has seemingly cancer symptoms (posterior class probability) and that an oncologist usually has the deepest understanding of cancer issues (estimated labeler accuracy given a specific class), we can leverage this information to better delegate a task to its respective expert. Note that this is indeed a real world challenge, as shown in our experiment with the Diabetes dataset in Section III.

Similarly, our new method estimates labeler accuracy on a per-class basis, or per subset-of-classes basis, providing a considerably new level of flexibility in proactive learning. A probabilistic approach to model annotator accuracy in a binary classification task was proposed by [12], but modeling annotator performance over the entire data in multi-classes is a complex task that requires a large number of training examples. Our method efficiently reduces the cost and complexity involved in estimating the labeler accuracy over multiple classes by employing the reduced per-class estimation method.

Another approach proposed by [13], [14], [15] that handles unreliable annotators in crowdsourcing scenarios is to query multiple annotators repetitively to estimate the ground truth label for each instance. The integration of judgements from crowd is typically done via majority vote or selective sampling, but these work do not comprise estimating individual per-task (per-class) expertise for selective recruitment of crowd members. In addition, these methods are not desirable in active learning scenarios because querying multiple annotators repeatedly for a single instance incurs multiple costs, whereas our method tries to find the one most cost-efficient expert who can answer the query reliably.

The framework that we propose is flexible and can work with any instance selection criterion or any supervised learning method. In order to further improve the effect of the proposed algorithm, we also propose a new density-based sampling strategy for multi-classes that considers the concept of *conflictivity* of the label distribution. We integrate our metrics as an ensemble method [16], [17], [18], and show that our selection criterion outperforms the traditional density-weighted sampling methods [19], [20], especially when there are multiple unreliable annotators.

The rest of the paper is organized as follows: Section II describes in detail the proposed proactive learning framework and presents the new density-based sampling strategy. The empirical results are reported and analyzed in Section III, and we give our concluding remarks and proposed future work in Section IV.

II. Method

In this section, we present a proactive learning method for multi-classification tasks when multiple domain experts are present. In our scenario, we assume that there exist multiple narrow experts and one meta oracle. A narrow expert has expertise in the subset of classes from the data, and each expert's expertise may or may not overlap. The probability of getting a correct answer given a query depends on the difficulty of the classification task for the expert. In other words, a narrow expert is more reliable in annotating the data for which the ground truth labels are within the expert's expertise. A meta oracle, on the other hand, has expertise in every category. The cost of each expert or a meta oracle varies depending on the difficulty of the task, the skewness of the data, and its range of expertise areas. We experiment with various combinations of cost ratios to simulate different real-world situations.

A. Proactive Learning with Multiple Domain Experts

In proactive learning, we jointly select the optimal oracle and the instance at which the current system's performance would best improve. As such, the solution to the problem is casted as a utility maximization subject to a budget constraint [8]. The objective of the problem can thus be formulated as:

$$\begin{aligned} &\max_{S \subset UL} E[V(S)] - \lambda(\sum_k t_k \cdot C_k) \\ &\text{s.t. } \sum_k t_k \cdot C_k \leq B, \sum_k t_k = |S| \end{aligned} \tag{1}$$

where S is the set of instances to be sampled, UL is the set of unlabeled samples, and $E[V(S)]$ is the expected value of information of the sampled data to the learning algorithm. $V(S)$ may be replaced by any active learning selection criterion, such as the uncertainty-based sampling [21]. $k \in K$ denotes the chosen oracle from the set of experts, and λ is a weighting parameter that determines how much the value of information is penalized by the oracle cost. C_k and t_k refer to the cost of the chosen expert k and the number of times it is queried, respectively. B is the total amount of budget for querying oracles. However, Equation 1 is a complex optimization problem because the learning function is updated at every iteration while the samples to be queried and their labels are unknown to the learner. Therefore, we employ a greedy approximation of the problem which chooses a small batch of samples to be queried at every iteration that maximizes the utility under the budget constraint:

Algorithm 1 Proctive Learning with Multiple Experts

Input: a multiclass classifier f, the pre-defined set of classes $\mathcal{C}$, labeled data L, unlabeled data UL, budget B, oracles $k \in K$ with cost C_k, each with expertise in some classes
Output: f
Obtain $P(ans|y = c, k)$, $\forall c \in \mathcal{C}, k \in K$ from Algorithm 2
Let C_T be the cost spent so far, $C_T = 0$
while $C_T < B$ **do**
  Train f on L
  Choose $(x^*, k^*) = \underset{k \in K, x \in UL}{\text{argmax}}\ U(x, k)$ (Eq. 3)
  Query the label $y^* = query(x^*, k^*)$
  $L = L \cup \{(x^*, y^*)\}$, $UL = UL - \{(x^*, y^*)\}$
  $C_T = C_T + C_k$
end while

$$(x^*, k^*) = \underset{x \in UL, k \in K}{\text{argmax}}\ U(x, k) \tag{2}$$

where $U(x, k)$ refers to a utility score when a sample x is annotated by an oracle k. We define the utility score such that it incorporates the reliability and the cost of an oracle as well as the base value of information of an instance. This ensures that the learner does not always choose the most reliable, and the most costly oracle, but encourages the learner to select the most cost-effective pair of an instance and an expert that is likely to give a correct answer. Thus, we can formulate the utility score as follows:

$$U(x, k) = \frac{V(x) \cdot P(ans|x, k)}{C_k} \text{ for } k \in K \tag{3}$$

where $V(x)$ is the value of the information of the sampled data to the learner, and $P(ans|x, k)$ is the probability of receiving the correct answer from an expert k given the sample x. We therefore assign a higher utility for the instances that have a higher value of information and a higher probability of being labeled correctly, while having a cheaper cost of annotation. Algorithm 1 describes the cost-optimized proactive learning process using the utility function formulated above. In most of the real world datasets, however, the accuracy information of the labeling sources $P(ans|x, k)$ is not given to the learner, and thus it needs to be estimated prior to the active selection process. While there may be various ways to estimate the accuracy of the labeling sources, the challenge is to minimize the number of queries that need to be made to each expert when there does not exist any query history a priori. The next section describes an efficient implementation of estimating expertise of labeling sources through selective sampling and the reduced per-class estimation.

B. Expertise Estimation

We assume that the oracle's expertise is distinctly aligned over a subset of classes rather than over the entire distribution of the data. We can then reduce the estimated labeling accuracy

Algorithm 2 Expertise Estimation for Multiple Experts

Input: Labeled data L, unlabeled data UL, oracles $k \in K$ each with expertise in some of the classes
Output: $P(ans|y = c, k) \ \forall c \in \mathcal{C},\ k \in K$
if $|L| > 0$ for each $c \in \mathcal{C}$ **then**
  for each x and ground truth label $(x, z) \in L$ **do**
    for each $k \in K$ **do**
      $y^{(k)} = query(x, k)$
      Update $P(ans|k, y = c)$ with $h(y^{(k)}, z)$
    end for
  end for
else
  Choose n samples randomly from UL
  for each sample x **do**
    Initialize $v(c) = 0\ \forall c \in C$
    for each $k \in K$ **do**
      $y^{(k)} = query(x, k)$
      $v(y^{(k)}) = v(y^{(k)}) + 1$
    end for
    $y^{maj} = \max_{c \in \mathcal{C}} v(c)$
    Set $P(ans|k, y = y^{maj})$ with $h(y^{(k)}, y^{maj})\ \forall k$
  end for
end if

$P(ans|x, k)$ as follows:

$$E[P(ans|x, k)] = \sum_{c \in \mathcal{C}} P(y = c|x) \cdot P(ans|k, y = c) \quad (4)$$

where $\mathcal{C}$ is the set of categories in a multi-classification task, $P(y = c|x)$ is the class posterior probability of the label for the sample x being c (predicted by the learner), which is an estimate of the true underlying label density. $P(ans|k, y = c)$ is the estimated probability of the expert k answering correctly for the label c. In other words, we integrate the learner's prediction of an instance with the expert's class-wide labeling accuracy. With the given formulation of $P(ans|x, k)$, the utility function favors the samples that have a higher probability of belonging to a certain label c, which an expert k is has expertise in. The meta-oracle will almost always have a higher value for $P(ans|x, k)$, but the overall utility will be dampened by a higher C_k as in Equation 3.

We consider two different scenarios for estimating $P(ans|k, y = c)$ (detailed in Algorithm 2): (1) when there are labeled samples already available (assuming ground truth), and (2) when there is no labeled sample at all. If we are given the ground-truth labels for n instances, we inquire for the labels of those instances to each expert and compute the labeler accuracy per class with the available ground-truth labels. Therefore, we define the empirical labeler accuracy per class as follows:

$$P(ans|k, y = c) = \frac{1}{n} \sum_{i=1}^{n} h(y_i^{(k)}, z_i)\ \forall k \in K \quad (5)$$

where $y_i^{(k)}$ is the prediction of x_i by an expert k, z_i is the ground-truth label of x_i, and $h(y_i^{(k)}, z_i) \in \{1, 0\}$ is an indicator function which is equal to 1 if $y_i^{(k)} = z_i$ and 0 otherwise. When there is no labeled sample available, we choose n samples from the unlabeled set, and inquire for the label of each sample to every expert. We estimate the ground-truth label of each instance by majority vote on experts ($= y^{maj}$), and compute $P(ans|k, y = c)$ with $h(y_i^{(k)}, y_i^{maj})$. Note that $P(ans|k, y = c)$ is independent of x, which thus gives only a brief class-sensitive belief about the expert's labeling accuracy. This simplified estimation allows for practical benefits in estimating labeler accuracy given the limited budget for the expertise discovery phase. In our experiments (Section III-B), we show that this brief knowledge of class-wide expertise greatly improves the performance when incorporated into the active learning selection formula. We also present the empirical analysis of the performance for varying degrees of errors for the estimated expertise.

C. Density-based Sampling for Multi-classification Tasks

While our framework is flexible and can work with any selection strategies, we propose a new density-based sampling method for multiple classes with multiple imperfect oracles to further improve the effect of the proposed algorithm.

Some of the most notable work done on the density-weighted uncertainty sampling (DWUS) strategies for active learning include the pre-clustering method [19], [20], which incorporates the prior density $p(x)$ of the data in the selection criterion. This method encourages the selection of more representative samples (e.g. centroids of denser clusters) at each query iteration, and avoids repetitively querying the samples that are in the same cluster.

We extend the previous work to accommodate for a multi-classification problem where labels are acquired from unreliable experts. First of all, if the expert that labeled a sample in a cluster is not reliable, we should in fact encourage querying samples from that cluster until we obtain a more credible label. Second, if a cluster encompasses conflicting opinions, or a cluster is placed over the decision boundaries, we should encourage querying from that cluster to better tune the decision boundaries between neighboring classes.

As such, we propose a new multi-class information density (MCID) as follows, which comprises of three components: (1) density, (2) unknownness, and (3) conflictivity. The density component measures how densely samples are positioned around a given point, and the unknownness component measures how many samples are labeled thus far. The conflictivity component measures how heterogeneous the label distribution is around a given sample. The conflictivity term encourages the learner to favor a cluster that still has conflicting and unresolved class distribution over a slightly denser cluster with unanimous class distribution.

A simple yet efficient implementation of MCID is to pre-cluster the dataset and calculate the three components in each cluster locally. For a given cluster $q \in Q$, where Q refers to a set of clusters of the dataset and q is a set of labeled and unlabeled samples within the cluster, the MCID of a sample is defined as follows:

$$\begin{aligned} \rho(q, x) = p(x) \cdot \frac{|q_{UL}|}{|q|} \\ \cdot (-\sum_{c \in \mathcal{C}} P(y = c|q) \cdot \log P(y = c|q)) \end{aligned} \quad (6)$$

TABLE I. OVERVIEW OF DATASETS.

Dataset	# Experts	# Classes	Size
Diabetes 130 U.S. Hospitals	3	3	13300
20 Newsgroups	5	20	7000
Landsat Satellite	3	6	3000
Image Segmentation	3	7	2310
Vehicle	4	4	946

where $\rho(q, x)$ is the MCID of a sample x in a cluster q, $p(x)$ is the density at a point x, q_{UL} is a set of unlabeled samples within the cluster, and $\mathcal{C}$ is the set of label classes. We induce $p(x)$ using a $|Q|$ Gaussian mixture model with weights $P(q)$, hence $p(x) = \sum_{q \in Q} p(x|q)P(q)$, where $p(x|q)$ is a multivariate Gaussian sharing the same variance σ^2 [19]:

$$p(x|q) = (2\pi)^{-d/2}\sigma^{-d}\exp\{-\frac{||x - c_q||^2}{2\sigma^2}\} \tag{7}$$

where c_q is the centroid of the cluster q. We estimate the cluster prior $P(q)$ via an EM procedure:

$$\begin{aligned} P(q|x_i) &= \frac{P(q)\exp\{\frac{-||x_i - c_q||^2}{2\sigma^2}\}}{\sum_{q \in Q} P(q)'\exp\{\frac{-||x_i - c_q||^2}{2\sigma^2}\}} \\ P(q) &= \frac{1}{N}\sum_{i=1\ldots N} P(q|x_i) \end{aligned} \tag{8}$$

where N is the size of the sample set. The second term in Equation 6 measures the proportion of samples known at each iteration. The last term is the entropy of class distribution within the cluster, which approximates the conflictivity of the cluster.

Note that the MCID measure does not contain any knowledge about how informative each individual point is. Therefore, the ultimate value function of an instance is given as a combination of the basis selection criteria $\phi(x)$ (e.g. the uncertainty-based selection [21], etc.) and the MCID. Therefore:

$$V(x) = \phi(x) \cdot \rho(q, x)^{\beta} \tag{9}$$

where $\beta \in (-\infty, \infty)$ is a weight parameter. For simplicity, in the following experiments, we use $\beta = 1$ and $\phi(x) = H(x) = -\sum_{y \in Y} P(y|x) \cdot \log P(y|x)$, or the entropy of the probability distribution [22].

III. EXPERIMENTAL EVALUATION

Table I shows the summary of the datasets we used in our experiments. The Diabetes 130 U.S. Hospitals dataset [23] contains the attributes that identify the medical specialty of each annotator, as well as the specific diagnosis type (label) and medical records (attributes) of each patient instance. For our experiment, we make a subset of the dataset by choosing the three frequent diagnosis types, and consider three major medical specialties (Internal Medicine, Family/General Practice, Surgery-General). Each instance is annotated by only one annotator with a single medical specialty, and therefore we assume that labels we query come from an expert classifier model which is trained over the instances that each respective expert has annotated.

The rest of the datasets in Table I do not have any annotator information, and thus we simulate multiple narrow experts as follows. We assume that the narrow experts' expertise does not overlap but together they cover every category. For example, we train 5 narrow experts for the 20 Newsgroups dataset, each specializing in 4 (=20/5) unique classes (See Table I). In order to simulate the reliability of the oracles with different expertise, we assume that a narrow expert resembles a classifier trained on the dataset of which the labels of the samples in its non-expertise categories are partially noised. The noise ratio was adjusted so that the overall labeling accuracy is around 50% for each non-expertise category. The meta oracle is trained on the entire dataset without any artificial noise. This simulates a realistic situation where every annotator has a varying degree of non-zero error rates on different classes. The results are averaged over 10 runs for every experiment.

TABLE II. COMPARISON OF ERROR RATES OF MCID VS DWUS VS US

Dataset	Cost	Classification Error Rates		
		MCID	DWUS	US
Diabetes	0.25	0.402	0.411	0.423
	0.50	0.374*	0.407	0.409
	0.75	0.362*	0.399	0.400
	1.00	0.354*	0.393	0.398
20 Newsgroups	0.25	0.508	0.521	0.516
	0.50	0.431*	0.470	0.488
	0.75	0.388*	0.428	0.453
	1.00	0.350*	0.381	0.388
Vehicle	0.25	0.333	0.335	0.350
	0.50	0.281	0.294	0.301
	0.75	0.260*	0.279	0.286
	1.00	0.242*	0.266	0.271

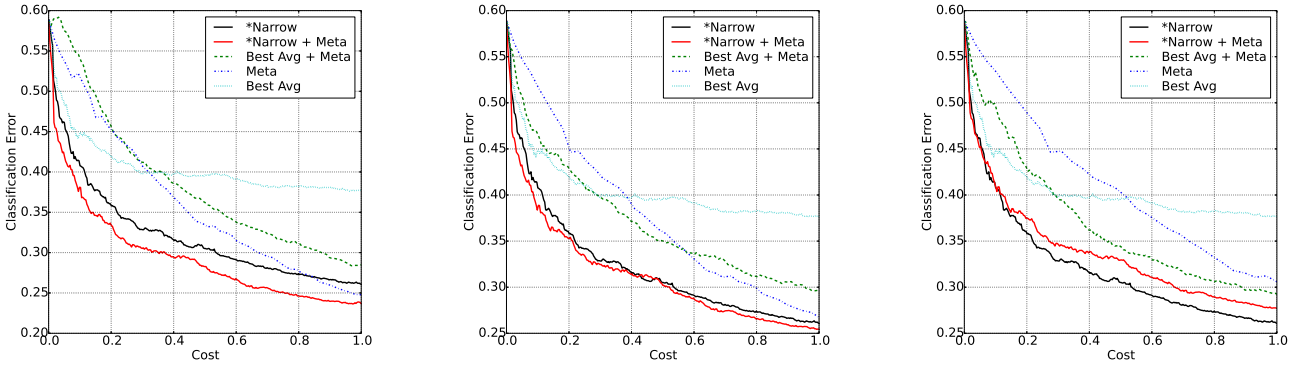
A. Multi-class Information Density

We compared the proposed multi-class information density (MCID) method (detailed in Section II-C) on several datasets with two other baseline selection criteria: (1) *US*, which uses the traditional uncertainty sampling (US) method, (2) *DWUS*, which employs the widely used density-only weighted uncertainty sampling (DWUS) method [19], [20].

Table II shows the classification error rates at four different stages of active learning (cost = 0.25, 0.50, 0.75, 1.0), where each label is obtained from a randomly chosen narrow expert to allow for a realistic and heterogeneous label distribution. Both *DWUS* and *MCID* methods outperform the baseline (*US*), which greatly saves the annotation cost to converge. There is a time-variant performance difference on these two methods: the *DWUS* method performs almost the same as the *MCID* method at the beginning, which indicates that the conflictivity component of the measurement does not improve the performance when not enough labels are given. Once enough labels are given (cost $\geq$ 0.5), the *MCID* method outperforms the previous density-only weighted baseline (*DWUS*). Statistically significant improvements ($p < .05$) over the baselines at each cost are marked as *.

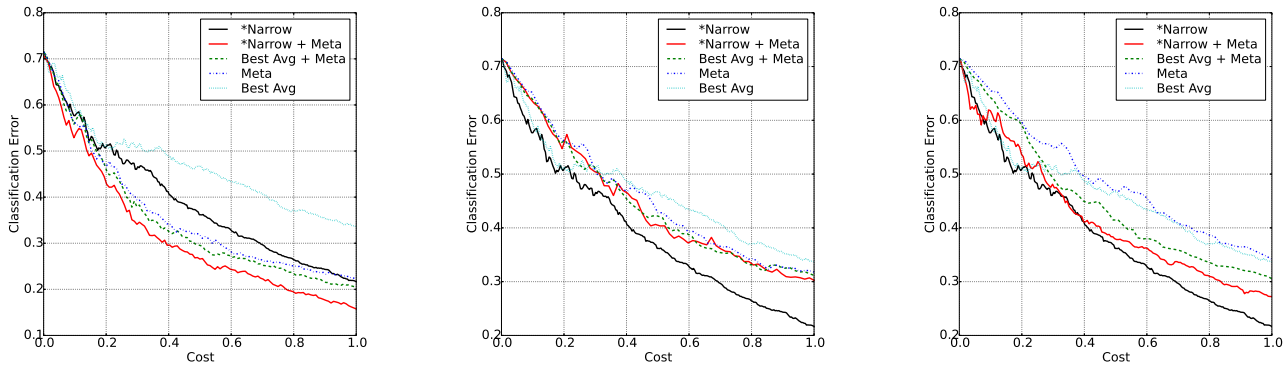
B. Multiple Experts

The following figures show the results for the proposed proactive learning algorithm on five different datasets: Diabetes 130 U.S. Hospitals, 20 Newsgroups, UCI Landsat Satellite, UCI Statlog Image Segmentation, and UCI Vehicle. For each dataset, we vary the cost ratio of a narrow expert to the meta



(a) Cost Ratio = 1:6 (b) Cost Ratio = 1:8 (c) Cost Ratio = 1:11

Fig. 1. Comparison of error rates on the Diabetes 130 U.S. Hospitals Dataset with different cost ratios (when expertise was estimated via ground truth samples). The X-axis denotes the normalized total cost, and the Y-axis denotes the classification error. Our proposed methods are marked as * in the legends.



(a) Cost Ratio = 1:2 (b) Cost Ratio = 1:4 (c) Cost Ratio = 1:5

Fig. 2. Comparison of error rates on the 20 Newsgroup Dataset with different cost ratios (when expertise was estimated via ground truth samples). The X-axis denotes the normalized total cost, and the Y-axis denotes the classification error. Our proposed methods are marked as * in the legends.

oracle, the initial number of labeled / unlabeled samples to estimate $P(ans|k, y = c)$, and the proactive learning methods.

Figures 1 and 2 show the performance of the proposed algorithm with varying cost ratios on the Diabetes dataset and the 20 Newsgroups dataset, respectively. Due to space constraints, we present the rest of the results in Tables III and IV.

There are three baselines that were considered: (1) *Best Avg* (Dotted Cyan), where the learner always asks one of the narrow experts that has the highest average $P(ans|x, k)$ under the assumption that labelers are uniformly accurate across multiple classes, (2) *Meta* (Dashdot Blue), where for a higher price, the learner always asks the meta oracle that has expertise in every category, and (3) *Best Avg+Meta* (Dashed Green), which jointly chooses between the more reliable meta oracle and the fallible narrow expert under the uniform reliability assumption. Note that the baseline (3) refers to the proactive learning method proposed earlier by [9]. For all of the baseline methods, we use our proposed MCID method as a criterion for the instance selection.

The two proactive learning methods that we propose (marked as *) are: (1) *Narrow* (Solid Black), where the learner selectively chooses the best pair of a sample and a narrow expert that yields the highest utility at each iteration, and (2) *Narrow+Meta* (Solid Red), which also includes the meta oracle in the pool of experts. We show that each proposed method has an advantage over each other depending on the availability and the affordability of the meta oracle.

In all of our experiments, *Narrow* and *Narrow+Meta* significantly outperform the *Best Avg* baseline. When the meta oracle is expensive (as in Figures 1(c), 2(b), 2(c)), *Narrow* significantly outperform the *Meta* and the *Best Avg+Meta* baseline ($p < .01$). In reality, the meta oracle would be significantly more expensive than the narrow experts or it may not exist at all. The results are thus promising because the proposed method can perform very well even in the absence of the meta oracle. When the meta oracle is cheaper (Figures 1(a), 2(a)), on the other hand, the joint *Narrow+Meta* method outperforms both the *Meta* baseline and the *Narrow* method ($p < .01$), which indicates that the proposed algorithm jointly optimizes between the meta oracle and the narrow experts in the most cost-efficient way. While the joint *Best Avg+Meta* baseline [8] outperforms the other two baselines when the cost ratio is high, the improvement is not as significant as

TABLE III. COMPARISON OF ERROR RATES ON THE UCI DATASETS (WHEN EXPERTISE WAS ESTIMATED VIA GROUND TRUTH SAMPLES)

Dataset	Cost Ratio	Cost	Classification Error				
			***Narrow**	***Narrow +Meta**	Best Avg +Meta	Meta	Best Avg
Landsat Satellite	1:2	0.25	0.311	0.329	0.335	0.331	0.326
		0.50	0.217	0.223	0.246	0.283	0.249
		0.75	0.133	0.155	0.166	0.237	0.195
		1.00	**0.069**	0.098	0.119	0.185	0.128
Image Segmentation	1:2	0.25	0.111	0.126	0.142	0.203	0.169
		0.50	0.064	0.080	0.081	0.060	0.130
		0.75	0.050	0.047	0.050	0.043	0.113
		1.00	0.045	**0.032**	0.041	**0.029**	0.118
Vehicle	1:1.5	0.25	0.302	0.260	0.262	0.252	0.325
		0.50	0.231	0.201	0.211	0.210	0.261
		0.75	0.166	0.141	0.179	0.168	0.229
		1.00	0.132	**0.103**	0.148	0.139	0.215

TABLE IV. COMPARISON OF ERROR RATES (WHEN EXPERTISE WAS ESTIMATED VIA MAJORITY VOTE)

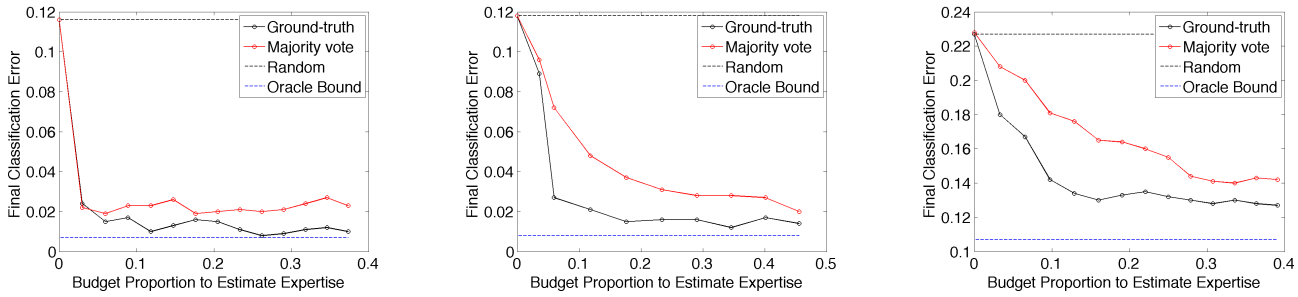
Dataset	Cost Ratio	Cost	Classification Error				
			***Narrow**	***Narrow +Meta**	Best Avg +Meta	Meta	Best Avg
Diabetes	1:11	0.25	0.355	0.375	0.430	0.471	0.421
		0.50	0.317	0.346	0.348	0.401	0.399
		0.75	0.276	0.301	0.306	0.343	0.372
		1.00	**0.269**	0.283	0.297	0.308	0.367
20 Newsgroups	1:5	0.25	0.461	0.490	0.501	0.559	0.521
		0.50	0.366	0.394	0.431	0.472	0.466
		0.75	0.289	0.335	0.343	0.397	0.395
		1.00	**0.221**	0.298	0.306	0.355	0.357
Landsat Satellite	1:2	0.25	0.321	0.331	0.351	0.329	0.334
		0.50	0.217	0.236	0.283	0.279	0.248
		0.75	0.138	0.184	0.223	0.234	0.199
		1.00	**0.078**	0.116	0.129	0.185	0.126
Image Segmentation	1:2	0.25	0.154	0.169	0.171	0.206	0.177
		0.50	0.087	0.060	0.062	0.062	0.135
		0.75	0.056	0.042	0.039	0.045	0.112
		1.00	0.042	**0.026**	**0.028**	**0.031**	0.118
Vehicle	1:1.5	0.25	0.301	0.251	0.246	0.250	0.326
		0.50	0.248	0.210	0.209	0.208	0.277
		0.75	0.182	0.167	0.170	0.169	0.263
		1.00	0.141	**0.112**	0.142	0.140	0.247

in our proposed methods for this experiment. This is because the previous work fails to capture the noisy labeler accuracy which varies by class. Tables III and IV show similar results on other UCI datasets. Note that the proposed algorithm works successfully even when there is no ground truth sample available to estimate expertise. While the ground truth case generally performs better than the majority vote estimation method, they eventually converge at almost the same accuracy level (Tables III and IV).

Figure 3 shows the difference in the final error rate at convergence as a function of the initial budget that was set aside to estimate expertise of each narrow expert. We assume that acquiring a label to estimate expertise incurs the same cost as querying an expert during the active learning process. If the learner spends a large enough budget to estimate expertise, it can more accurately delegate an instance to a narrow expert that has expertise for the chosen instance. *Ground truth* (Solid Black) represents the convergence accuracy when we employ the *Narrow* method, where the expertise was estimated using the ground truth samples with the marked proportion of the budget. *Majority vote* (Solid Red) refers to the *Narrow* method where the expertise was estimated using the majority vote method. As a baseline, we present the final accuracy when there is no prior knowledge of expertise, thus randomly choosing an expert at each iteration (Dotted Black). We also present an oracle bound (Dotted Blue), where we assume that we have perfect estimation of expertise of each expert, thus delegating an instance to the correct expert every time. For all of the UCI datasets that were tested, the results show that the proposed method works significantly better than the baseline even with a limited budget to estimate expertise. This result shows that even with the imperfect estimation of expertise we can still improve the performance greatly. The *ground truth* method utilizes improved estimation of expertise, thus outperforming the *majority vote* method.

IV. CONCLUSION

The novel contributions of the this work are as follows: we proposed an efficient proactive learning algorithm for which there are multiple class-sensitive experts with varying costs whose expertise are distinctly aligned over multiple classes. The proposed method formulates a cost-driven decision frame-



(a) Landsat Satellite (b) Image Segmentation (c) Vehicle

Fig. 3. Final error rate at convergence as a function of the initial budgets set aside for expertise estimation on the UCI datasets, in proportion to the total budget spent to acquire labels until it reaches convergence.

work which maximizes the utility across oracle-example pairs. We showed that our algorithm efficiently delegates each narrow expert to an unlabeled instance that the chosen expert is most likely to have expertise in. The empirical results on the datasets with both real and simulated experts demonstrate the effectiveness of this approach under different cost conditions. Specifically, when there exists an affordable meta oracle, the proposed algorithm jointly optimizes between the meta oracle and the narrow experts. Our approach works sufficiently well even with the imperfect estimation of expertise due to a limited budget. We also implemented a new density metric for multi-class classification which considers the *conflictivity* of the label distribution. The result shows an improvement over the traditional density-only-weighted method, especially when the annotators are not reliable.

This paper allows for a practical and efficient application of active learning in real world tasks such as crowdsourcing or data mining. To continue this work, we will investigate the theoretical min-max bounds of the proposed algorithm under different reliabilities and costs of the experts. We will also extend our work to a new crowdsourcing scenario with a larger pool of experts, where the challenge is to efficiently estimate the labeler expertise as a group, rather than as individuals. We also plan on formulating the theoretical condition in which the conflictivity metric would improve the performance over the density-only-weighted method.

REFERENCES

[1] D. Lewis and W. Gale, "Training text classifiers by uncertainty sampling," in *Proceedings of ACM-SIGIR Conference on Information Retrieval*, 1994.

[2] N. Roy and A. Mccallum, "Toward optimal active learning through sampling estimation of error reduction," in *In Proc. 18th International Conf. on Machine Learning*, pp. 441 – 448, Morgan Kaufmann, 2001.

[3] M. Bilgic, L. Mihalkova, and L. Getoor, "Active learning for networked data," *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[4] A. McCallum and K. Nigam, "Employing EM and pool-based active learning for text classification," *ICML 98*, pp. 359 – 367, 2001.

[5] V. Ambati, S. Vogel, and J. G. Carbonell, "Active learning and crowd-sourcing for machine translation," *LREC 10*, 2010.

[6] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney, "Economical active feature-value acquisition through expected utility estimation," *KDD 05 Workshop on Utility-based data mining*, 2005.

[7] S. Moon, C. McCarter, and Y.-H. Kuo, "Active learning with partially featured data," in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pp. 1143–1148, 2014.

[8] P. Donmez and J. G. Carbonell, "Proactive learning: Cost-sensitive active learning with multiple imperfect oracles," *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 2008.

[9] P. Donmez and J. G. Carbonell, "From Active to Proactive Learning," *Advances in Machine Learning I*, vol. 262, pp. 97 – 120, 2010.

[10] L. Yang and J. Carbonell, "Cost complexity of proactive learning via a reduction to realizable active learning," *Tech report CMU-ML-09-113*, 2010.

[11] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Who should label what? instance allocation in multiple expert active learning," in *In Proc. of the SIAM International Conference on Data Mining (SDM*, 2011.

[12] Y. Yan, R. Rosales, G. Fung, and J. Dy, "Active Learning from Crowds," *Proceedings of the 28th International Conference on Machine Learning*, pp. 1161 – 1168, 2011.

[13] P. Dai, Mausam, and D. S. Weld, "Artificial intelligence for artificial artificial intelligence," *AAAI*, 2011.

[14] P. Viappiani, S. Zilles, H. J. Hamilton, and C. Boutilier, "Learning complex concepts using crowdsourcing: A bayesian approach," *Algorithmic Decision Theory*, 2011.

[15] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima, "Accurate Integration of Crowdsourced Labels Using Workers' Self reported Confidence Scores," *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[16] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," *ICML 03*, pp. 19 – 26, 2003.

[17] P. Donmez, J. G. Carbonell, and P. Bennett, "Dual strategy active learning," *In Proceedings of the European Conference on Machine Learning*, pp. 116 – 127, 2007.

[18] P. Melville and R. Mooney, "Diverse ensembles for active learning," *International Conference on Machine Learning (ICML) '04*, 2004.

[19] H. Nguyen and A. Smeulders, "Active learning using pre-clustering," *International Conference on Machine Learning (ICML)*, 2004.

[20] J. Zhu, H. Wang, B. Tsou, and M. Ma, "Active Learning With Sampling by Uncertainty and Density for Data Annotations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, 2010.

[21] D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," *Proceedings of the International Conference on Machine Learning (ICML) '94*, pp. 148 – 156, 1994.

[22] B. Settles and M. Craven, "Training text classifiers by uncertainty sampling," *In Proceedings of the Conference on Empirical Methods in Nat ural Language Processing (EMNLP)*, pp. 1069 – 1078, 2008.

[23] B. Strack, J. Olmo, DeShazo, C. Jennings, K. Cios, and J. Clore, "Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," *BioMed Research International*, 2014.