Asking the Right Questions to the Right Users: Active Learning with Imperfect Oracles

Shayok Chakraborty
Department of Computer Science
Florida State University

Abstract

Active learning algorithms automatically identify the salient and exemplar samples from large amounts of unlabeled data and tremendously reduce human annotation effort in inducing a machine learning model. In a traditional active learning setup, the labeling oracles are assumed to be infallible, that is, they always provide correct answers (in terms of class labels) to the queried unlabeled instances. However, in real-world applications, oracles are often imperfect and provide incorrect label annotations. Oracles also have diverse expertise and while they may be noisy, certain oracles may provide accurate annotations to certain specific instances. In this paper, we propose a novel framework to address the challenging problem of active learning in the presence of multiple imperfect oracles. We pose the optimal sample and oracle selection as a constrained optimization problem and derive a linear programming relaxation to select a batch of (sample-oracle) pairs, which can potentially augment maximal information to the underlying classification model. Our extensive empirical studies on 9 challenging datasets (from a variety of application domains) corroborate the usefulness of our framework over competing baselines.

Introduction

Supervised learning algorithms require a large amount of labeled data to induce a reliable model. However, while unlabeled data is cheap and easily available, obtaining class labels requires extensive human effort, often from experts with very limited availability. Thus, developing intelligent learning algorithms under the constraint of weak human supervision has attracted significant research attention. *Active Learning (AL)* algorithms address this challenge by automatically identifying the salient and exemplar samples from large amounts of unlabeled data. This tremendously reduces human annotation effort as only a few samples, that are identified by the algorithm, need to be labeled manually. Further, since the model gets trained on the highly informative samples from the underlying data population, it typically depicts better generalization capability than a passive learner, where training samples are obtained at random from the data population. In a serial query AL setting, the learner queries only a single instance in each iteration, which is annotated and appended to the training set. This results in frequent model updates and also cannot exploit the presence of multiple labeling oracles. In batch mode active learning (BMAL), the learner queries batches of unlabeled samples and can leverage the availability of multiple labeling agents. Common applications of AL include computer vision (Chakraborty et al. 2015), text mining (Tong and Koller 2001), spam filtering (Sculley 2007) and bioinformatics (Osmanbeyoglu et al. 2010) among others.

Traditional active learning algorithms assume the labeling oracles to be infallible, i.e. they always provide correct answers to the queried instances. This assumption may not hold for several real-world applications, where we usually have multiple labelers providing different qualities of annotation. For instance, with the advent of crowdsourcing platforms like the Amazon Mechanical Turk (AMT), hundreds of annotators are available over the Internet to provide varying degrees of noisy annotations. As another example, consider a medical image classification application, where we have expert doctors as well as novice residents providing varying degrees of noisy labels. In such situations, some annotators may be more reliable (and hence more expensive) than others; there may exist different prior knowledge about annotators; more importantly, annotator effectiveness may vary depending on the specific data instance considered. This diversity among annotators has rendered typical supervised learning algorithms sub-optimal and has motivated the development of algorithms that are annotator-aware (Urner, Ben-David, and Shamir 2012).

In this paper, we propose a novel framework to address the problem of batch mode active learning (which queries a batch of unlabeled samples simultaneously) in the presence of multiple imperfect labeling oracles. Our objective is to select a batch of exemplar unlabeled samples and identify the optimal labeling oracle for each queried sample. We pose the sample selection based on the uncertainty and redundancy criteria and the oracle selection based on the labeling cost and the error probability conditions. The active sample and oracle selection problem is solved using a single optimization framework to select a batch of informative (sample-oracle) pairs. By utilizing the presence of multiple

labeling oracles that can label samples simultaneously, but provide imperfect annotations, this paper takes a step towards filling in a gap between active learning and real-world tasks, to make active learning reach practical applications.

Related Work

In this section, we present a brief survey of active learning (AL) algorithms in general, followed by a survey of AL in the presence of noisy oracles.

Active Learning: Active Learning is a well-studied problem in machine learning (Settles 2010). In a typical pool-based setup, the learner is exposed to a pool of unlabeled samples and it iteratively queries batches of informative samples for manual annotation. The most common query strategy in active learning is uncertainty sampling, where unlabeled samples with the highest classification uncertainties are queried for annotation. The uncertainty of an unlabeled sample can be quantified by its Shannon's entropy (Holub, Perona, and Burl 2008), its distance from the decision boundary in the feature space for SVM classifiers (Tong and Koller 2001), the disagreement among a committee of classifiers about the label of the sample (Freund et al. 1997) and also by combining multiple criteria such as uncertainty, representativeness and diversity (Shen et al. 2004). The Fisher information matrix has also been exploited to compute classification uncertainty and to develop active learning algorithms (Hoi et al. 2008). Guo and Schuurmans (Guo and Schuurmans 2007) proposed a discriminative AL strategy where the active sampling criterion was based on the uncertainty of the future learner. Guo further used a matrix partitioning approach to develop an AL algorithm independent of the underlying classification model (Guo 2010). Adaptive active learning schemes have been proposed where the goal is to automatically compute the batch size based on the complexity of a data stream (Chakraborty, Balasubramanian, and Panchanathan 2015). Deep active learning is a recently researched topic in this field, which bridges the ideas of deep learning and active learning; the objective is to automatically learn a discriminating feature set using deep neural networks and simultaneously identify the salient unlabeled samples for manual annotation (Ranganathan et al. 2017). Adversarial techniques using GANs have also been used for active learning (Zhu and Bento 2017) with promising results.

Active Learning with Imperfect Oracles: All the aforementioned AL techniques assume that the labeling oracles are infallible. A few research efforts have focused on the development of active learning algorithms in the presence of noisy annotators, where the goal is to select informative samples, as well as annotators for labeling them. The algorithms proposed in (Zhang and Chaudhuri 2015) (Donmez and Carbonell 2008) assumed the presence of two labeling oracles, one of which always returns the correct label and the other returns incorrect annotations with a fixed probability. Yan *et al.* (Yan, Chaudhuri, and Javidi 2016) proposed an AL algorithm using a single oracle, which can provide incorrect labels and can also abstain from labeling. These algorithms assume a very simplistic setting and do not generalize to multiple oracles with diverse expertise.

In the presence of multiple noisy annotators, a common strategy is to use relabeling, where an actively queried sample is relabeled multiple times using crowdsourcing and the final label is obtained using majority voting (Zhao, Sukthankar, and Sukthankar 2011). Zheng *et al.* (Zheng, Scott, and Deng 2010) addressed the problem of active learning with multiple labelers, where each labeler has a different (known) cost and a different (unknown) accuracy; the label was obtained from a subset of labelers using majority voting. However, the labelers were globally selected for all instances and not actively for every individual unlabeled sample. Along similar lines, (Ipeirotis et al. 2014) proposed a repeated labeling (re-labeling) strategy, together with majority voting and uncertainty-preserving labeling to integrate the information from multiple labels. Donmez *et al.* (Donmez, Carbonell, and Schneider 2009) selected a subset of confident oracles (based on the proximity of their upper confidence interval to the maximum upper confidence interval) and used majority voting on the selected oracles to label each queried sample. Ambati *et al.* (Ambati, Vogel, and Carbonell 2010) proposed the Active Crowd Translation (ACT) framework for active learning using crowdsourcing to enable automatic translation for low-resource language pairs. A common drawback in all these methods is that the same unlabeled sample is labeled multiple times by different annotators (and the best label is selected using an aggregation mechanism), which results in sub-optimal usage of available resources. Fang *et al.* (Fang, Yin, and Tao 2014) proposed a knowledge transfer mechanism from auxiliary domains for computing labelers' expertise. However, the framework necessitated access to a different source dataset, which may not be readily available in real-world applications. Yan *et al.* (Yan et al. 2011) (Yan et al. 2012) proposed a probabilistic multi-labeler model to compute the accuracy of each labeler, and select the most confident labeler for each queried unlabeled sample. Huang *et al.* (Huang et al. 2017) proposed a Cost Effective Active Learning (CEAL) framework for active sample selection in the presence of multiple noisy oracles. However, even though these algorithms considered the presence of multiple noisy oracles, they queried only a single unlabeled instance in each AL iteration. This results in inefficient usage of labeling resources, as only a single annotator is being utilized at any given point of time; myopically extending the single instance selection to multi-instance selection produces sub-optimal results (as evidenced by our empirical studies).

In this paper, we propose a novel algorithm to address this challenging problem. Our framework queries a batch of informative samples simultaneously and identifies the optimal labeling oracle for each queried sample. Through its batch selection strategy, the framework can leverage the presence of multiple labeling oracles with varying degrees of imperfections - a situation commonly encountered in real-world applications. We now describe our framework.

Proposed Framework

Consider an active learning problem, where we are given a labeled training set L and an unlabeled set U ($|L| \ll |U|$). Let w be the model trained on L and Z be the number of

classes in the problem. We have access to k labeling oracles $\{O_1, O_2, \ldots O_k\}$ with corresponding cost $\{C_1, C_2, \ldots C_k\}$, denoting the price to be paid to get one unlabeled sample labeled. The oracles are imperfect and may provide incorrect annotations to queried samples. We are further given a labeling budget B. Our objective is to select a batch of unlabeled samples together with a set of oracles to label each sample, such that the total cost incurred equals the budget, and the selected samples with the provided labels augment maximal information to the classification model. To address this problem, we propose to perform active selection of both samples and oracles. These are detailed below.

Active Sample Selection

In order to identify the optimal set of samples to be queried, we need a metric to quantify the utility score of a batch of samples. In this research, we used the *informativeness* and *redundancy* criteria to compute the utility score. A sample selection framework driven by these two conditions ensures that the selected samples are individually informative and they have minimal redundancy (duplication) among them. Such criteria has been used in previous active learning research (Shen et al. 2004).

Computing informativeness: The informativeness of an unlabeled sample x_i was computed as the classification uncertainty of x_i using the model w. We used the Shannon's entropy to compute the uncertainty of an unlabeled sample in this research:

$$E(x_i) = -\sum_{j=1}^{Z} p_{ij} \log p_{ij} \tag{1}$$

where p_{ij} is the posterior probability of x_i with respect to class j, computed by the current model w. A high value of entropy denotes high classification uncertainty, and thus a more informative sample from an active learning perspective.

Computing redundancy: We also computed a redundancy matrix $R \in \Re^{|U|\times|U|}$, where R_{ij} denotes the redundancy between samples x_i and x_j in the unlabeled set. This is necessary to avoid selecting samples that are individually informative, but mutually redundant. We used the cosine similarity to quantify the redundancy between a pair of samples, where a low value of the similarity denotes low redundancy. We thresholded the similarity values at 0, so that R contains only non-negative entries. The matrix R was computed as follows:

$$R(i,j) = \min(0, cos(x_i, x_j)) \tag{2}$$

Active Oracle Selection

For a given unlabeled sample, we estimate the optimal labeling oracle based on two conditions: the *error probability* of the oracle on the unlabeled sample; and the overall *labeling cost* of the oracle. These are described below.

Computing Oracle Error Probability: As mentioned previously, in a real-world application labelers have varying knowledge and expertise and may provide noisy annotations to queried instances. However, even though the labelers are noisy and have diverse expertise, it is reasonable to assume that certain labelers may be good at labeling certain specific instances. For instance, an infant labeler may not be able to correctly label a large variety of images, but may label images of common animals accurately. In medical image classification, a novice resident may not provide accurate annotations to all samples, but may be able to correctly identify certain specific medical abnormalities. We therefore exploited a data-driven strategy to select the optimal oracle for a given unlabeled sample. Each oracle was asked to label the samples in the labeled set L. Since the ground truth labels in L are known, the errors committed by each oracle on L can be determined. A binary logistic regression (LR) classifier was then trained for each oracle separately to model their error patterns; given a particular unlabeled sample, the trained model returns the probability of committing a labeling error on the sample by the corresponding oracle. We denote the error probability of unlabeled sample x_i when labeled by oracle O_j as q_{ij}.

Computing Oracle Cost: The cost of an oracle denotes the price to be paid to purchase a label from that oracle. It is important to consider the labeling cost of an oracle because, if two oracles furnish approximately the same error probability on a particular unlabeled sample, the one with lower cost should be preferred. The cost is directly proportional to the reliability / accuracy of the oracle. The cost of oracle O_j was computed as:

$$C_j = \alpha A_j \tag{3}$$

where A_j is the accuracy of O_j on the labeled set L and α is the constant of proportionality. Depending on the application and domain knowledge, other strategies can be used to compute these terms.

Active (Sample-Oracle) Selection

Given $E(x_i)$, q_{ij} and C_j, we compute a matrix $P \in \Re^{k\times|U|}$ (k is the number of labeling oracles), where each column represents an unlabeled sample and each row represents an oracle. Our objective is to select a batch of unlabeled samples which furnish high entropy values (high uncertainties), and get them labeled by oracles which have low cost and furnish low labeling error probabilities for the corresponding samples. The matrix P is defined to capture all these conditions:

$$P(j,i) = \frac{q_{ij} * C_j}{E(x_i)}, \quad i = 1, \ldots |U|, \quad j = 1, \ldots k \tag{4}$$

Also, we would like to minimize the redundancy among the selected samples, as given by the entries in the matrix R. We define a binary matrix $M \in \{0,1\}^{|U|\times k}$ where each row corresponds to an unlabeled sample and each column corresponds to an oracle. A value of 1 in a row denotes that the sample should be selected for annotation, and the position of 1 in a particular row of M denotes the oracle that should be used to label the sample. The active (sample-oracle) selection is thus posed as the following optimization problem:

$$\begin{aligned} \min_{M} \quad & \text{trace}(MP) + \lambda (Me)^\top R(Me) \\ \text{s.t.} \quad & M_{ij} \in \{0,1\}, \forall i,j \\ & M_i.e \leq 1, \forall i \\ & \langle M, E \rangle = B \end{aligned} \tag{5}$$

where λ is a weight factor governing the relative importance of the two terms, e is a vector of length k with all entries 1, M_i denotes row i of matrix M, $\langle \cdot, \cdot \rangle$ denotes the matrix inner product operator, E is a matrix of the same dimension as M ($|U| \times k$) with the cost value C_j of oracle O_j in the entire column j, and B is the labeling budget. The first term in the objective function denotes that the selected samples have high entropy, low cost and low labeling error probability from the corresponding labeling oracles; the second term ensures that the selected samples have minimal redundancy among them. The first constraint denotes that M is a binary matrix; the second constraint signifies the each row of M can have at most one entry as 1, since each selected unlabeled sample can be labeled by exactly one oracle; and the third constraint denotes that the total cost incurred by labeling the selected samples equals the specified budget B. Such a formulation enables us to utilize the presence of multiple imperfect labeling oracles simultaneously (corroborating its usefulness in real-world applications), contrary to the methods proposed in (Huang et al. 2017) (Yan et al. 2011), which query only a single unlabeled sample and utilize a single labeling oracle in each AL iteration. We now discuss an efficient strategy to solve this optimization problem, as detailed in the following theorem.

Theorem 1. *The optimization problem defined in Equation (5) can be expressed as an equivalent linear programming (LP) problem.*

Proof. The first term in the objective function can be expressed as a linear term: $trace(MP) = \sum_{i,j} P_{ij}.M_{ji}$. The second term can be simplified as follows:

$$\begin{aligned} (Me)^\top R(Me) &= \sum_{i,j} R_{ij} (Me)_i (Me)_j = \sum_{i,j} R_{ij} \langle M_i.e, M_j.e \rangle \\ &= \sum_{i,j} R_{ij} \langle M_i, M_j.ee^\top \rangle = \sum_{i,j} R_{ij} \langle M_j^\top M_i, ee^\top \rangle \\ &= \sum_{i,j} R_{ij} \sum_{a,b} M_{ia}.M_{jb} = \sum_{i,j} \sum_{a,b} R_{ij} M_{ia}.M_{jb} \\ &= \sum_{i,j} \sum_{a,b} R_{ij} V_{ijab} \end{aligned}$$

where $V_{ijab} = M_{ia}.M_{jb}$ (we use the algebra of inner product operations and the fact that $ee^\top$ is a matrix of all 1's in this derivation). Since M is a binary matrix with only 0 and 1 entries, V_{ijab} will equal 1 when both M_{ia} and M_{jb} are 1 and will equal 0 otherwise. Due to the binary constraints on M, the quadratic equality $V_{ijab} = M_{ia}.M_{jb}$ can be expressed as the following equivalent linear inequality:

$$M_{ia} + M_{jb} \leq 1 + 2V_{ijab} \tag{6}$$

A simple observation reveals that when M_{ia} and M_{jb} are both 1, V_{ijab} has to be equal to 1. When M_{ia} and M_{jb} are both 0 or, one of them is 0 and the other one is 1, V_{ijab} is free to be both 0 and 1. However, we are solving a minimization problem with a term $\sum_{i,j} \sum_{a,b} R_{ij} V_{ijab}$ in the objective function (R has only non-negative entries). These conditions will force V_{ijab} to be 0, as it will lead to a better (lower) value of the objective function. Hence, the quadratic equality $V_{ijab} = M_{ia}.M_{jb}$ and the linear inequality $M_{ia} + M_{jb} \leq 1 + 2V_{ijab}$ produce the exact same values of V_{ijab} under all conditions. The optimization problem in Equation (5) can thus be expressed as follows:

$$\begin{aligned} \min_{M,V} \quad & \sum_{i,j} P_{ij}.M_{ji} + \lambda \sum_{i,j} \sum_{a,b} R_{ij} V_{ijab} \\ \text{s.t.} \quad & M_{ij}, V_{ijab} \in \{0,1\}, \forall i,j,a,b \\ & M_i.e \leq 1, \forall i \\ & \langle M, E \rangle = B \\ & M_{ia} + M_{jb} \leq 1 + 2V_{ijab} \end{aligned} \tag{7}$$

In this optimization problem, both the objective function and the constraints are linear in the variables M and V. It is thus a linear programming (LP) problem.

□

We vectorize the variables M and V, append them one below the other and express the objective function and the constraints in terms of this new variable. The integer constraints on M and V are then relaxed into continuous constraints and the problem is solved using an off-the-shelf LP solver. After obtaining the continuous solution, we recover the integer solution of our variable of interest M, using a greedy approach where the highest entries in each row of M are reconstructed as 1 and the other entries as 0, observing the constraints.

Experiments and Results

We conducted an extensive set of experiments to study the active learning performance of our framework against competing baselines, the labeling accuracy and the effects of the number of labeling oracles and query budgets. These are detailed below.

Datasets and Experimental Setup

We used 9 challenging datasets (binary and multi-class) from a variety of application domains (handwritten digits, objects, face and emotion recognition, medical diagnosis and spam filtering) to study the performance of our framework: MNIST (LeCun et al. 1998), SVHN (Netzer et al. 2011), CIFAR-10 (Krizhevsky 2009), VidTIMIT (Sanderson 2008), MindReading (El-Kaliouby and Robinson 2004), MMI (Pantic et al. 2005), Spambase, Sensor and Breast Cancer (the last three from the UCI Machine Learning Repository). Our objective was to test the performance of AL algorithms and not to outperform the best error rates in these datasets; we therefore did not follow the exact train / test splits mentioned for some of these datasets.

We simulated 5 labeling oracles (O_1–O_5) using 5 common classification models: k-nearest neighbors, naïve Bayes, SVM, Random Forest and Adaboost respectively. When an unlabeled sample needed to be labeled using a particular oracle, it was passed as a test sample to the corresponding classifier, and the predicted label was interpreted as the label provided by the oracle in a real-world setup.

Each dataset was divided into 5 parts: (i) an oracle training set (25%); (ii) an oracle test set (5%); (iii) an initial training set (10%); (iv) an unlabeled set (40%); and (v) a test set (20%). The first two parts were used to train and test the labeling oracles and the other three were used for evaluating the active learning algorithms. Table 1 shows the percentage accuracy of each of the oracles (on the oracle test set) for each of the datasets. As evident from the table, the accuracy of the oracles vary significantly for a given dataset, which aptly captures a practical scenario. Similar to previous research (Huang et al. 2017), each oracle was assigned an integer between 1 and 5 in increasing order of labeling accuracy, which was interpreted as the labeling cost C_j of oracle O_j for a given dataset, that is, the price to be paid to get one unlabeled sample from the dataset labeled by the particular oracle. This appropriately simulates a real-world setting, where more reliable and experienced annotators are more expensive.

Dataset	O_1	O_2	O_3	O_4	O_5
MNIST	81.33	87.33	82.33	85.00	81.33
SVHN	66.66	71.66	63.33	66.33	59.00
CIFAR	40.60	48.00	36.80	41.60	31.00
VidTIMIT	83.00	88.00	69.00	86.66	50.66
MindReading	55.00	67.00	63.00	65.00	59.50
MMI	65.00	88.50	81.00	82.00	74.00
Spambase	71.00	91.00	70.40	95.00	74.00
Sensor	97	83.6	93.4	98.8	82.6
Breast Cancer	62.2	93.2	62.2	94.4	62.2

Table 1: Accuracy (in percentage) of the Oracles on the Datasets used

In each AL iteration, a query budget B was imposed; each algorithm queried a batch of unlabeled samples and their labels were obtained from the selected oracles, such that the total cost of purchasing the labels from the oracles equalled the budget B. The selected samples, together with the labels predicted by the respective oracles, were then appended to the training set; the model was updated and tested on the test set. The process was continued iteratively until a stopping condition was satisfied (taken as 25 iterations in this work). The objective was to study the improvement in performance on the test set with increasing sizes of the training set. The query budget B in each AL iteration was set as 100. All the results were averaged over 5 runs (with different initial training, unlabeled and test sets) to rule out the effects of randomness. The parameter λ was taken as 0.3 (based on preliminary experiments). Logistic Regression (LR) was used as the classification model in our experiments.

Comparison Baselines

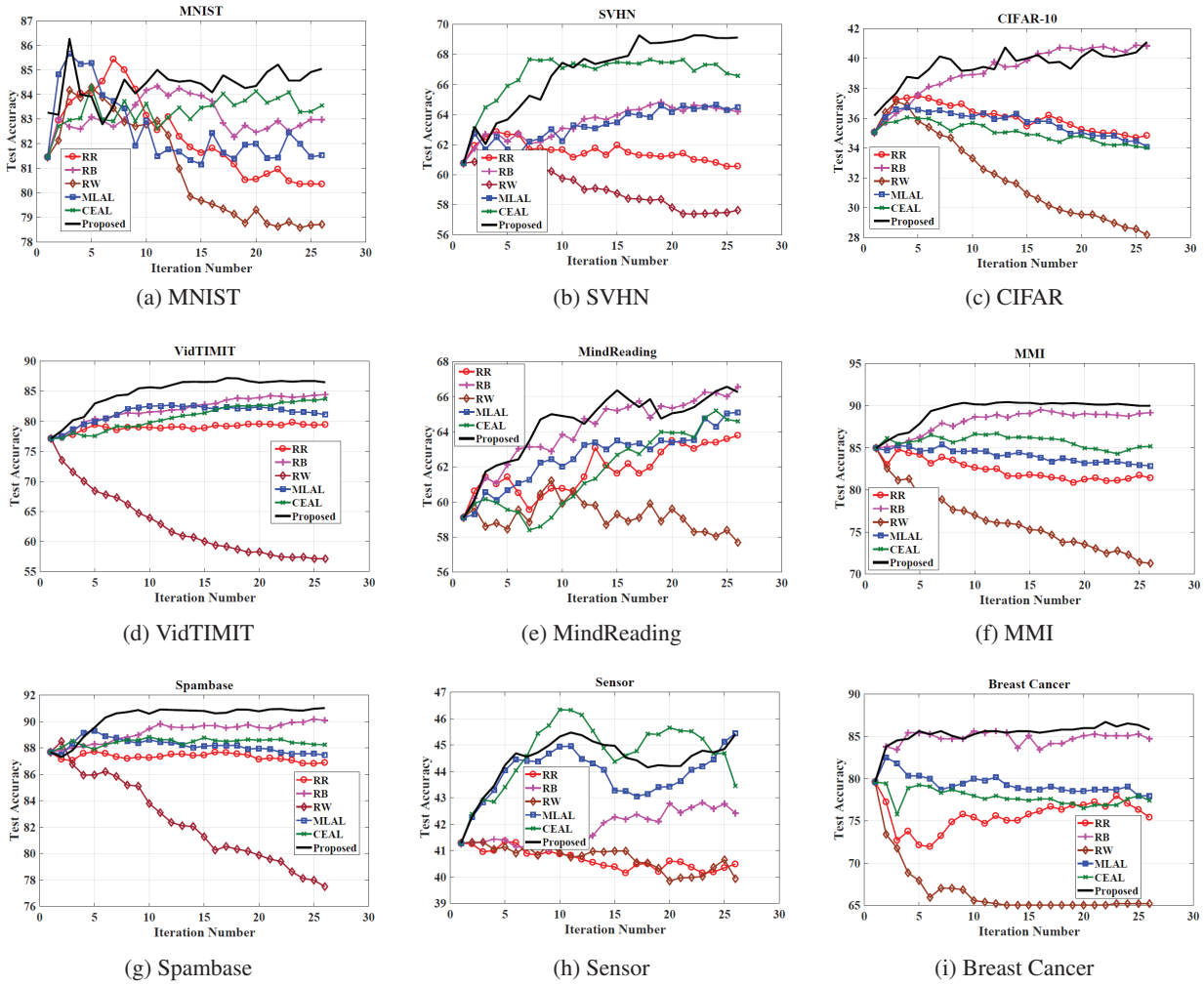
We compared the proposed algorithm against the following baselines: Cost Effective Active Learning (**CEAL**) proposed by Huang *et al.* (Huang et al. 2017); Multi-Labeler Active Learning (**MLAL**) proposed by Yan *et al.* (Yan et al. 2011). These two are the best-performing algorithms for this problem (Huang et al. 2017) and were hence selected as comparison baselines. In addition, we also compared the performance against **R**andom sampling (where a batch of unlabeled samples was selected at random) together with three oracle selection strategies: **R**andom, **B**est and **W**orst (defined by their accuracies in Table 1). This gave us 3 more baselines: **RR**, **RB** and **RW**. For instance, **RB** will select an unlabeled sample at random and always use the best oracle to get its label.

Active Learning Performance

The AL performance results are depicted in Figure 1. In each graph, the x-axis denotes the iteration number and the y-axis denotes the accuracy on the test set. The *CEAL* method depicts competitive performance in the SVHN and Sensor datasets; but it is not consistent across datasets in its performance (such as MMI, Spambase and Breast Cancer). The same holds for the *MLAL* algorithm which also depicts inconsistent performance across the datasets. As mentioned earlier, both these algorithms query only a single unlabeled sample in each AL iteration; thus, myopically extending the single query framework to multiple queries produces sub-optimal results. Random sampling using the best oracle (*RB*) can sometimes produce impressive performance, as in CIFAR-10 and Breast Cancer; however, its performance in the other datasets is much worse. Thus, always querying the best oracle does not necessarily produce optimal results. The *RW* method shows a consistent degradation in accuracy due to the addition of noisy data to the training set. The *RR* method depicts performance in between *RB* and *RW*. Our framework consistently depicts impressive performance across all the datasets. It shows a consistent upward trend in the accuracy, even in the presence of imperfect labeling oracles. At any given iteration number in any dataset, the proposed algorithm furnishes the highest, or very close to the highest accuracy value, compared to all the baselines. This shows that our framework is efficiently identifying the most informative unlabeled samples and the optimal annotators to get them labeled, in order to induce a robust model with minimal human effort. The results unanimously corroborate the potential of our method for real-world active learning applications in the presence of multiple noisy labeling oracles.

Study of Labeling Accuracy

In this experiment, we studied the correctness of the labels returned by the labeling oracles for all the methods. Table 2 reports the percentage accuracy in labeling the queried unlabeled instances by all the methods for all the datasets. The proposed method achieves the highest labeling accuracy in 7 out of the 9 datasets. This shows that, besides identifying the exemplar unlabeled samples to be queried, our algorithm also identifies the optimal oracles to annotate each sample.



(a) MNIST (b) SVHN (c) CIFAR

(d) VidTIMIT (e) MindReading (f) MMI

(g) Spambase (h) Sensor (i) Breast Cancer

Figure 1: Active Learning performance comparison. **RR** selects samples at random and queries oracles at random; **RB** selects samples at random and always queries the best oracle; **RW** selects samples at random and always queries the worst oracle. **MLAL** denotes the multiple labeler active learning algorithm (Yan et al. 2011); **CEAL** denotes the cost effective active learning algorithm (Huang et al. 2017). Best viewed in color.

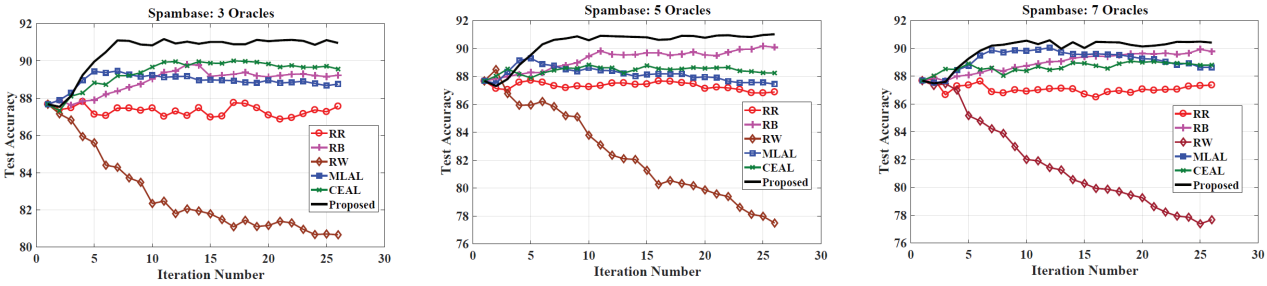
Dataset	RW	RR	RB	MLAL	CEAL	Proposed
MNIST	62.12	70.83	67.72	83.33	83.40	**84.09**
SVHN	49.00	55.69	53.70	63.63	**75.92**	71.66
CIFAR	28.75	40.40	38.50	29.83	42.20	**49.00**
VidTIMIT	46.18	45.61	65.45	67.43	76.53	**82.90**
MindReading	46.03	53.77	61.33	62.22	61.63	**70.13**
MMI	60.71	60.86	66.81	72.00	80.63	**85.18**
Spambase	65.84	69.64	84.28	79.64	77.14	**91.42**
Sensor	81.66	91.62	95.50	88.88	**98.50**	97.91
Breast Cancer	64.72	75.77	94.85	83.55	70.65	**96.57**

Table 2: Accuracy (in percentage) in labeling the queried unlabeled instances. Best results are shown in **bold**.

Thus, using a binary LR, our framework can accurately identify error patterns in labeling among the oracles and can thus obtain high quality labels, which accounts for its superior performance in Figure 1.

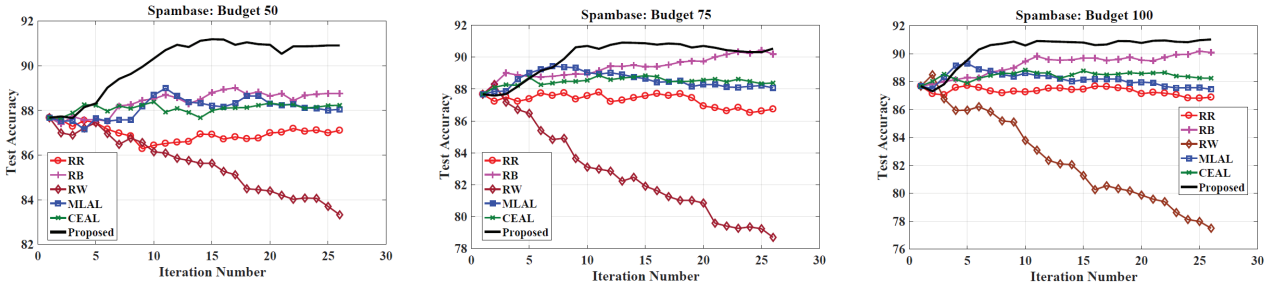
Effect of Number of Labeling Oracles

The goal of this experiment was to study the effect of the number of labeling oracles on the active learning performance. Besides the default setting with 5 oracles, we studied the performance with 3 oracles (SVM, Random Forest and Adaboost) and 7 oracles (k-NN, naïve Bayes, SVM, Random Forest, Adaboost, discriminant analysis and decision trees). The results on the Spambase dataset are shown in Figure 2. The proposed method once again depicts the fastest accuracy growth with increasing number of iterations compared to all the baselines, for all three experiments. This shows the robustness of our framework to varying number of oracles and its capability to find the optimal labeling oracle for a given unlabeled sample, regardless of the number of oracles.

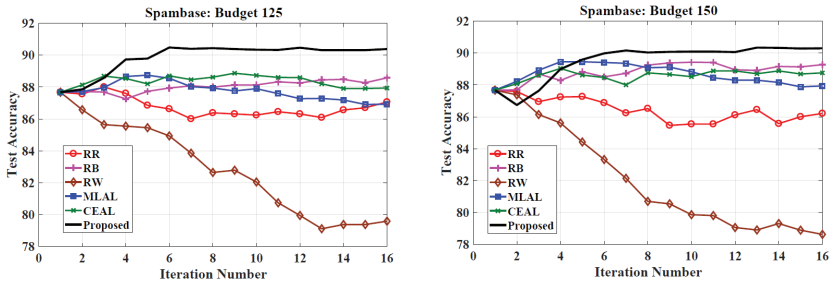


(a) 3 Oracles (b) 5 Oracles (c) 7 Oracles

Figure 2: Effect of the number of labeling oracles on the Spambase dataset. Best viewed in color.



(a) Budget 50 (b) Budget 75 (c) Budget 100



(d) Budget 125 (e) Budget 150

Figure 3: Effect of varying query budgets on the Spambase dataset. Best viewed in color.

Effect of Query Budgets

In this experiment, we studied the effect of varying query budgets B. The results on the Spambase dataset for query budgets 50, 75, 100, 125 and 150 are depicted in Figure 3. The number of labeling oracles was fixed at 5. The results follow a similar pattern as Figure 2, with the proposed method outperforming the baselines. Our algorithm thus depicts impressive performance across varying query budgets. This corroborates its potential for real-world applications, where the budget is governed by the time and available resources of an application.

Conclusion and Future Work

In this paper, we studied the problem of batch mode active learning in the presence of multiple imperfect labeling oracles. This captures a practical real-world setting and is contrary to most existing active learning algorithms, where the oracles are assumed to be infallible or, which utilize only a single oracle at any given point of time. We proposed a novel framework to simultaneously select a batch of informative unlabeled samples, together with the best annotator to label each sample. The active sample-oracle selection was posed as a constrained optimization problem, based on the confidence, redundancy, labeling cost and oracle error probability criteria, and was shown to be equivalent to a linear programming problem. Our extensive empirical studies on 9 challenging datasets from different application domains demonstrated the merit of our algorithm over competing baselines, in terms of active learning performance, labeling accuracy, number of labeling oracles and query budgets. As part of future research, we plan to study the performance of our framework on other problems, such as multi-label learning.

References

Ambati, V.; Vogel, S.; and Carbonell, J. 2010. Active learning and crowd-sourcing for machine translation. In *Interna-*

tional Conference on Language Resources and Evaluation.

Chakraborty, S.; Balasubramanian, V.; and Panchanathan, S. 2015. Adaptive batch mode active learning. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)* 26(8):1747 – 1760.

Chakraborty, S.; Balasubramanian, V.; Sun, Q.; Panchanathan, S.; and Ye, J. 2015. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 37(10):1945–1958.

Donmez, P., and Carbonell, J. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *ACM Conference on Information and Knowledge Management (CIKM)*.

Donmez, P.; Carbonell, J.; and Schneider, J. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

El-Kaliouby, R., and Robinson, P. 2004. Mind reading machines: Automated inference of cognitive mental states from video. In *IEEE International Conference on System, man and Cybernetics*.

Fang, M.; Yin, J.; and Tao, D. 2014. Active learning for crowdsourcing using knowledge transfer. In *AAAI Conference on Artificial Intelligence*.

Freund, Y.; Seung, S.; Shamir, E.; and Tishby, N. 1997. Selective sampling using the query by committee algorithm. *Machine Learning* 28(2-3):133–168.

Guo, Y., and Schuurmans, D. 2007. Discriminative batch mode active learning. In *Neural Information Processing Systems (NIPS)*.

Guo, Y. 2010. Active instance sampling via matrix partition. In *Neural Information Processing Systems (NIPS)*.

Hoi, S.; Jin, R.; Zhu, J.; and Lyu, M. 2008. Semi-supervised SVM batch mode active learning for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Holub, A.; Perona, P.; and Burl, M. 2008. Entropy-based active learning for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*.

Huang, S.; Chen, J.; Mu, X.; and Zhou, Z. 2017. Cost-effective active learning from diverse labelers. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Ipeirotis, P.; Provost, F.; Sheng, V.; and Wang, J. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. In *Technical Report*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11).

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. 2011. Reading digits in natural images with unsupervised feature learning. In *Neural Information Processing Systems (NIPS) Workshop*.

Osmanbeyoglu, H.; Wehner, J.; Carbonell, J.; and Ganapathiraju, M. 2010. Active machine learning for transmembrane helix prediction. *BMC Bioinformatics* 11(1).

Pantic, M.; Valstar, M.; Rademaker, R.; and Maat, L. 2005. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*.

Ranganathan, H.; Venkateswara, H.; Chakraborty, S.; and Panchanathan, S. 2017. Deep active learning for image classification. In *IEEE International Conference on Image Processing (ICIP)*.

Sanderson, C. 2008. *Biometric Person Recognition: Face, Speech and Fusion*. VDM Verlag.

Sculley, D. 2007. Online active learning methods for fast Label-Efficient spam filtering. In *Fourth Conference on Email and AntiSpam*.

Settles, B. 2010. Active learning literature survey. In *Technical Report 1648, University of Wisconsin-Madison*.

Shen, D.; Zhang, J.; Su, J.; Zhou, G.; and Tan, C. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Tong, S., and Koller, D. 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research (JMLR)* 2:45–66.

Urner, R.; Ben-David, S.; and Shamir, O. 2012. Learning from weak teachers. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Yan, Y.; Fung, G.; Rosales, R.; and Dy, J. 2011. Active learning from crowds. In *International Conference on Machine Learning (ICML)*.

Yan, Y.; Rosales, R.; Fung, G.; Farooq, F.; Rao, B.; and Dy, J. 2012. Active learning from multiple knowledge sources. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Yan, S.; Chaudhuri, K.; and Javidi, T. 2016. Active learning from imperfect labelers. In *Neural Information Processing Systems (NIPS)*.

Zhang, C., and Chaudhuri, K. 2015. Active learning from weak and strong labelers. In *Neural Information Processing Systems (NIPS)*.

Zhao, L.; Sukthankar, G.; and Sukthankar, R. 2011. Incremental relabeling for active learning with noisy crowdsourced annotations. In *International Conference on Social Computing*.

Zheng, Y.; Scott, S.; and Deng, K. 2010. Active learning from multiple noisy labelers with varied costs. In *IEEE International Conference on Data Mining (ICDM)*.

Zhu, J., and Bento, J. 2017. Generative adversarial active learning. In *arXiv:1702.07956*.