Pattern Recognition Letters 43 (2014) 98–108

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Active learning with uncertain labeling knowledge ☆

Meng Fang [a,*], Xingquan Zhu [b]

[a] QCIS – Centre for Quantum Computation & Intelligent Systems, University of Technology Sydney, P.O. Box 123, Broadway, NSW 2007, Australia
[b] Dept. of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA

ARTICLE INFO

Article history:
Available online 30 October 2013

Communicated by Ken-ichi Maeda

Keywords:
Active learning
Classification
Uncertain knowledge

ABSTRACT

Traditional active learning assumes that the labeler is capable of providing ground truth label for each queried instance. In reality, a labeler might not have sufficient knowledge to label a queried instance but can only guess the label with his/her best knowledge. As a result, the label provided by the labeler, who is regarded to have uncertain labeling knowledge, might be incorrect. In this paper, we formulate this problem as a new "uncertain labeling knowledge" based active learning paradigm, and our key is to characterize the knowledge set of each labeler for active learning. By taking each unlabeled instance's information and its likelihood of belonging to the uncertain knowledge set as a whole, we define an objective function to ensure that each queried instance is the most informative one for labeling and the labeler should also have sufficient knowledge to label the instance. To ensure label quality, we propose to use diversity density to characterize a labeler's uncertain knowledge and further employ an error-reduction-based mechanism to either accept or decline a labeler's label on uncertain instances. Experiments demonstrate the effectiveness of the proposed algorithm for real-world active learning tasks with uncertain labeling knowledge.

1. Introduction

Obtaining labeling information for training instances, in a supervised learning task, is a nontrivial process which normally involves expensive costs. Instead of labeling the entire training set or randomly selecting a subset for labeling, active learning (Cohn et al., 1994) represents a family of methods which identify some most informative instances for the oracle[1] to label. A large number of studies have shown that active learning can significantly reduce the labeling costs in various fields (Settles, 2010). However, most existing active learning methods rely on a strong assumption that the oracle has perfect labeling knowledge and can provide correct labels for each queried instances.

In reality, it is possible that an oracle may have insufficient knowledge in labeling some instances (Raykar et al., 2009; Rzhetsky et al., 2009). This normally happens in two situations: (1) the instances to be labeled are not clearly correlated to the underlying labeling concepts, so labelers cannot provide accurate labels based on the limited information in the instances; and (2) the labelers have insufficient knowledge and therefore cannot identify the differences between some closely correlated instances, although the instance itself already contains enough labeling information.

In Fig. 1, we illustrate three examples to demonstrate the labeling uncertainty in several image labeling domains, where each row represents a labeling task and the labeler is required to accurately label each single image as one of the given concepts. In Fig. 1(A), the labeling concept is "mountain" vs. "not mountain". While the images in the left and the middle panels are easy to label without confusion, the image on the right panel is difficult to label mainly because the picture itself has limited information for labelers to provide accurate label. In this situation, the labeler may have good labeling knowledge, but the limitation of the instance results in the label uncertainty. In Fig. 1(B), the labeling concept is "camel" vs. "not camel". The image on the left panel can be clearly identified as a "camel" by a nonexpert, whereas the image on the middle panel might not be easy to identify without good domain knowledge (The animal is a Rama and is therefore not a camel). The image on the right panel is even more difficult because it only shows a portion of the animal, yet very skilled labeler is still able to identify it as Appace (and therefore is not a camel). In Fig. 1(C), the labeling task, "California gull" vs. "not California gull", is even more challenging. Although domain experts may have general knowledge, such as "medium-sized gull with head and underparts white but black ring near the tip and red spot on lower mandible, and yellow-green leges", it is still very difficult to accurately label the images in the middle and the right panels, which are Western Gull (middle) and Herring Gull (right). There is little differences between the heads and tails and they are also hard to observe.

☆ This paper was presented, in part, at the Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba Science City, Japan, November 2012.

* Corresponding author. Tel.: +61 416669269.
E-mail addresses: Meng.Fang@student.uts.edu.au (M. Fang), xzhu3@fau.edu (X. Zhu).

[1] In this paper, labelers, oracles, and domain experts are equivalent terms.

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.011

Fig. 1. Examples of image labeling with uncertainty.

The similar labeling uncertainty also common exits in other domains, such as text labeling. In Table 1, we show three real-world scientific text annotation examples (Rzhetsky et al., 2009), where the experts are required to mark (*i.e.*, label) to polarity of each sentence as either "Positive" or "Negative", and the labelers are allowed to express their confidence with a certain degree of Certainty score (0–3). If the labeler has knowledge for labeling an instance, he/she may provide a label (Positive or Negative) with a high certainty score. If the labeler does not have the labeling knowledge, he/she may guess the most possible label, but indicating a low certainty score (Certainty = 0).

For the above examples, regardless of whether the labeler has limited knowledge or the instance contains limited information, the consequence is that the labeler cannot provide accurate labels for some queried instances. Under such circumstances, for a queried instance on which the labeler has insufficient labeling knowledge, the labeler may simply answer "I don't know the label" and then guesses the most likely label based on the existing knowledge. As a result, the label information provided by the labeler is essentially uncertain and needs to be carefully verified during the active learning process.

Motivated by the above observations, in this paper, we formulate the problem as an active learning paradigm with uncertain knowledge. In this new setting, the oracle is no longer perfect but has uncertain knowledge, such that the instances within the uncertain knowledge set may be incorrectly labeled. For a queried instance x, the oracle may correctly label x (if the oracle has sufficient labeling knowledge) or only guess a label for x (if x falls into the oracle's uncertain knowledge set). Accordingly, an effective active learning framework should clearly address the following two major issues:

- **Uncertain knowledge characterization**: The active learning process should be able to identify the oracle's uncertain knowledge and carefully avoid querying instances which may fall into the oracle's uncertain knowledge set.

Table 1
Examples of text annotation with uncertain labeling knowledge. The three sentences are extracted verbatim from the scientific texts annotated by experts with personal knowledge (Rzhetsky et al., 2009). Each labeler marks the polarity of a sentence as Positive (P) or Negative (N) with a Certainty score (0, 1, 2, 3), with a score "0" indicating completely unsure. For the first and the third sentences, the labeler marks their polarity as "Positive" but with different degrees of uncertainty.

Sentence	Annotation	
	Positive	Certainty
Putative transmembrane domains are highlighted with gray	1	3
No interconversion of the two forms was detected after purification	0	2
The function of this gene, necessary for surfactin production, is still unclear	1	0

- **Uncertain label utilization**: Because the labels of uncertain instances are predicted based on the oracle's existing knowledge, the active learning process should carefully determine whether to accept the label provided by the oracle to update and retrain the model for future learning process.

The inherent technical challenges associated to the problem is threefold:

- **Instance selection for labeling:** How to select the instance mostly needed for active learning by considering both the informativeness of the instance and the labeler's uncertain knowledge;
- **Uncertain knowledge characterization:** How to characterize the labeler's uncertain knowledge and avoid select instances on which the labeler has insufficient knowledge;
- **Label confirmation or rejection:** How to make the best use of the labels provided by the labeler, if the he/she is uncertain about the labels of the instances and provides some guessed labels based on the existing knowledge.

To address the above challenges for active learning with uncertain knowledge, we propose to take each unlabeled instance's information and its likelihood of belonging to the uncertain knowledge set as a whole, and define an objective function to ensure that (1) each queried instance submitted to the oracle is the one most informative for labeling and (2) the oracle should also have sufficient knowledge to label the instance.

Notice that without properly characterizing the labeler's knowledge, it is hardly possible to determine whether the label has sufficient knowledge to label an instance or not. Accordingly, we use Diverse Density to characterize the labeler's uncertain knowledge, through which we can identify each unlabeled instance's likelihood of falling into the labeler's uncertain knowledge set. To handle an uncertain instance, on which the labeler does not know the ground truth label but can only guess, we propose an error-reduction estimation based controlling mechanism to either accept or reject the label guessed by the labeler, through which the active learning can maximize the utilization of the labeler's knowledge as well as control errors introduced from the labeler's uncertain knowledge.

The remainder of the paper is organized as follows. Section 2 discusses the related work. In Section 3, we introduce the proposed active learning paradigm with uncertain knowledge, including methods for characterizing labeler's uncertain knowledge and utilization of the labels provide by the labelers for active learning. Experimental results are reported in Section 4, and we conclude the paper in Section 5.

2. Related work

Classical active learning strategy is to query instances which are most uncertain to the learner (s) (or classifiers) trained from labeled instances (Lewis and Gale, 1994). Alternatively, one can select the instance on which a committee of classifiers mostly disagree (Freund et al., 1997). Another general uncertainty sampling strategy is based on information-theoretic measure which queries the instance minimizing the posterior entropy (MacKay, 1992; Tong and Koller, 2002). For learners considering data distributions in their decision models, such as Support Vector Machines, one can choose to label instances which are close to the learner's decision boundaries (Tong and Koller, 2002). All these methods share the same view as the version space reduction (Schohn and Cohn, 2000; Tong and Koller, 2002). Another set of methods directly minimize the empirical risk by querying instances to reduce future classification error as much as possible (Roy and McCallum, 2001). For all these active learning methods, an important assumption is that the oracle always knows the true labels of the queried instances.

Recently, several works argue the assumption that the oracle can always behave perfectly is too strong for real-world applications (Rashidi and Cook, 2011). Some studies focus on the problem of multiple weak labelers who might provide noisy labels (Donmez and Carbonell, 2008; Yan et al., 2010, 2012). In order to handle noisy labelers, solutions exist to learn the qualities of multiple labelers in tandem with learning values of classifier parameters (Dekel and Shamir, 2009; Yan et al., 2011), to repeatedly acquire labels over multiple rounds in order to reach a consensus sources/labelers (Sheng et al., 2008), or to select the optimal labeler from multiple weak labelers by solving an optimization problem with a fixed budget constraint (Donmez and Carbonell, 2008).

Although the above existing works have taken into account the scenarios where oracles in active learning are imperfect, their problem settings and solutions are subject to a major limitation: all these methods actually assume that oracles are subject to different levels of expertise and inherently disregard whether an oracle can label an instance or not. In other words, they realize that oracles might be weak and noisy, so their active learning solutions mainly focus on how to combine oracles' noisy labels in order to gain better label quality. There is, however, no mechanism (or solution) to characterize the oracle's knowledge. As a result, there is no treatment to avoid an oracle's weakness, and their methods would still require all oracles to label the selected instances, even though the instances might be out of the labelers' domain knowledge.

Another limitation of the existing noisy-labeler-based active learning methods is that they all require multiple oracles and are not intended to be used in a single oracle scenario. In their problem settings, multiple weak/cheap labelers can provide redundant information which help choose a high quality labeler or label which mostly agreed by the labelers. In our problem setting, there is only one oracle who has limitations in labeling some samples. Meanwhile, the oracle in our problem setting is not assumed to be a "cheap" labeler, but involves certain labeling costs. So for each uncertain instance, active learning should avoid repeatedly querying multiple times to refine the label information. As a result, existing noisy labelers based active learning methods still cannot handle our problem.

A recent work (Tuia and Muñoz-Marí, 2013) considers the labeler with uncertain knowledge in the remote sensing domains, with assumption that the labeler will not provide labels for instances which the labeler does not have the labeling knowledge. This work is close to our problem setting, but our problem is more general in the sense that we allow labelers to provide guessed labels, even though the labeler do not have the labeling knowledge for the instances. In other words, we allow labeler to provide uncertain labeling information, and our method will directly model the labeler's knowledge based on his/her labeling outputs.

3. Active learning with uncertain labeling knowledge

In this section, we first formulate the problem definition and define the objective function, and then propose a method to characterize a labeler's uncertain knowledge and utilize the uncertain labels provided by the label. After that, we propose the uncertain labeling knowledge based active learning framework.

3.1. Problem formulation

Consider a data set with n instances $\{x_1, x_2, \ldots, x_n\}$, where the label for the ith instance is denoted by y_i. In a generic active learning setting, the oracle is able to provide ground truth label for every queried instance, so the objective of the uncertainty sampling based active learning (Freund et al., 1997) is to query the instance with the highest uncertainty value (*e.g.* entropy). Accordingly, given the labeled data, we have

$$\arg\max_{x_i \in \mathcal{U}} H(y_i; \hbar(\mathcal{L})) \tag{1}$$

where $\mathcal{U}$ denotes the set of unlabeled instances and H represents the entropy of instance x_i with respect to the class labels predicted by a classifier $\hbar(.)$ trained from labeled set $\mathcal{L}$. For a data set with two class labels, the most informative instance selected by Eq. (1) is the one with equal likelihood of belonging to both classes.

In the above setting, the oracle has unbounded knowledge to label any instances. In real-world scenarios, such as scientific text annotation (Rzhetsky et al., 2009), the oracle may have limited domain knowledge or uncertain knowledge so cannot provide correct labels for some instances. The set of instances, which the oracle does not know the ground truth labels, form the oracle's uncertain knowledge set. Formally, we define that

Definition. *Uncertain knowledge* represents a set of unlabeled instances which the oracle does not know the ground truth labels; an *Uncertain Instance* is an instance which belongs to the uncertain knowledge set.

Because the oracle does not know the genuine labels of uncertain instances, when sending an uncertain instance to the oracle to query the label, the oracle will answer "I don't know the label" and guess the most likely label based on the labeler's existing knowledge. On the other hand, when sending an instance which does not belong to the uncertain knowledge set of the oracle, the oracle will return the genuine label of the instance. The set of queried instances thus form the knowledge base of the oracle. Formally, we define that

Definition. The *knowledge base* ($\mathcal{B}$) is defined as the union of a set of instances ($\mathcal{B}^+$) which have been labeled by the oracle and a set of instances ($\mathcal{B}^-$) which the oracle has confirmed that it does not have knowledge to label.

Denote by $\mathcal{O}$ the oracle of the underlying active learning task, $\mathcal{B}^+$ can be regarded as the set of knowledge that the oracle has already acquired from the active learning process and $\mathcal{B}^-$ represents the set of uncertain knowledge of the oracle. Therefore, after querying, the new unlabeled data set for the oracle becomes $\mathcal{U} = \mathcal{U} \setminus \mathcal{B}$. The expected entropy of an unlabeled instance x_i with respect to sets $\mathcal{B}^+$ and $\mathcal{B}^-$ is given by

$$H(y_i; \hbar(\mathcal{L})) = P(x_i \in \mathcal{B}^+) H(y_i | x_i \in \mathcal{B}^+; \hbar(\mathcal{L})) + P(x_i \in \mathcal{B}^-) H(y_i | x_i \in \mathcal{B}^-; \hbar(\mathcal{L})) \quad (2)$$

It is clear that knowledge base $\mathcal{B} = \mathcal{B}^+ \cup \mathcal{B}^-$, and

$$P(x_i \in \mathcal{B}^+) + P(x_i \in \mathcal{B}^-) = 1 \quad (3)$$

If the oracle does not know the genuine label of instance x_i (*i.e.*, x_i falls into the uncertain knowledge set), x_i is regarded as an out-of-domain instance for both the oracle and the underlying classifier $\hbar(\mathcal{L})$, which is trained based on the oracle's knowledge. In this case, the entropy which is conditioned on $x_i \in \mathcal{B}^-$ becomes

$$H(y_i | x_i \in \mathcal{B}^-; \hbar(\mathcal{L})) = 0, \quad \text{if } x_i \in \mathcal{B}^- \quad (4)$$

This is because, if $x_i \in \mathcal{B}^-$, the value of y_i is completely determined by x_i (*i.e.*, $y_i \equiv$ *unknown*) and, according to the definition of the conditional entropy, the conditional entropy is 0. Combining the oracle's knowledge set and the instance's information, the objective function in Eq. (1) can be rewritten as

$$\begin{aligned} &\arg\max_{x_i \in \mathcal{U}} P(x_i \in \mathcal{B}^+) H(y_i | x_i \in \mathcal{B}^+; \hbar(\mathcal{L})) \\ &= \arg\max_{x_i \in \mathcal{U}} (1 - P(x_i \in \mathcal{B}^-)) H(y_i | x_i \in \mathcal{B}^+; \hbar(\mathcal{L})) \end{aligned} \quad (5)$$

Eq. (5) represents the trade-off between minimizing the probability of falling into the oracle's uncertain knowledge set and maximizing the entropy of the instance. We expect that unifying an instance's information and its likelihood of belonging to the oracle's uncertain knowledge set will help select right instances for active learning.

3.2. Uncertain knowledge characterization

To estimate $P(x_i \in \mathcal{B}^+)$ in Eq. (5), we employ the diverse density concept (Maron and Lozano-Pérez, 1998) to build knowledge model for the oracle. The diverse density was first introduced in multi-instance learning, where a bag is labeled positive if one or more instances in the bag are positive and negative only if all instances in the bag are negative. The diverse density defines the density of the instances, in terms of how many positive bags are within a region and how far is the region to the negative bags, to help predict whether an instance is positive or not.

We assume that there exists a concept set $\mathcal{C}$ which represents the oracle's knowledge. Once $\mathcal{C}$ is properly captured, we will use diverse density to transform instances from their original feature space to a new feature space. Then, we build a classifier in the new feature space to estimate $P(x_i \in \mathcal{B}^+)$.

Given an active learning process, we denote the set of instances, which have been labeled by the oracle, by $\mathcal{B}^+ = \{\mathfrak{b}_1^+, \ldots, \mathfrak{b}_p^+\}$. Similarly, the set of instances, which the oracle has confirmed that it does not know the labels, form $\mathcal{B}^- = \{\mathfrak{b}_1^-, \ldots, \mathfrak{b}_q^-\}$. Then, we define the diverse density (DD) of $\mathcal{C}$ as the probability of $\mathcal{C}$ being the target concept given set ($\mathcal{B}^+$) and ($\mathcal{B}^-$) for the oracle

$$DD(\mathcal{C}) = P(\mathcal{C} | \mathfrak{b}_1^+, \ldots, \mathfrak{b}_p^+, \mathfrak{b}_1^-, \ldots, \mathfrak{b}_q^-) \quad (6)$$

The concept $\mathcal{C}$ with the maximum diverse density value will be selected as the target. Assume instances in $\mathcal{B}^+$ and $\mathcal{B}^-$ are conditionally independent, given the real target concept. By using Bayes' rules and rearranging Eq. (6), we have

$$\begin{aligned} DD(\mathcal{C}) &= \frac{P(\mathcal{C}) \prod_{i=1}^{p} P(\mathfrak{b}_i^+ | \mathcal{C}) \prod_{j=1}^{q} P(\mathfrak{b}_j^- | \mathcal{C})}{P(\mathfrak{b}_1^+, b_2^+, \ldots, \mathfrak{b}_p^+, \mathfrak{b}_1^-, \mathfrak{b}_2^-, \ldots, \mathfrak{b}_q^-)} \\ &= \left[\frac{\prod_{i=1}^{p} P(\mathfrak{b}_i^+) \prod_{j=1}^{q} P(\mathfrak{b}_j^-)}{P(\mathfrak{b}_1^+, \ldots, \mathfrak{b}_p^+, \mathfrak{b}_1^-, \ldots, \mathfrak{b}_q^-) P(\mathcal{C})^{p+q-1}} \right] \left[\prod_{i=1}^{p} P(\mathcal{C} | \mathfrak{b}_i^+) \prod_{j=1}^{q} P(\mathcal{C} | \mathfrak{b}_j^-) \right] \end{aligned} \quad (7)$$

Assuming target concept set $\mathcal{C}$ consists of a number of small concepts $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$, the conditional probability of each small concept c_k, given an instance $\mathfrak{b}_\tau$ in the knowledge base $\mathcal{B}$, can be defined as a feature value of $\mathfrak{b}_\tau$ (Chen et al., 2006). As a result, we can form a new set of features for $\mathfrak{b}_\tau$ as follows:

$$\mathbf{f}_{\mathcal{C}}(\mathfrak{b}_\tau) = \left[f_{c_1}(\mathfrak{b}_\tau), \ldots, f_{c_m}(\mathfrak{b}_\tau) \right]^T = \left[P(c_1 | \mathfrak{b}_\tau), \ldots, P(c_m | \mathfrak{b}_\tau) \right]^T \quad (8)$$

Indeed, instances in the knowledge base $\mathcal{B}$ are the best candidates for determining the target concept set $\mathcal{C}$. Accordingly, we use all instances in $\mathcal{B}$ to approximate $\mathcal{C}$ as follows:

$$\mathcal{C} = \{\mathfrak{b}_1^+, \ldots, \mathfrak{b}_p^+, \mathfrak{b}_1^-, \ldots, \mathfrak{b}_q^-\} \quad (9)$$

Therefore, the number of small concepts in $\mathcal{C}$ exactly equals to the sum of the number of instances in $\mathcal{B}^+$ and in $\mathcal{B}^-$, *i.e.*, $m = p + q$. Given target concept set as defined in Eq. (9), our next step is to estimate the conditional probability $P(c_k | \mathfrak{b}_\tau)$. Intuitively, because both c_k and $\mathfrak{b}_\tau$ are individual instances, the conditional probability $P(c_k | \mathfrak{b}_\tau)$ is proportional to the Gaussian distance between them, as Maron and Lozano-Pérez (1998) have suggested in their *most-likely-cause estimator* for conditional probability estimation. As a result, for a single concept c_k, its conditional probability with respect to instance $\mathfrak{b}_\tau$ is defined as follows:

$$P(c_k | \mathfrak{b}_\tau) \propto \bar{d}(c_k, \mathfrak{b}_\tau) = exp\left(-\frac{|c_k - \mathfrak{b}_\tau|^2}{\sigma^2} \right) \quad (10)$$

In Eq. (10), $\bar{d}(.)$ measures the Gaussian distance between an instance and a concept. If an instance is closer to a concept c_k, it will have a higher probability value of belonging to c_k. Given the concept set $\mathcal{C}$, the conditional probability of all small concepts in $\mathcal{C}$, $P(c_k | \mathfrak{b}_\tau)$, $k = 1, \ldots, m$, will provide useful information to differentiate an instance's likelihood of belonging to the uncertain knowledge spot $\mathcal{B}^-$ and the acquired knowledge set $\mathcal{B}^+$. For the given knowledge base $\mathcal{B}$ with $m = p + q$ instances, we can generate a mapped instance set in the new feature space $\mathbb{R}_c$ as follows:

$$\begin{aligned} \mathbf{f}_{\mathcal{C}}(\mathcal{B}) &= [\mathfrak{b}_1^+, \mathfrak{b}_2^+, \ldots, \mathfrak{b}_p^+, \mathfrak{b}_1^-, \mathfrak{b}_2^-, \ldots, \mathfrak{b}_q^-] \\ &= [\mathbf{f}_{\mathcal{C}}(\mathfrak{b}_1^+), \mathbf{f}_{\mathcal{C}}(\mathfrak{b}_2^+), \ldots, \mathbf{f}_{\mathcal{C}}(\mathfrak{b}_p^+), \mathbf{f}_{\mathcal{C}}(\mathfrak{b}_1^-), \mathbf{f}_{\mathcal{C}}(\mathfrak{b}_2^-), \ldots, \mathbf{f}_{\mathcal{C}}(\mathfrak{b}_q^-)] \end{aligned}$$

For each instance in the knowledge base $\mathcal{B}$, we can use a sign function to define a new class label as follows:

$$sign(\mathfrak{b}_\tau) = \begin{cases} 1 & \text{if } \mathfrak{b}_\tau \in \mathcal{B}^+ \\ -1 & \text{if } \mathfrak{b}_\tau \in \mathcal{B}^- \end{cases} \tag{11}$$

Then

$$(\mathcal{B}) = [sign(\mathfrak{b}_1^+), \ldots, sign(\mathfrak{b}_p^+), sign(\mathfrak{b}_1^-), \ldots, sign(\mathfrak{b}_q^-)]^T$$

provides new labeling information for all instances in $\mathcal{B}$. Combining mapped new feature values $\mathbf{f}_\mathcal{C}(\mathcal{B})$ and labels $(\mathcal{B})$, we can form a well defined binary classification task. By using any existing learning algorithms, we will be able to train a learner $\hbar(\mathbf{f}_\mathcal{C}(\mathcal{B}), (\mathcal{B}))$ and predict a new instance x_i's likelihood of belonging to the acquired knowledge $\mathcal{B}^+$, as defined in Eq. (12).

$$P(x_i \in \mathcal{B}^+) = \hbar(\mathbf{f}_\mathcal{C}(\mathcal{B}), (\mathcal{B}))[\mathbf{f}_\mathcal{C}(x_i); 1] \tag{12}$$

In Eq. (12), $\mathbf{f}_\mathcal{C}(x_i)$ denotes the transformed instance of x_i in the new feature space $\mathbb{R}_c$, and $\hbar(.)[\mathbf{f}_\mathcal{C}(x_i); 1]$ denotes the class distribution of the classifier $\hbar(.)$ in classifying $\mathbf{f}_\mathcal{C}(x_i)$ into class "1". One can use any learning algorithm to train $\hbar(.)$.

3.3. Uncertain label utilization

When querying the oracle for the label of an instance, we may face the situation that the oracle may not be able to provide ground truth label for the queried instance, but can only provide a prediction with a certain degree of uncertainty. Such reality raises a dilemma that (1) if we discard the labels provided by the labelers, we might not be able to take the full advantage of the labelers' knowledge for active learning; on the other hand, (2) if we unconditionally accept the labels provided by the labelers, the incorrectly predicted labels by the labelers may deteriorate the active learning performance, because including mislabeled samples in the training set is considered mostly harmful for supervised learning.

To ensure the label quality and tackle uncertain answers from the oracle, we propose to optimize expected future error to decide whether to accept the label provided by the labeler for learning.

To verify the prediction from the oracle, we propose the following strategy: For instance x^* which the oracle does not have knowledge to label (where x^* is selected using Eq. (5)), we decide whether it is intractable to accept the prediction of x^*.

After the oracle confirms that it has no knowledge to label x^*, it is asked to provide a prediction $\hat{y}^*$ on x^* based on its existing knowledge. Then we face the problem to either reject or accept the label predicted by the oracle. Ideally, if the predicted label is the same as the ground truth label, the most informative instance and its predicted label should be consistent with the learner's prior belief over the majority (but not all) of unlabeled instances (Roy and McCallum, 2001). On the other hand, if the oracle predicted label is different from the ground truth label, the most informative instance will act as a noisy sample and may result in additional errors for any classifier trained from the labeled set. Accordingly, we can try to verify if the predicted instance $(x^*, \hat{y}^*)$ can result in a lower expected error by

$$\Delta\mathcal{E} = \mathcal{E}_{\widehat{P}_{\mathcal{L}\cup(x^*,y^*)}} - \mathcal{E}_{\widehat{P}_\mathcal{L}} < 0 \tag{13}$$

where $\mathcal{E}$ is the expected error of the learner, which is defined as

$$\mathcal{E}_p = \int_x \ell(P(y|x), \widehat{P}(y|x))P(x) \tag{14}$$

In Eq. (14), ℓ is the loss function that measures the difference between the true distribution $P(y|x)$ and the learner's prediction $\widehat{P}(y|x)$. We employ two commonly used loss functions, and rewrite Eq. (14) as

$$\hat{\mathcal{E}}_{\widehat{P}_\mathcal{L}} = \frac{1}{|\mathcal{L}|}\sum_{x\in\mathcal{L}}\sum_{y\in\mathcal{Y}}\widehat{P}_\mathcal{L}(y|x)\log\widehat{P}_\mathcal{L}(y|x) \tag{15}$$

for Log loss and

$$\hat{\mathcal{E}}_{\widehat{P}_\mathcal{L}} = \frac{1}{|\mathcal{L}|}\sum_{x\in\mathcal{L}}\left(1 - \max_{y\in\mathcal{Y}}\widehat{P}_\mathcal{L}(y|x)\right) \tag{16}$$

for 0–1 loss. Then we vote this verification by

$$V = \begin{cases} 1 & \text{if } \Delta\mathcal{E} < 0 \\ -1 & \text{otherwise} \end{cases} \tag{17}$$

In other words, we compare the error of the learner trained from the original data set and the learner trained from the set which includes the instance predicted by the oracle. If the latter has a lower error than the former, we will accept the oracle's prediction; otherwise, we will reject the oracle's prediction.

In order to reduce variance and avoid overfitting, we adopt bootstrap sampling (Breiman, 1996) to sample several times. For the original labeled training set, which has size of n, we generate a new training data set by sampling n times, and the test data set is made of the remaining instances in the sampling process. The new classifier is trained based on this new sampling data set. The same process is repeated t rounds. We will vote for each sampling round according to Eq. (17) and calculate a final score as follows:

$$score = sign\sum_{i\leqslant t} V_i \tag{18}$$

If $score > 0$ we accept $(x^*, \hat{y}^*)$; otherwise, we will reject $(x^*, \hat{y}^*)$.

3.4. The algorithm

Algorithm 1 shows the detailed process of the proposed active learning paradigm with uncertain knowledge. The algorithm first selects the instance x^* to optimize the objective function (Lines 4–6), and then queries the label of instance x^* from the oracle $\mathcal{O}$. If the oracle $\mathcal{O}$ does not have certain labeling information, the verification process (Lines 7–17) will be triggered.

More specifically, if the oracle $\mathcal{O}$ does not know the label of x^*, the algorithm will collect the prediction $\hat{y}^*$ of the given instance by the oracle $\mathcal{O}$ (Line 8), and then verifies the prediction $\hat{y}^*$ of the given instance x^* by using error-reduction sampling estimation (Line 9). The instance $(x^*, \hat{y}^*)$ is accepted only if it results in expected error reduction based on current labeled data set $\mathcal{L}$ in several sampling sets, as defined by the score in Eq. (18) (Lines 10–12), otherwise, the algorithm will reject instance $(x^*, \hat{y}^*)$ and does not include it into the training set. Meanwhile, the process on Line 13 will include x^* into $\mathcal{B}^-$, so the uncertain knowledge modeling process (Section 3.2) can accurately characterize the oracle's knowledge in the next round.

Our query strategy is a traditional pool-based active learning framework, where the computational complexity of our algorithm is $O(n^2)$. We use the empirical risk minimization (ERM) strategy for verifying the guess from the labeler. This is another computationally expensive part that takes $O(m^2)$, where m is the number of nodes in the labeled data set, for each uncertain answer. In the active learning processing, the size m of labeled data is small, with $m \ll n$.

Algorithm 1. Active Learning with Uncertain Labeling Knowledge

Require: (1) Unlabeled instances set: $\mathcal{U}$; (2) the oracle $\mathcal{O}$; (3) A learner $\hbar(.)$; and (4) The number (or the percentage) of instances required to be labeled by the oracle (*reqLabeled*)
Ensure: Labeled instance set $\mathcal{L}$

$\mathcal{L} \leftarrow$ Randomly label a tiny potion of instances from $\mathcal{U}$
numLabeled $\leftarrow |\mathcal{L}|$; *numQueries* $\leftarrow 0$
while *numLabeled* $\leqslant$ *reqLabeled* **do**
  $\hbar(\mathcal{L}) \leftarrow$ Train a leaner from labeled set $\mathcal{L}$
  $\hbar(\mathbf{f}_{\mathcal{O}}(\mathcal{B}), (\mathcal{B})) \leftarrow$ Model the oracle $\mathcal{O}$'s knowledge
  Calculate optimal x^* via Eq. (5)
  if the labeler answers "I don't know the label" **then**
    $\hat{y}^* \leftarrow$ Predicted by the oracle $\mathcal{O}$
    Calculate *score* via Eq. (18)
    if *score* > 0
      $\mathcal{L} \leftarrow \mathcal{L} \cup (x^*, \hat{y}^*)$
    end if
    $\mathcal{B}^- \leftarrow \mathcal{B}^- \cup x_{i^*}$
  else
    $\mathcal{L} \leftarrow \mathcal{L} \cup (x^*, \hat{y}^*)$; $\mathcal{B}^+ \leftarrow \mathcal{B}^+ \cup x^*$
    numLabeled $\leftarrow$ *numLabeled* $+ 1$
  end if
  $\mathcal{U} \leftarrow \mathcal{U} \setminus x_{i^*}$
  numQueries $\leftarrow$ *numQueries* $+ 1$
end while

4. Experiments

We evaluate the performance of the proposed method based on a real-world data set and four benchmark data sets. The real-world text annotation data set (Rzhetsky et al., 2009) was labeled by annotators with uncertain knowledge. The four benchmark data sets, including Vertebral Column, Liver Disorders, Ecoli, and Blood Transfusion, are downloaded from the UCI Machine Learning Repository (Frank and Asuncion, 2010). For real-world text annotation data set, the labelers' uncertain knowledge is explicitly given in each labeled instance (detailed information will be introduced in Section 4.2). For benchmark data sets, we generate synthetic oracles to simulate an active learning scenario involving oracles with uncertain knowledge. To investigate the empirical performance of our approach, we compare our method with several baseline active learning methods. Because there is no existing work considering the same problem setting for one oracle with uncertain knowledge, we adopt following algorithms for comparison:

- **EIAL:** The proposed active learning method which selects instances by considering the instance's entropy and the oracle's uncertain knowledge, and learns from the uncertain information by using error-reduction sampling estimation approach.
- **PIAL:** An active learning method which selects instances according to Eq. (5). When handling an uncertain instance with label predicted by the oracle, PIAL accepts the instances if the labeler's certain (which is provided by the oracle or by the classifier learned from labeled instances) is greater than a threshold value (we use 0.75 for our benchmark data sets).
- **INAL:** An active learning method which selects instances by using Eq. (5). There is no error-reduction-based sampling estimation, so all uncertain instances (and their labels predicted by the oracle) are ignored.
- **TRAL:** Traditional active learning approach, which merely chooses the most informative instances for labeling (*i.e.*, disregard the uncertain knowledge of the oracle).
- **RAND:** Randomly choose instances for querying. There is no active learning process involved.

4.1. Experimental settings

We use 10-fold cross-validation in our experiments and report the average results. For each algorithm, we randomly label a small data set (3% of train data set) to kick off the active learning process. Then the instances from the unlabeled data set are selected to query their class labels from the labelers. In each fold of cross-validation, all methods are compared based on the same oracle, which has its own knowledge. We use logistic regression to train classifiers from instances labeled by different methods, and compare the classifiers on the same test set. To evaluate the effectiveness of different algorithms, we mainly compare the accuracies of classifiers and the labeling costs by using different methods, given a fixed number of queries (*i.e.*, the budget).

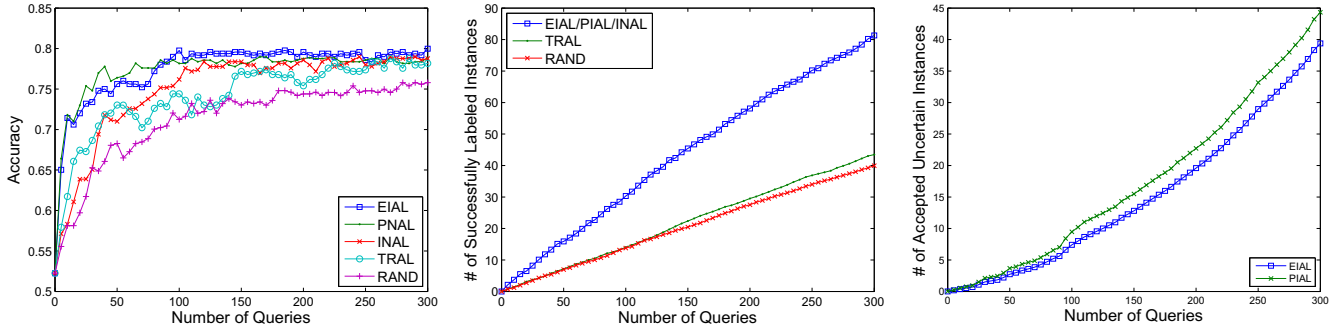
Given the same number of queries, an active learning algorithm is considered more effective if it can successfully label more instances than other methods (in our experiments, "successfully label" an instance means that the algorithm can obtain the ground truth label of the instance. This does not include the label predicted by the oracle). Accordingly, we define the number of successfully labeled instances (*numLabeled*) as the cost in the paper. In addition to evaluating the efficiency, as querying process, we also use the oracle's knowledge set to train a classifier respectively and compare the average accuracy by different methods (According to Definitions in Section 3, the oracle's knowledge set is formed by a set of labeled instances, so we can use those labeled instances to train a classifier).

4.2. Results on the real-world data set

Our real-world data set contains publicly available corpus with 1000 sentences from scientific texts annotated by annotators with explicitly specified uncertain knowledge (Rzhetsky et al., 2009). The selected sentences are annotated by experts with different background and personal knowledge. In our experiments, we choose annotator 3 as the oracle for querying. For each sentence, the labeler is required to label/mark five dimensions, including Focus, Evidence, Polarity, Certainty, and Trend, based on the labeler's personal knowledge. Among those annotations, the Polarity is described as Positive (P) and Negative (N), with a Certainty score within the range [0, 1, 2, 3] to indicate labeler's confidence. So 0 is completely uncertain and 3 is absolutely certain for both positive and negative, respectively. In other words, certainty 3 indicates that the oracle (expert) can give certain label (P or N) for the queried instance, and Certainty 0–2 indicates that expert is not exactly sure about the label of the instance but can give a prediction with reasonable possibility.

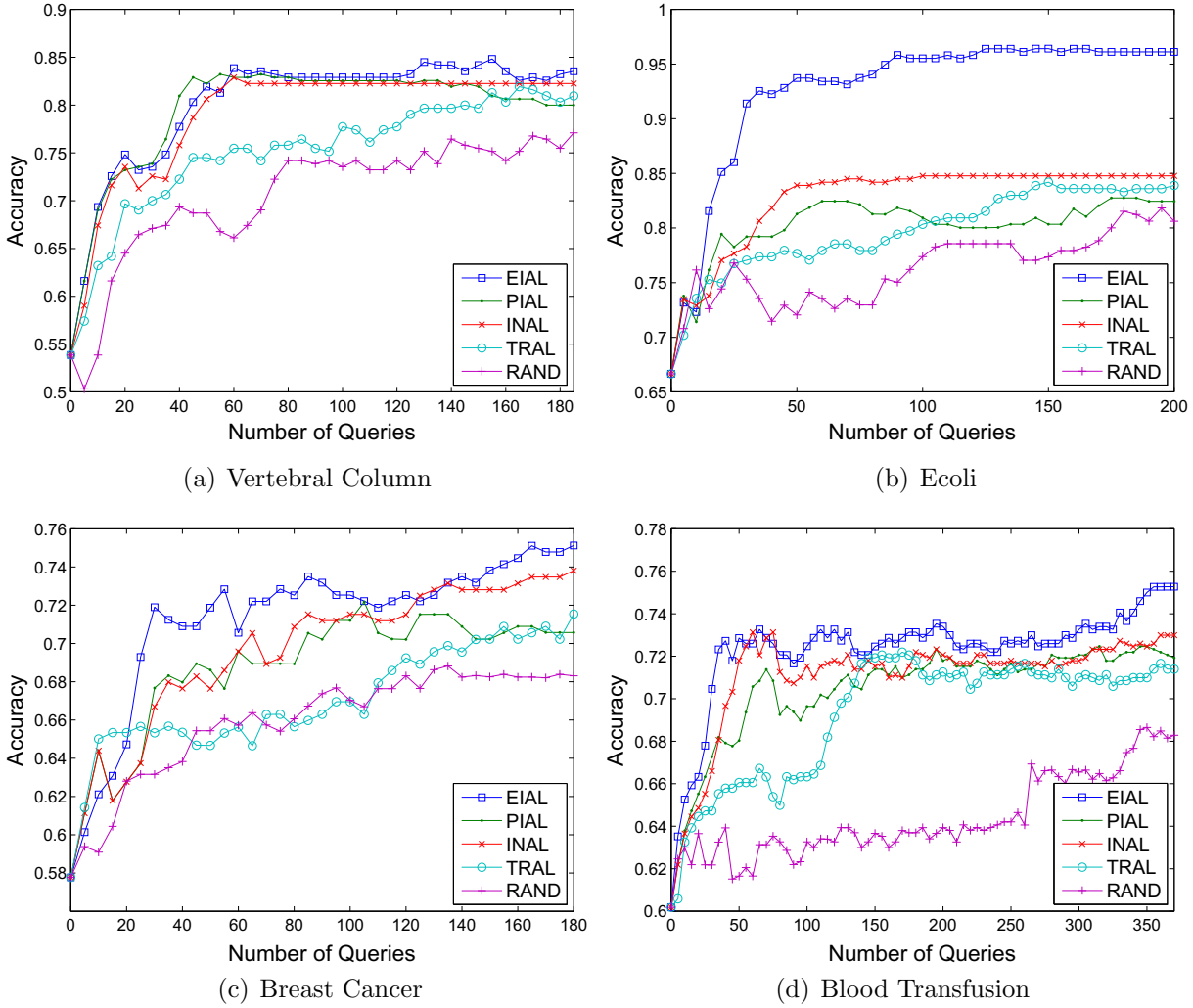
In our experiments, the active learning task is to learn a classifier for Polarity prediction (*i.e.*, a binary classification problem). We preprocess the data set and choose the fragments which have enough length. In addition, we use term frequency and its inverse document frequency (tf–idf) of the fragments to extract most common valuable words. As the result of the above process, we obtain 504 instances, each containing 153 features. During the active learning process, we will use labels and confidence scores provided by the labeler to model labeler's knowledge set and choose informative instances to query for the labels.

To test the proposed EIAL approach, we report the number of successfully labeled instances and test accuracies between different methods in Fig. 2(a) and (b). The results indicate that EIAL performs as well as PIAL and they outperform all others methods. In addition, EIAL, PIAL, and INAL have the same number of successfully labeled instances and they also have the largest number of



(a) Text:Polarity (b) Text:Polarity (c) Text:Polarity

Fig. 2. (a) The accuracy (*y*-axis) *vs.* the number of active learning iterations (*x*-axis). (b) The number of successfully labeled instances (*y*-axis) *vs.* the number of active learning iterations (*x*-axis). And (c) The number of accepted instances *vs.* the number of active learning iterations (*x*-axis).



(a) Vertebral Column (b) Ecoli

(c) Breast Cancer (d) Blood Transfusion

Fig. 3. The accuracies of the classifiers (*y*-axis) trained from the data set $\mathcal{L}$ labeled by different methods *w.r.t.* different number of queries (*numQueries*).

labeled instances because they both choose the instances based on the same method. Because TRAL and RAND do not have any treatment to handle instances falling into the uncertain knowledge of the oracle, the number of successfully labeled instances by these two methods are much smaller than EIAL, PIAL, and INAL.

EIAL and INAL both have the same mechanism to characterize the uncertain knowledge of the oracle and their major difference is that EIAL employs an error-reduction-based approach to utilize uncertain labels predicted by the oracle whereas INAL inherently ignores the uncertain instances. Our results in Fig. 2(a) and (b) clearly show that although both methods have the same labeled instances, EIAL's accuracy is better than INAL, which confirms the benefits of utilizing uncertain information provided by the oracle for active learning. Meanwhile, we notice that INAL's performance

is better than TRAL. This is because INAL avoids to query instances which belong to the oracle's uncertain knowledge. This asserts that properly modeling oracle's knowledge is beneficial for active learning.

Comparing PIAL to EIAL and INAL, PIAL accepts instances when the oracle's certainty is greater than a threshold (we set the threshold as certainty 2 for the real-world data set). Fig. 2(c) shows that while the number of accepted instances by PIAL is greater than EIAL, EIAL still has a better accuracy than PIAL. This indicates that simply relying on oracle's certainty of prediction to include an uncertain instance into the ladled set is risky (because it may include incorrectly labeled instances which, in turn, severely deteriorates the learner performance (Zhu and Wu, 2004)). By using error-reduction validation, EIAL can carefully select high quality uncertain instances predicted by the oracle and avoid including mislabeled noisy instances into the training set.

4.3. Results on benchmark data sets

In this subsection, we report the algorithm performance on four UCI benchmark data sets. Because these data sets are not labeled by the oracle with uncertain knowledge, we generate one synthetic oracle with limited knowledge sets, for each data set, to simulate our active learning with uncertain labeling knowledge scenarios. In our experiments, we use k-means clustering algorithm to cluster the data into three subsets and randomly choose one cluster for the oracle, with the setting that instances in the selected cluster can be accurately labeled by the oracle. For instances in the remaining two clusters, they are regarded as the oracle's uncertain knowledge, which means that the oracle does not have sufficient knowledge to label instances in these two clusters, but can only guess the label for instances in the remaining two clusters.

To closely simulate real-world environments where oracles may have incorrect, unknown, and uncertain knowledge, for instances belonging to the oracle's uncertain knowledge (*i.e.*, the remaining two clusters in the above analysis), we employ the following three approaches to simulate oracles' response:

- **Incorrect knowledge:** We randomly choose 10% instances to form incorrect knowledge. When an instance in the incorrect knowledge is sent to the oracle, the oracle will return a random label with random certainty value.
- **Unknown knowledge:** We randomly choose 20% instances to form unknown knowledge. When an instance in the unknown knowledge is sent to the oracle, the oracle will return an unknown label with random certainty value.

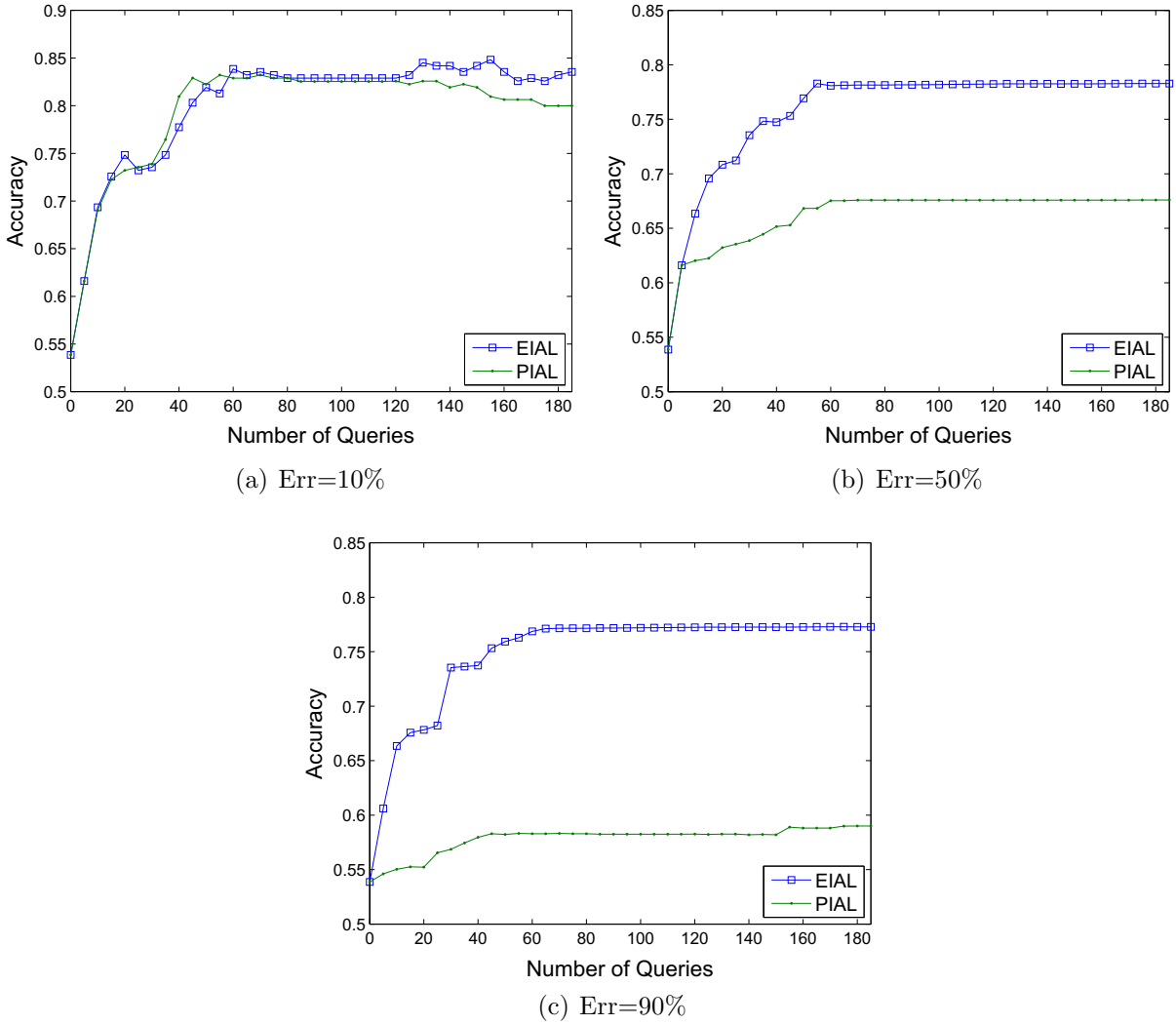


Fig. 4. The average accuracies of each labelers' classifiers (y-axis) trained from corresponding data set $\mathcal{L}$ during the process of querying, given different percentage of erroneous labels, for the vertebral column data set.

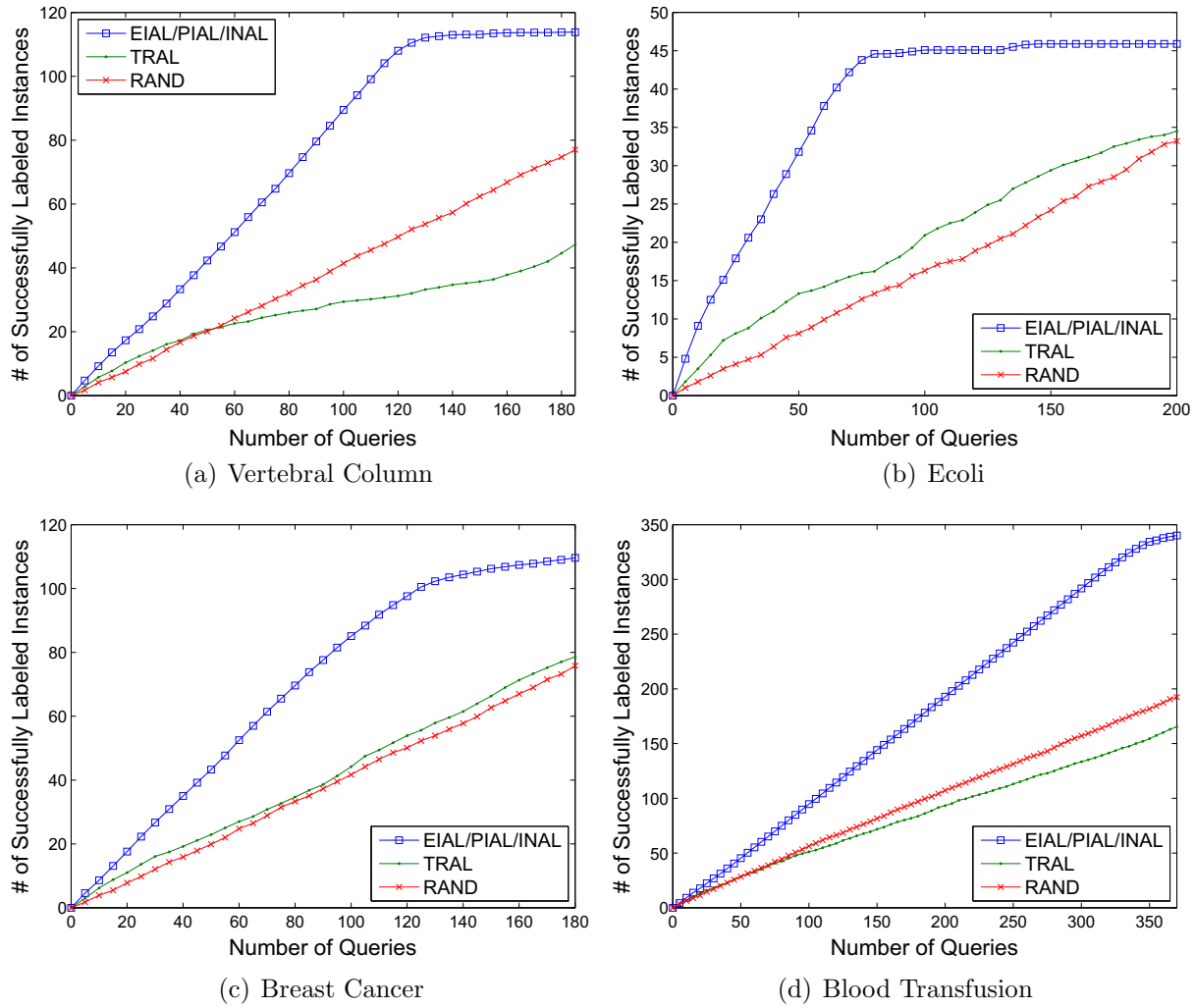
- **Uncertain knowledge:** For the remaining 70% of instances (which form the uncertain knowledge), the oracle will return a predicted label (along with the highest class probability value) based on the classifier trained from the labeled instances of the oracle.

In other words, we assume that the oracle only knows the ground true labels of the instances in the selected cluster, and the instances in the remaining two clusters belong to the oracle's uncertain knowledge. We then build a classifier f based on the instances in the cluster assigned to the oracle (this is the existing knowledge of the oracle). For 70% of the instances in the remaining two clusters, the oracle will predict a label based on the classifier f. For 20% of the instances in the remaining two clusters, the oracle can not provide any label information about the querying instance. For 10% of the remaining instances, the oracle will make a mistake and provide a random label.

4.3.1. Classification accuracies

Figs. 3(a)–(d) report the learning curves of the classifiers trained from instance sets labeled by different active learning approaches. The results demonstrate that EIAL has the best performance among all methods, which assert that information from uncertain data can help improve the classification accuracy. Comparing PIAL with other methods (except RAND), PIAL appears to be more unstable. One possible reason is that although PIAL can utilize uncertain instances through a simple strategy, simply accepting labels of the labeled uncertain instances predicted by the labeler may result in class errors and deteriorate the classification accuracy (Zhu and Wu, 2004). In Fig. 4, it shows that PIAL is affected by the percentage of erroneous labels. The more the erroneous labels the worse performance. However EIAL is still better in all the cases. This is because verifying strategy keeps the good quality of labeling. As a result, its performance varies significantly depending on the amount of noisy data included in the labeled data set.

Among all methods, RAND has the worst performance and TRAL shows slightly better results than RAND. Indeed, while TRAL follows the active learning principle to query the most informative instances, it may query instances which the oracle can not provide a ground truth label. As a result, its performance is inferior to EIAL, PIAL, and INAL. While it is true that the oracle may predict incorrect labels for uncertain instances, the error-reduction sampling estimation approach, employed by the EIAL, can effectively avoid including incorrectly predicted instances. It indicates that although an oracle may have uncertain labeling knowledge, the error reduction sampling can help refine uncertain instances to improve the active learning. In comparison, PIAL accepts instances by using the certainty values provided by the oracle. It may result in more noisy instances to be included in the training set and deteriorate the classifier performance.



(a) Vertebral Column (b) Ecoli

(c) Breast Cancer (d) Blood Transfusion

Fig. 5. The number of successfully labeled instances (y-axis) trained from corresponding data set $\mathcal{L}$ labeled by different methods during the process of querying.

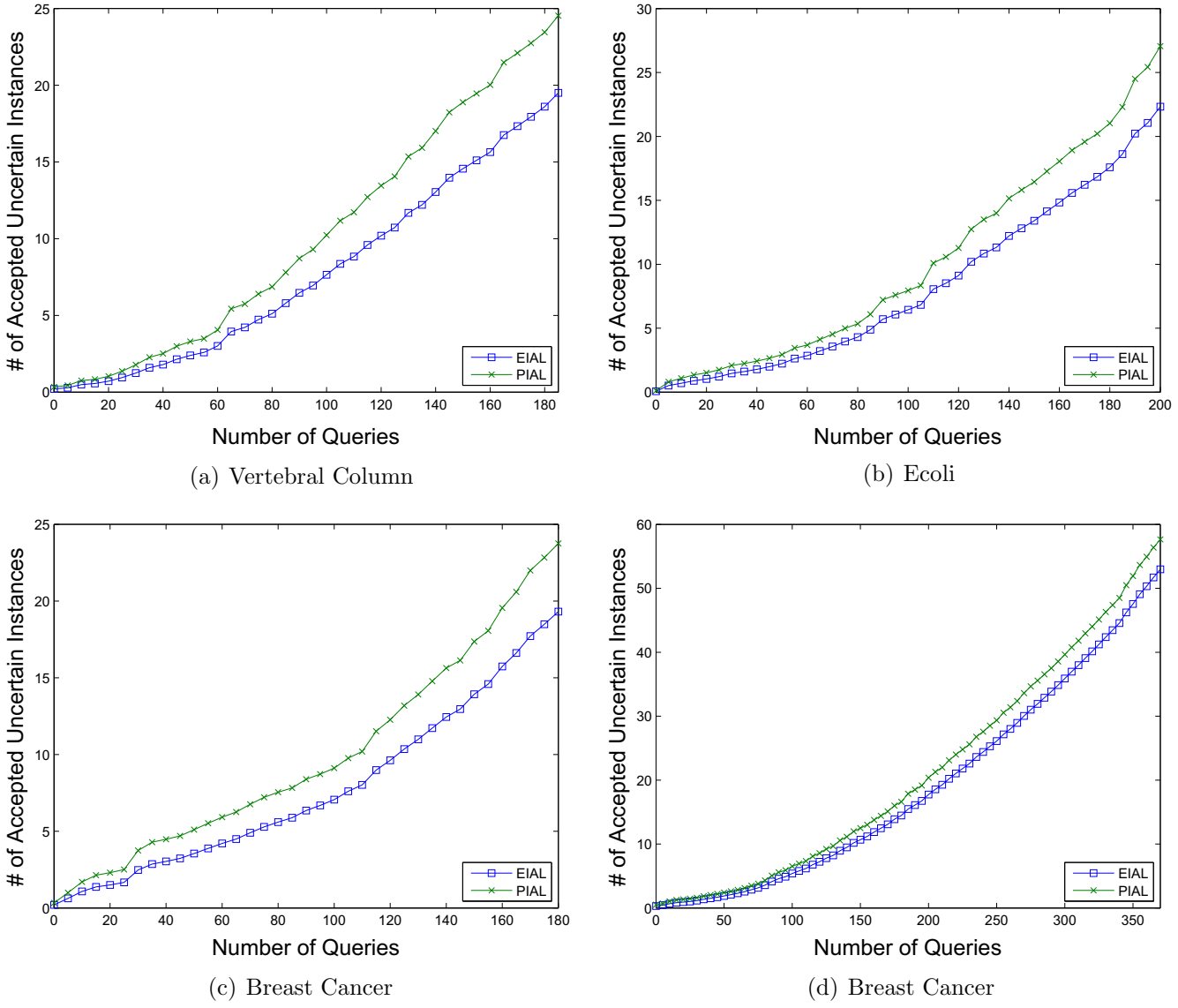


Fig. 6. The accepted of uncertain instances (y-axis) verified by error estimation from corresponding data set $\mathcal{L}_o$ during the process of querying.

Figs. 5(a)–(d) report the number of successfully labeled instances by using different methods, which show that EIAL, PIAL, and MIAL have the most successful labeled instances. TRAL and RAND do not consider the uncertain knowledge of oracle, so they may query instance which the oracle may do not know the labels and, in turn, reduce the number of successfully labeled instances.

Figs. 6(a)–(d) report the number of accepted uncertain instances by EIAL and PIAL, which demonstrate that although the number of accepted instances by PIAL is larger than EIAL, EIAL still has a better performance as shown in Fig. 3. For each querying, our EIAL has the verification process which can refuse the wrongly labeled instance. However, PIAL tends to accept instance that will be wrongly labeled. This observation indicates that error-reduction sampling estimation can help select better quality instances than PIAL.

4.3.2. Evaluation of the uncertain label utilization

In order to evaluate the effectiveness of the proposed uncertain label utilization approach which is used to decide whether to accept (or reject) the uncertain instances predicted by the oracle, we report the average error rate during the reject/accept process. Because we do know the ground truth labels of all instances, we can validate the error rate of reject/accept process by comparing with the ground true labels. Two types of mistakes during the re-

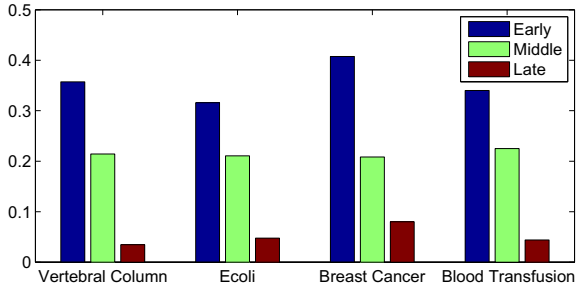


Fig. 7. The verification error rates with respect to different time stages during the whole active learning process. The error rate (y-axis) calculates that out of all instances rejected and accepted by our approaches the percentage of the errors resulted from: (1) accepting a wrongly predicted label by the oracle, and (2) rejecting correctly predicted label by the oracle. The three stages (Early, Middle, and Late) corresponds to the beginning, the middle, and the late stage of the active learning process.

ject/accept process include (1) the algorithm accepts a wrongly predicted label by the oracle, and (2) the algorithm rejects correctly predicted label by the oracle.

In our experiments, we collect all verified instances and divide them into three segments (groups) with respect to the querying time periods: early stage, middle stage, and late stage. They all have the same interval and we compute the error rate based on the same intervals. Fig. 7 presents the results for each benchmark data sets. The results shows that the verification error rapidly declines from early stage to late stage. This is mainly because that as the size of labeled data set increases, it provides more information about the data distributions which helps reduce variance and overfitting and can produce better classifier for verification. The error rates of four data sets are below 10% which indicates that our verification can maintain low error rates.

5. Conclusion

In this paper, we formulated a new active learning paradigm where the oracle, or the labeler used for labeling, may be incapable of labeling some query instances. When querying for the label of an instance, for which the oracle does not know the true label, the oracle can only provide a guessed label which could be wrong. So the active learning goal, in our new setting, is to carefully select most informative instances which the oracle is highly capable of labeling. To achieve the goal, the major challenge is to (1) properly characterize the uncertain knowledge of the oracle as well as (2) carefully utilize the instance labels predicted by the oracle to improve the active learning. In the paper, we used diverse density to model the oracle's uncertain knowledge, and combined the entropy of each unlabeled instance and its likelihood of belonging to the uncertain knowledge to select instances for labeling. Meanwhile, because instance labels predicted by the oracle could be wrong, we proposed to use error-reduction sampling estimation to either accept or reject the oracle's prediction on the uncertain instances. Experiments and comparisons demonstrate that the proposed design can result in a much higher success rate in obtaining ground truth label for each queried instance. Results also demonstrate that the quality of labeled instance set is better than other baseline approaches.

References

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Chen, Y., Bi, J., Wang, J.Z., 2006. MILES: multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12), 1931–1947. http://dx.doi.org/10.1109/TPAMI.2006.248.

Cohn, D., Ladner, R., Waibel, A., 1994. Improving generalization with active learning. Machine Learning, 201–221.

Dekel, O., Shamir, O., 2009. Good learners for evil teachers. In: Proceedings of ICML, New York, NY, USA, pp. 233–240.

Donmez, P., Carbonell, J.G., 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08. ACM, New York, NY, USA, ISBN 978-1-59593-991-3, pp. 619–628.

Frank, A., Asuncion, A., 2010. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>.

Freund, Y., Seung, H.S., Shamir, E., Tishby, N., 1997. Selective sampling using the query by committee algorithm. Machine Learning, 0885-6125 28, 133–168. http://dx.doi.org/10.1023/A:1007330508534.

Lewis, D.D., Gale, W.A., 1994. A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94. Springer-Verlag New York Inc., New York, NY, USA, ISBN 0-387-19889-X, pp. 3–12.

MacKay, D.J.C., 1992. Information-based objective functions for active data selection. Neural Computation, 0899-7667 4, 590–604. http://dx.doi.org/10.1162/neco.1992.4.4.590.

Maron, O., Lozano-Pérez, T., 1998. A framework for multiple-instance learning. In: Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10, NIPS '97. MIT Press, Cambridge, MA, USA, ISBN 0-262-10076-2, pp. 570–576.

Rashidi, P., Cook, D., 2011. Ask me better questions: active learning queries based on rule induction. In: Proceedings of the 17th Annual International ACM SIKDD Conference on Knowledge Discovery and Data Mining, KDD '11, San Diego, USA.

Raykar, V.C., Yu, S., Zhao, L.H., Jerebko, A., Florin, C., Valadez, G.H., Bogoni, L., Moy, L., 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, New York, NY, USA, pp. 889–896.

Roy, N., McCallum, A., 2001. Toward optimal active learning through sampling estimation of error reduction. In: Proceedings of the 18th International Conference on Machine Learning, ICML '01. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ISBN 1-55860-778-1, pp. 441–448.

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

Schohn, G., Cohn, D., 2000. Less is more: active learning with support vector machines. In: Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ISBN 1-55860-707-2, pp. 839–846.

Settles, Burr, 2010. Active Learning Literature Survey. University of Wisconsin, Madison.

Sheng, V.S., Provost, F., Ipeirotis, P.G., 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of KDD, pp. 614–622.

Tong, S., Koller, D., 2002. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 1532-4435 2, 45–66.

Tuia, D., Muñoz-Marí, J., 2013. Learning user's confidence for active learning. IEEE Transaction on Geoscience and Remote Sensing 51, 872–880.

Yan, Y., Rosales, R., Fung, G., Dy, J., 2010. Modeling multiple annotator expertise in the semi-supervised learning scenario. In: Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10). AUAI Press, pp. 674–682.

Yan, Y., Rosales, R., Fung, G., Dy, J., 2011. Active learning from crowds. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 1161–1168.

Yan, Y., Rosales, R., Fung, G., Farooq, F., Rao, B., Dy, J., 2012. Active learning from multiple knowledge sources. Journal of Machine Learning Research–Proceedings Track 22, 1350–1357.

Zhu, X., Wu, X., 2004. Class noise vs. attribute noise: a quantitative study of their impacts. Artificial Intelligence Review 22 (3), 177–210.