21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

I don't know the label: Active Learning with Blind Knowledge

Meng Fang and Xingquan Zhu
Centre for Quantum Computation & Intelligent Systems, FEIT, University of Technology Sydney
Meng.Fang@student.uts.edu.au, Xingquan.Zhu@uts.edu.au

Abstract

Active learning traditionally assumes that the oracle is capable of providing labeling information for each query instance. In reality, the oracle might have no information for some queries and cannot provide accurate label but only answers "I don't know the label". We focus on this problem and provide a unified objective function to ensure that each query instance submitted to the oracle is the one mostly needed for labeling and the oracle should also have sufficient knowledge to label. Experimental results on real-world and benchmark data sets demonstrate the effectiveness of the proposed design for supporting active learning using oracles with blind knowledge.

1. Introduction

Obtaining labeling information for instances is normally a time consuming process with expensive costs. Instead of labeling the entire training set or a randomly selected instance subset, active learning [3] represents a family of methods which select most informative instances for the oracle to label. Most existing active learning methods rely on a strong assumption that the oracle is perfect and can provide correct labels for each queried instances.

In reality, it is possible that the oracle may have insufficient knowledge to label some instances [9]. For example, Figure 1 demonstrates a possible situation on which an oracle may not have sufficient knowledge for labeling. For a queried instance x, the oracle can either correctly label it (if the oracle has sufficient labeling knowledge) or answer "I don't know the label" for x (if x falls into the oracle's blind knowledge). The inherent challenges associated to the oracle with blind knowledge is twofold: (1) how to characterize the oracle's blind knowledge; and (2) how to select the instance for the oracle by considering both information of instance and the oracle's blind knowledge. Recently, several works argue that it is too strong to assume that oracles may always behave perfectly . Some works are to repeatedly acquire labels from multiple sources/labelers [5] [10]. Some works try to learn the quality of a labeler in tandem with learning values of classifier parameters [4] [8] [11]. Those works' assumption is that there are lots of cheap labelers and they can try to avoid uncertainty labels by asking labelers several times or choosing other labelers. In our problem setting, we only consider one imperfect oracle where answering each query is subject to a certain amount of costs. In addition, although existing methods realize that oracles might be weak and noisy, they have no mechanism to characterize the oracle's knowledge. As a result, they have no solution to avoid an oracle's weakness, and would still require all oracles to label instances which may be out of the oracle's domain knowledge.

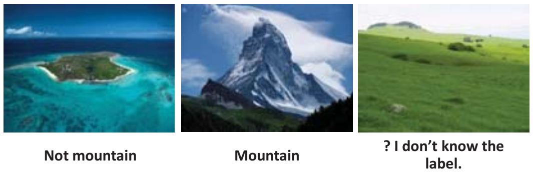


Figure 1: Examples of image labeling with insufficient labeling knowledge.

Motivated by the above observations, we propose, in this paper, a mutual information theory based framework to query the instance which has the maximum mutual information according the knowledge of the oracle. To characterize the oracle's blind knowledge, we use diverse-density framework to transform instances into a new feature space, through which we can accurately assess the likelihood of each unlabeled instance belonging to the oracle's blind knowledge. Experiments demonstrate that, given a fixed query number, the proposed method can label more instances and also result in more accurate classification models than the baseline methods.

2. Approach

In this section, we first formulate the problem and define the objective function, and then propose a method for knowledge characterization. The algorithm framework is reported in Section 2.3 .

2.1. Information-theoretic Model

Consider a data set with n instances $\{x_1, \ldots, x_n\}$, where the label for the ith instance is denoted by y_i. In a generic active learning setting, the oracle is able to provide label for every queried instance, so the objective of the uncertainty sampling based active learning [3] is to query the instance with the highest entropy (*i.e.* uncertainty). So given the labeled data, we have

$$\underset{x_i \in \mathcal{U}}{argmax}\, H(y_i; \hbar(\mathcal{L})) \tag{1}$$

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 from a classifier $\hbar(.)$ trained from labeled set $\mathcal{L}$.

In real-world scenarios, such as image or scientific text annotation, the oracle may have limited knowledge or blind knowledge and cannot provide correct labels for some instances.

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.

The expected entropy of an unlabeled instance x_i with respect to sets $\mathcal{B}^+$ and $\mathcal{B}^-$ is given by

$$\begin{aligned} H(y_i; \hbar(\mathcal{L})) = {} & P(x_i \in \mathcal{B}^+) H(y_i | x_i \in \mathcal{B}^+; \hbar(\mathcal{L})) \\ & + P(x_i \in \mathcal{B}^-) H(y_i | x_i \in \mathcal{B}^-; \hbar(\mathcal{L})) \end{aligned} \tag{2}$$

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

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

If the oracle does not know the label for an instance x_i (*i.e.*, x_i falls into the blind 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. That is 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, *i.e.*

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

Combining the oracle's knowledge set and the instance's information, the objective function in Eq.(1) can be rewritten as

$$\underset{x_i \in \mathcal{U}}{argmax}\, P(x_i \in \mathcal{B}^+) H(y_i | x_i \in \mathcal{B}^+; \hbar(\mathcal{L})) \tag{5}$$

which represents the trade-off between minimizing the probability of falling into the oracle's blind knowledge set and maximizing the entropy of instance.

2.2. Characterizing Blind Knowledge

To estimate $P(x_i \in \mathcal{B}^+)$ in Eq. (5), we employ the diverse density concept [7] to build knowledge model for the oracle.

We assume that some regions or concept set $\mathcal{C}$ exist to represent an oracle's blind knowledge. We then define the diverse density of the target concept $\mathcal{C}$ as the probability that concept $\mathcal{C}$ is the target concept given the acquired knowledge set ($\mathcal{B}^+$) and the observed knowledge blind set ($\mathcal{B}^-$) of the oracle.

$$DD(\mathcal{C}) = P(\mathcal{C} | \mathfrak{b}_1^+, \mathfrak{b}_2^+, \ldots, \mathfrak{b}_p^+, \mathfrak{b}_1^-, \mathfrak{b}_2^-, \ldots, \mathfrak{b}_q^-) \tag{6}$$

Assume target concept set $\mathcal{C}$ consists of a number of small concepts $\mathcal{C} = \{c_1, \ldots, c_m\}$, 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 $\mathfrak{b}_\tau$ [2]. Then the feature for $\mathfrak{b}_\tau$ is defined as

$$\begin{aligned} \mathbf{f}_{\mathcal{C}}(\mathfrak{b}_\tau) &= [f_{c_1}(\mathfrak{b}_\tau), \ldots, f_{c_m}(\mathfrak{b}_\tau)]^T \\ &= [P(c_1 | \mathfrak{b}_\tau), \ldots, P(c_m | \mathfrak{b}_\tau)]^T \end{aligned} \tag{7}$$

To estimate $P(c | \mathfrak{b}_\tau)$ for various concept classes, the most-likely-cause estimator is defined as

$$P(c_k | \mathfrak{b}_\tau) \;\propto\; \text{đ}(c_k, \mathfrak{b}_\tau) = exp(-\frac{\|c_k - \mathfrak{b}_\tau\|^2}{\sigma^2}) \tag{8}$$

We can use a sign function to define new labeling information for all instances in $\mathcal{B}$ as $\textbf{ł}(\mathcal{B}) = [sign(\mathfrak{b}_1^+), \ldots, sign(\mathfrak{b}_p^+), sign(\mathfrak{b}_1^-), \ldots, sign(\mathfrak{b}_q^-)]^T$. As a result, we form a well defined binary classification problem as

$$P(x_i \in \mathcal{B}^+) = \hbar(\mathbf{f}_{\mathcal{C}}(\mathcal{B}), \textbf{ł}(\mathcal{B}))[\mathbf{f}_{\mathcal{C}}(x_i); 1] \tag{9}$$

where $\hbar(.)[\mathbf{f}_{\mathcal{C}}(x_i); 1]$ denotes the class distribution of the classifier $\hbar(.)$ in classifying $\mathbf{f}_{\mathcal{C}}(x_i)$ into class "1" and one can use any learning algorithm to train $\hbar(.)$.

2.3. Active Learning with Blind Knowledge

Algorithm 1 lists major steps of the proposed framework for active learning with blind knowledge. In

Algorithm 1 Active Learning with Blind Knowledge

```
Input: (1) Unlabeled instances set: $\mathcal{U}$; (2) The oracle $\mathcal{O}$; (3)
    A learner $\hbar(.)$; and (4) The number of instances required
    to be labeled by the oracle $\mathcal{O}$ $(reqLabeled)$
Output: Labeled instance set $\mathcal{L}$
 1: $\mathcal{L} \leftarrow$ Randomly label a tiny potion of instances from $\mathcal{U}$
 2: $numLabeled \leftarrow |\mathcal{L}|$;  $numQueries \leftarrow 0$
 3: $\mathcal{B}^- \leftarrow \mathcal{L}$; $\mathcal{B}^+ \leftarrow \emptyset$;  $\mathcal{B} \leftarrow \mathcal{B}^+ \cup \mathcal{B}^-$
 4: while $numLabeled \leq reqLabeled$ do
 5:     $\hbar(\mathcal{L}) \leftarrow$ Train a leaner from labeled set $\mathcal{L}$
 6:     $\hbar(\mathbf{f}_{\mathcal{C}}(\mathcal{B}), \text{ł}(\mathcal{B})) \leftarrow$ Modeling the blind knowledge
 7:     for each $x_i$ in $\mathcal{U}$ do
 8:         $\mathbf{f}_{\mathcal{C}}(x_i) \leftarrow$ Transform $x_c$ to new feature space $\mathbb{R}_c$
 9:         $P(x_i \in \mathcal{B}^+) \leftarrow$ Estimate likelihood of $x_i$ belonging to $\mathcal{O}$'s blind knowledge (Eq.(9))
10:         $\mathcal{H}[x_i] \leftarrow$ Calculate expected entropy of $x_i$ (Eq.(2))
11:     end for
12:     $i^* \leftarrow argmax_{x_i \in \mathcal{U}} \mathcal{H}[.]$
13:     $y_{i^*} \leftarrow$ Query the label of $x_{i^*}$ from the oracle $\mathcal{O}$
14:     if the oracle answers “I don’t know the label” then
15:         $\mathcal{B}^- \leftarrow \mathcal{B}^- \cup x_{i^*}$;
16:     else
17:         $\mathcal{L} \leftarrow \mathcal{L} \cup (x_{i^*}, y_{i^*})$;    $\mathcal{B}^+ \leftarrow \mathcal{B}^+ \cup x_{i^*}$
18:         $numLabeled \leftarrow numLabeled + 1$
19:     end if
20:     $\mathcal{U} \leftarrow \mathcal{U} \setminus x_{i^*}$;  $\mathcal{B} \leftarrow \mathcal{B}^+ \cup \mathcal{B}^-$  Update knowledge $\mathcal{B}$
21:     $numQueries \leftarrow numQueries + 1$
22: end while
```

each query, the algorithm builds a benchmark learner $\hbar(\mathcal{L})$ from labeled instance subset to estimate uncertainty of each unlabeled instance and builds a classifier $\hbar(\mathbf{f}_{\mathcal{C}}(\mathcal{B}), \text{ł}(\mathcal{B}))$ to model oracle $\mathcal{O}$'s blind knowledge (Lines 5-6). The instance x_{i^*} with the largest utility value is selected and submitted to the oracle to query for the label (Lines 7-12). If the oracle $\mathcal{O}$ answers that it does not know the label for x_{i^*}, the algorithm will include x_{i^*} into the oracle $\mathcal{O}$'s blind knowledge set ($\mathcal{B}^-$). The knowledge model of the oracle will be updated for the next query (Line 20).

3. Experiments

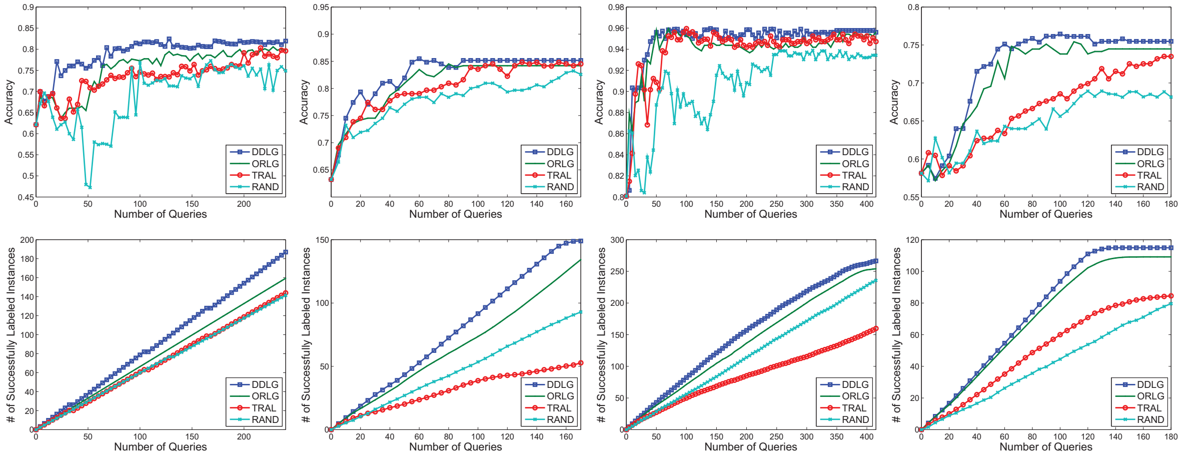
We report the empirical study results of the proposed method based on a real-world data set [1] and three UCI benchmark data sets [6]. We use 10-fold cross validation for our experiments and report average results in the paper. In our experiments, we compare our method with several baseline active learning methods: **DDLG:** Our framework: instances are transformed into new feature space $\mathbb{R}_c$ by using diverse density, and $P(x_i \in \mathcal{B}^+)$ is calculated by using logistic classifier; **ORLG:** We use instances in $\mathcal{B}^+$ and $\mathcal{B}^-$ to train a logistic regression model, in the original feature space, through which we can calculate $P(x_i \in \mathcal{B}^+)$; **TRAL:** A traditional active learning algorithm [3] only considers the uncertainty of the instance; and **RAND:** An algorithm which randomly selects instances to ask for the oracle's labels. There is no mechanism for handling blind knowledge of the oracle in TRAL and RAND.

Nature scene data set. The data set was first used in pattern classification problem [1]. We use the Mountain category and transfer the data set into a binary classification problem (mountain vs. no-mountain). All instances containing multiple labels are regarded as the samples belonging to the labeler's blind knowledge. In the experiment, we first converted images into CIE Luv color space with 294 dimensional feature vector [1], and then collect 2,407 images and extract 142 features by using Principle Component Analysis.

In Figure 2(a), the proposed DDLG maintains the best performance with respect to the classification accuracy and the number of successfully labeled instances. The second best methods are ORLG and TRAL with ORLG slightly better than TRAL. RAND shows the worst performance. The above results assert that by avoiding instances which belong to the blind knowledge of the oracle, our method can acquire sufficient labeled instances for training which helps build an accurate classifier. Although ORLG also has the mechanism to avoid the blind knowledge, DDLG is better than ORLG which works on the original feature space. We believe that this explains the rationality of diversity density for modeling oracle's blind knowledge. For the number of successfully labeled instance, there is a little difference between TRAL and RAND because they do not have any mechanism to avoid the blind knowledge and do not have enough labeled instances to train a good classifier. Because TRAL always try to select the most informative instances, it is still better than RAND.

Benchmark data sets. For benchmark data sets, we use a synthetic approach to simulate oracles with limited knowledge sets. In our experiments, we use k-means to cluster the data into three subsets and randomly choose one cluster for the oracle and assume that all instances in this cluster can be accurately labeled by the oracle. By doing so, we leave instances in the other two clusters as the oracle's blind knowledge, which means that the oracles can not provide true labels for instances in the remaining two clusters.

The results in Figures 2(b)–(d) demonstrate that, overall, DDLG outperforms all other methods for the accuracy and the number of successfully labeled instances. In most cases, the accuracy of ORLG is slightly better (or much better in Figure 2(b)(d)) than TRAL mainly because the latter does not have effective mechanisms to avoid a labeler's blind knowledge although



(a) Scene:Mountain (b) Vertebral Column (c) Breast Cancer (d) Haberman's Survival

Figure 2: The accuracies of the classifiers (upper ones) and number of successfully labeled instances (lower ones) trained from the data set $\mathcal{L}$ labeled by different methods *w.r.t.* different number of queries (*numQueries*).

it does have an active learning module. In addition, Figure 2(b)–(c) show that the number of labeled instances in TRAL is less than ORLG and the accuracy of TRAL is much better than RAND. The results show that active learning does improve the model performance compared with random sampling. By avoiding a labeler's blind knowledge, the DDLG can acquire most labeled instances than other baseline methods for the same number of queries. An active learner with a proper modeling of the blind knowledge shows clear benefits to improve the classification accuracy.

4. Conclusion

We formulated a new active learning paradigm where the oracle used for labeling may be incapable of labeling some query instances. The active learning goal, in our new setting, is to carefully avoid the oracle's blind knowledge and select the most informative instances for labeling. In the paper, we used diverse-density framework to model the oracle's blind knowledge, and combined the uncertainty of each unlabeled instance and its likelihood of belonging to the blind knowledge to select instances for labeling. Empirical results demonstrate that our proposed design can model the oracle's blind knowledge for active learning.

Acknowledgments: This work is supported by the Australian Research Council (ARC) Future Fellowship under Grant No. FT100100971.

References

[1] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[2] Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-Instance Learning via Embedded Instance Selection. *IEEE Transactions on PAMI*, 28:1931–1947, 2006.

[3] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *J. Artif. Int. Res.*, 4:129–145, 1996.

[4] O. Dekel and O. Shamir. Good learners for evil teachers. In *Proceedings of ICML*, pages 233–240, New York, NY, USA, 2009.

[5] P. Donmez and J. G. Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of CIKM*, pages 619–628, New York, NY, USA, 2008.

[6] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[7] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Proceedings of NIPS*, pages 570–576, 1998.

[8] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322, 2010.

[9] A. Rzhetsky, H. Shatkay, and W. J. Wilbur. How to get the most out of your curation effort. *PLoS Comput Biol*, 5:e1000391, 2009.

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

[11] Y. Yan, R. Rosales, G. Fung, and J. Dy. Active learning from crowds. In *Proceedings of ICML*, pages 1161–1168, 2011.