Active Learning for Crowdsourcing Using Knowledge Transfer

Meng Fang[†] and **Jie Yin**[‡] and **Dacheng Tao**[†]
[†]Centre for Quantum Comp. & Intelligent Sys, University of Technology, Sydney, Australia
[‡]Computational Informatics, CSIRO, Australia
Meng.Fang@student.uts.edu.au, Jie.Yin@csiro.au, Dacheng.Tao@uts.edu.au

Abstract

This paper studies the active learning problem in crowdsourcing settings, where multiple imperfect annotators with varying levels of expertise are available for labeling the data in a given task. Annotations collected from these labelers may be noisy and unreliable, and the quality of labeled data needs to be maintained for data mining tasks. Previous solutions have attempted to estimate individual users' reliability based on existing knowledge in each task, but for this to be effective each task requires a large quantity of labeled data to provide accurate estimates. In practice, annotation budgets for a given task are limited, so each instance can be presented to only a few users, each of whom can only label a few examples. To overcome data scarcity we propose a new probabilistic model that transfers knowledge from abundant unlabeled data in auxiliary domains to help estimate labelers' expertise. Based on this model we present a novel active learning algorithm that: a) simultaneously selects the most informative example and b) queries its label from the labeler with the best expertise. Experiments on both text and image datasets demonstrate that our proposed method outperforms other state-of-the-art active learning methods.

Introduction

Active learning is an effective way of reducing labeling effort in various data mining tasks by selectively labeling the most informative instances. Conventional active learning algorithms have mainly relied on an omniscient labeler or oracle to provide the correct label for each query. More recently, crowdsourcing is becoming an increasingly important methodology for collecting labeled data, since there is a need to label large-scale and complex data. This has fostered the development of new active learning frameworks that make use of crowdsourced annotations from multiple imperfect labelers (Sheng, Provost, and Ipeirotis 2008; Yan et al. 2011). Crowdsourcing systems, such as the Amazon Mechanical Turk (AMT), have made it easy for a large number of labelers around the world to perform labeling tasks at low cost. The simple majority vote approach has been widely used by crowdsourcing services to aggregate the most reliable label for each instance.

In these crowdsourcing settings, different annotators provide the labels based on their own expertise and knowledge in specific domains. Since annotators usually have different competencies with respect to a given task, data labels from different annotators may be very noisy or, in some cases, inaccurate. The labels provided by less competent labelers are more error-prone. Consequently, simply taking the majority vote without considering the reliability of different labelers can adversely impact on subsequent classification. To solve this problem, previous studies have focused on estimating user reliability by individually assessing knowledge for each task (Yan et al. 2010; Fang et al. 2012; Dekel, Gentile, and Sridharan 2012). However, in real-world applications, due to the limited budget for a given task each instance is often only presented to a few labelers, each of whom can only annotate a few instances. This is particularly true for active learning, where the learning process usually starts with a very small amount of labeled data, with very few annotations from each labeler. With such insufficient data, existing approaches may fail to infer labelers' expertise accurately and hence degrade the classification accuracy of active learning.

Although labeled data can be very difficult and expensive to obtain in certain domains, fortunately there often exists an abundance of unlabeled data from other different, but related domains. One typical example is in image classification, where an image needs to be classified according to its visual content, for instance, whether it contains a seagull or not. There may be very few images specifically labeled as seagulls, but it is cheap and easy to download a large number of unlabeled images of other types of birds or animals from numerous information sources. Although these images may originate from different domains they still contain basic visual patterns, such as edges and areas, that are similar to those in images of seagulls. Similarly, for text classification, there may only be a few blog documents labeled with blog types, but collecting a large quantity of mainstream news from online sources is easy. Although news and blog documents are from different domains it is very likely that they share common latent topics. Therefore, if we can learn to recognize the patterns common to two domains they can be used to help model the labelers' expertise and facilitate supervised learning tasks in the target domain.

Motivated by these observations, we propose a novel probabilistic model that addresses the active learning problem in crowdsourcing settings, where multiple cheap labelers work together for data annotation. The proposed model effectively transfers knowledge from unlabeled data in related domains to help model the expertise of labelers; in this way active learning in the target domain is enhanced. Specifically, our approach uses sparse coding (SC) techniques to learn a succinct, higher-level feature representation of the inputs shared by the two domains (e.g., visual patterns in images or latent topics in texts). These patterns serve as a higher-level abstraction of the target data and better reveal the labelers' hidden labeling expertise; this helps to reduce errors in estimating the labelers' expertise when labeled data are limited in the target domain. Based on this model we present a new active learning algorithm that simultaneously determines the most informative instance to label and the most reliable labeler to query, making active learning more effective. Experiments on real-world data demonstrate that our proposed method outperforms other state-of-the-art multi-labeler active learning methods, and transferring knowledge across domains boosts the performance of active learning.

Related Work

According to the query strategy used, there are three categories of active learning techniques: 1) uncertainty sampling (MacKay 1992; Tong and Koller 2002), which focuses on selecting the instance that the classifier is most uncertain about; 2) query by committee (Freund et al. 1997; Melville and Mooney 2004), in which the most informative instance is the one that a committee of classifiers find most disagreement; and 3) expected error reduction (Roy and McCallum 2001), which aims to query instance that minimizes the expected error of the classifier. Most existing studies have focused on a single domain and have assumed that an omniscient oracle always exists that provides an accurate label for each query.

More recently, exploiting the "wisdom of crowds" has gained attention for the purpose of multi-labeler learning (Raykar et al. 2010); using these methods, labels are collected from multiple imperfect labelers rather than a single omniscient labeler. As a result, the labels are inherently subjective and noisy with substantial variation between different annotators. There have been attempts to improve the overall quality of crowdsourced labeling from imperfect labelers. For instance, Sheng, Provost, and Ipeirotis (2008) used repeated labeling strategies to improve the label quality inferred via a majority vote. Donmez and Carbonell (2008) introduced different costs to the labelers and selected an optimal labeler-instance pair to maximize the information gain for a pre-defined budget. These methods have assumed that the labelers' expertise is known through available domain information, such as associated costs or expert salaries.

Other attempts at multi-labeler active learning have focused on estimating the reliability of different labelers. One approach has been to learn a classifier for each labeler and approximate each labeler's quality using confidence scores (Crammer, Kerns, and Wortman 2008; Raykar et al. 2009). Other recent studies have focused on modeling individual users' reliability based on the knowledge in each task. Yan et al. (2010) directly used the raw features of instances to calculate the reliability of labelers, while Fang et al. (2012) modeled the reliability of the labelers using a Gaussian mixture model with respect to some concepts. However, these methods require large amounts of labeled data for accurate estimation of the labelers' expertise. In practice, limited budgets mean that few labeled data are available from each labeler, limiting the use of these methods.

To further reduce labeling efforts for a given learning task, transfer learning and active learning can be combined to build an accurate classifier for a target domain by utilizing labeled data from other related domains. Two studies (Shi, Fan, and Ren 2008; Saha et al. 2011) used the source domain classifier to answer the target queries as often as possible so that the target domain labelers are only queried when necessary. These methods, however, assume that the classification problem in the source domain shares the same label space with that in the target domain. In our prior work (Fang, Yin, and Zhu 2013), we used the labeled data from a similar source domain to help model the labelers' expertise for active learning. On the contrary, in this work we focus on transferring knowledge from abundant unlabeled data in auxiliary domains to help estimate the expertise of labelers and improve active learning accuracy.

Problem Definition

We consider active learning in a multi-labeler setting, where the target data $\mathcal{X}_t = \{\mathbf{x}_1^t, \cdots, \mathbf{x}_{N_t}^t\}$ and the source data $\mathcal{X}_s = \{\mathbf{x}_1^s, \cdots, \mathbf{x}_{N_s}^s\}$ exist. In the target domain there are a total of M labelers $(l_1, \cdots, l_M)$ who provide the labels for any instance $\mathbf{x}_i \in \mathbb{R}^n$. For any selected instance $\mathbf{x}_i$ we denote the labeler set as M_i, the label provided by the corresponding labeler l_j as $l_{i,j}$, and its ground truth label (unknown) as z_i. In the source domain each instance $\mathbf{x} \in \mathcal{X}_s$ is unlabeled and may be drawn from a distribution different from that of the target data. Unlike semi-supervised learning, we do not assume that the unlabeled data from the source domain is assigned to the same class labels of the learning task in the target domain.

To characterize a labeler's labeling capability, we assume that each labeler's reliability of labeling an instance $\mathbf{x} \in \mathcal{X}_t$ is determined by whether the labeler has the necessary expertise with respect to some higher-level representation $\mathbf{x}'$ of instance $\mathbf{x}$. Essentially, the higher-level representation $\mathbf{x}'$ is a learned projection of $\mathbf{x}$ onto a set of high-level patterns (e.g., visual patterns in images or latent topics in texts). In this way we define a labeler's expertise as a weighted combination of these high-level patterns and by doing so we select a labeler with the best expertise to label a selected instance.

Given the target data $\mathcal{X}_t$ and M labelers in the target domain, and a set of unlabeled data $\mathcal{X}_s$ from the source domain, the objective of active learning is to select the most informative instance from the target data pool $\mathcal{X}_t$ and to query the most reliable labeler to label the instance. In this way the classifier trained from the labeled instances has the highest classification accuracy in the target domain.

Modeling Expertise of Multiple Labelers

Modeling the expertise of multiple labelers aims to select a labeler with the best labeling capability for labeling a queried instance. We therefore propose a probabilistic graphical model to learn from multiple labelers, as shown in Figure 1. The variables X, where $\mathbf{x}_i \in X$ represents an instance, and Y, where $y_{i,j} \in Y$, denotes the label provided by labeler l_j to instance $\mathbf{x}_i$, are directly observable. The other variables—the ground truth label z_i, and a labeler's expertise $\mathbf{e}_i$—are hidden, and their values need to be inferred from observed variables.

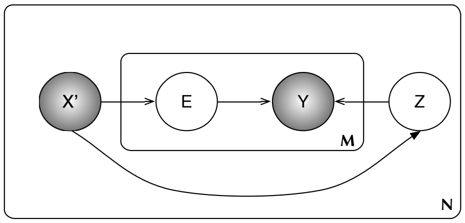


Figure 1: Probabilistic graphical model for modeling multiple labelers with different expertise. Observed variables: X' – a learned higher-level representation of X (X – instances); Y – labels provided by the labelers. Unobserved variables: E – expertise of labelers; Z – ground truth labels.

To characterize the labelers' expertise, we introduce another variable X' to denote a higher-level feature representation of instances X. Since X' is a learned projection of X it better reveals the labelers' hidden expertise with respect to some latent areas. Accordingly, our probabilistic graphical model can be represented using the following joint probability distribution:

$$p = \prod_{i}^{N} p(z_i|\mathbf{x}'_i) \prod_{j}^{M_i} p(e_{i,j}|\mathbf{x}'_i) p(y_{i,j}|z_i, e_{i,j}). \quad (1)$$

In our model, we assume that the expertise of a labeler is dependent on the higher-level representation $\mathbf{x}'_i$ of instance $\mathbf{x}_i$. Since instance $\mathbf{x}'_i$ is a new feature vector projected onto a K-dimensional space $\mathbf{x}'_i = (x_i'^{(1)}, \cdots, x_i'^{(K)})$, we represent a labeler's expertise $e_{i,j}$ as a weighted linear combination of $\mathbf{x}'_i$'s higher-level features:

$$e_{i,j} = \sum_{k=1}^{K} e_j^k {x'}_i^{(k)} + \nu_j. \quad (2)$$

Given an instance's higher-level representation $\mathbf{x}'_i$, we use logistic regression to define the expertise of labeler l_j with respect to $\mathbf{x}'_i$ as a conditional probability distribution:

$$p(e_{i,j}|\mathbf{x}'_i) = (1 + \exp(\sum_{k=1}^{K} e_j^k x_i'^{(k)} + \nu_j))^{-1}. \quad (3)$$

For an instance $\mathbf{x}'_i$, its ground truth label is assumed to solely depend on the instance itself. For simplicity, we use a logistic regression model to compute the conditional probability $p(z_i|\mathbf{x}'_i)$ as:

$$p(z_i|\mathbf{x}'_i) = (1 + \exp(-\gamma^T \mathbf{x}'_i - \lambda))^{-1}. \quad (4)$$

We also assume that for a given instance $\mathbf{x}'_i$ the actual label $y_{i,j}$ provided by the labeler l_j is subject to both the labeler's expertise $e_{i,j}$ and the ground truth label z_i of $\mathbf{x}'_i$. We model the offset between the actual label $y_{i,j}$ and the instance's genuine label z as a Gaussian distribution:

$$p(y_{i,j}|e_{i,j}, z_i) = \mathcal{N}(z_i, e_{i,j}^{-1}). \quad (5)$$

Intuitively, if a labeler has the better capability e_{ij} of labeling instance $\mathbf{x}_i$, the variance $e_{i,j}^{-1}$ of the Gaussian distribution would be smaller. Thus, the actual label $y_{i,j}$ provided by the labeler would be closer to the ground truth label z_i of $\mathbf{x}'_i$.

So far we have discussed the calculation of conditional probabilities $p(z_i|\mathbf{x}'_i)$, $p(e_{i,j}|\mathbf{x}'_i)$, and $p(y_{i,j}|z_i, e_{i,j})$ in Eq. (1). Now we focus on how to compute the higher-level representation $\mathbf{x}'$ of instance $\mathbf{x}$.

Knowledge Transfer From Unlabeled Data

We exploit transfer learning to extract useful knowledge from unlabeled data in auxiliary domains to help learning in the target domain. This meets two objectives: first, to estimate labelers' expertise more accurately (as defined in Eq. (3)); and second, to improve the performance of the active learning task. In particular, we aim to discover a succinct, higher-level feature representation of the instances that retains certain strong correlations between raw features in order to minimize the divergence of the source and target domains. By mapping the data onto a basis space, vital human characteristics are better retained from the previous data, thus making the supervised learning task easier (Raina et al. 2007; Pan and Yang 2010).

Formally, given the source data $\mathcal{X}_s$ and the target data $\mathcal{X}_t$, our objective is to learn a new feature representation $\mathbf{x}'_i$ for each instance $\mathbf{x}_i \in \mathcal{X}_t$ in the target data.

For this purpose we use SC (Olshausen and others 1996; Lee et al. 2007), which was first used as an unsupervised computational model for finding succinct representations of low-level sensory data. The basic goal of SC is to approximately represent input vectors as a weighted linear combination of a number of (unknown) basis vectors. These basis vectors thus capture high-level patterns in the input data. Specifically, given unlabeled data $\{\mathbf{x}_1, \cdots, \mathbf{x}_{N_s}\}$ with $\mathbf{x}_i \in \mathbb{R}^n$, the following optimization problem is solved:

$$\min \sum_i \|\mathbf{x}_i - \sum_j a_i^j \mathbf{b}_j\|^2 + \beta\|\mathbf{a}_i\|_1,$$
$$s.t. \ \|\mathbf{b}_j\| \leq 1, \forall j \in 1, \cdots, K. \quad (6)$$

Above, $\{\mathbf{b}_1, \cdots, \mathbf{b}_K\}$, where $\mathbf{b}_j \in \mathbb{R}^n$, are a set of basis vectors. For input data $\mathbf{x}_i$, $\mathbf{a}_i = \{a_i^1, \cdots, a_i^K\}$, where $a_i \in \mathbb{R}^K$, is a sparse vector of coefficients with each a_i^j corresponding to the basis vector $\mathbf{b}_j$. The second term is a L_1 regularization that enforces the coefficient vector $\mathbf{a}_i$ to be sparse (Ng 2004).

In our problem, we would like to find high-level patterns shared by the source and target domains. Therefore, we apply SC simultaneously to the two domains by reconstructing $\mathcal{X}_t$ and $\mathcal{X}_s$ using the same set of basis vectors. To achieve this, we define our optimization objective function as:

$$\begin{aligned} \min \sum_{\mathbf{x}_i^s \in \mathcal{X}_s} & (\|\mathbf{x}_i^s - \sum_j a_i^j \mathbf{b}_j\|^2 + \beta_1 \|\mathbf{a}_i\|_1) \\ + \sum_{\mathbf{x}_i^t \in \mathcal{X}_t} & (\|\mathbf{x}_i^t - \sum_j a_i^j \mathbf{b}_j\|^2 + \beta_2 \|\mathbf{a}_i\|_1), \\ & s.t. \|\mathbf{b}_j\| \leq 1, \forall j \in 1, \cdots, K, \end{aligned} \tag{7}$$

where $\{\mathbf{b}_1, \cdots, \mathbf{b}_K\}$ with $\mathbf{b}_j \in \mathbb{R}^n$, is a set of common basis vectors shared by the two domains. This optimization problem is convex and can be efficiently solved using iterative methods (Lee et al. 2007). We can then obtain a new feature representation of the target data, $\mathbf{x}'_i = \{a_i^1, \cdots, a_i^K\}$, as a sparse linear combination of the shared basis vectors $\{\mathbf{b}_1, \cdots, \mathbf{b}_K\}$.

Parameter Estimation

We next discuss the learning process for estimating the parameters of our proposed graphical model. Given the observed variables, including instances, their labels provided by labelers, and the higher-level representation via transfer learning, we would like to infer two groups of hidden variables $\Omega = \{\Theta, \Phi\}$, where $\Theta = \{\gamma, \lambda\}$, $\Phi = \{\mathbf{e}_j, \nu_j\}_{j=1}^M$. This learning task can be solved using a Bayesian treatment of the EM algorithm (Dempster, Laird, and Rubin 1977).

E-step: We first compute the expectation of the log-likelihood with respect to the distribution of the hidden variables derived from the current estimation of model parameters. Given current estimation of the parameters, we compute the posterior on the estimated ground truth:

$$\hat{p}(z_i) = p(z_i|\mathbf{x}', \mathbf{e}_i, \mathbf{y}_i) \propto p(z_i, \mathbf{e}_i, \mathbf{y}_i|\mathbf{x}'_i), \tag{8}$$

where

$$p(z_i, \mathbf{e}_i, \mathbf{y}_i|\mathbf{x}'_i) = p(z_i|\mathbf{x}'_i) \prod_j^{M_i} p(e_{i,j}|\mathbf{x}'_i) p(y_{i,j}|z_i, e_{i,j}). \tag{9}$$

M-step: To estimate the model parameters, we maximize the expectation of the logarithm of the posterior on z with respect to $\hat{p}(z_i)$ from the E-step:

$$\Omega^* = \underset{\Omega}{argmax}\, \mathcal{Q}(\Omega, \hat{\Omega}), \tag{10}$$

where $\hat{\Omega}$ is the estimate from the previous iteration, and

$$\begin{aligned} \mathcal{Q}(\Omega, \hat{\Omega}) &= \mathbb{E}_{\mathbf{z}}[\log(p(\mathbf{x}'_i, \mathbf{e}_i, \mathbf{y}_i|z_i))] \\ &= \sum_{i,j} \mathbb{E}_{z_i}[\log p(e_{i,j}|\mathbf{x}'_i) + \log p(y_{i,j}|z_i, e_{i,j}) \\ &+ \log p(z_i|\mathbf{x}'_i)]. \end{aligned} \tag{11}$$

We can solve the above optimization problem using the L-BFGS quasi-Newton method (Nocedal and Wright 1999) to compute the updated parameters.

Active Learning with Multiple Labelers

Based on our probabilistic model, in multiple labelers setting, active learning can select the most informative instance and the most reliable labeler to query its label.

Instance Selection Given an instance $\mathbf{x}_i$ in unlabeled pool $\mathcal{X}_t^u$ of target data, we first calculate its new feature representation $\mathbf{x}'_i = \{a_i^1, \cdots, a_i^K\}$ by solving Eq. (7). To select the most informative instance, we employ the commonly used uncertainty sampling strategy using the posteriori probability $p(z_i|\mathbf{x}'_i)$ from our graphical model:

$$i^* = \underset{i}{argmax}\, H(z_i|\mathbf{x}'_i), \tag{12}$$

where

$$H(z_i|\mathbf{x}'_i) = -\sum_{z_i} p(z_i|\mathbf{x}'_i) \log{(z_i|\mathbf{x}'_i)}. \tag{13}$$

Since the calculation of the posteriori probability $p(z|\mathbf{x}')$ takes multiple labelers and their expertise into account, the instance selected using Eq. (12) represents the most informative instance from the perspective of all the labelers.

Labeler Selection Given an instance selected using Eq. (12), labeler selection aims to identify the labeler who can provide the most accurate label for the queried instance. According to Eq. (2), we compute the confidence of each labeler as

$$C_j(\mathbf{x}'_i) = \sum_{k=1}^K e_j^k a_i^k + \nu_j. \tag{14}$$

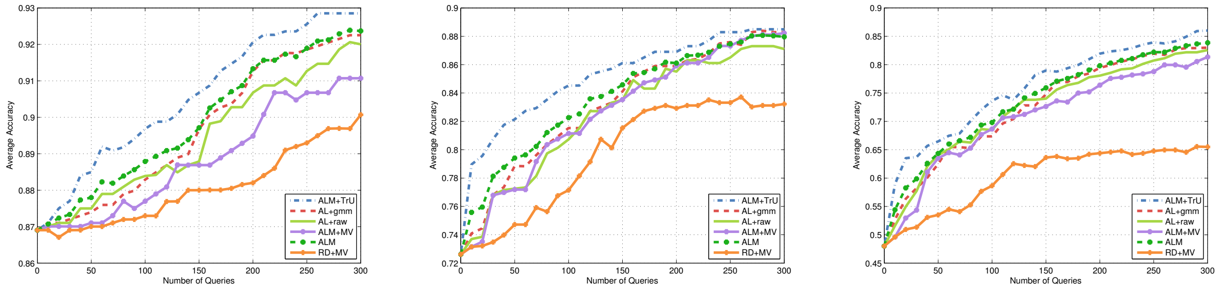
We then rank the confidence values of all the labelers and select the labeler with the highest confidence score to label the selected instance:

$$j^* = \underset{j}{argmax}\, C_j(\mathbf{x}'_{i^*}). \tag{15}$$

After selecting the best instance and labeler, we query the label for the instance. The active learning algorithm is summarized in Algorithm 1.

Algorithm 1 Active Learning with Multiple Labelers

```
Input: (1) Target dataset $\mathcal{X}_t$ (unlabeled set $\mathcal{X}_t^u$ and labeled set $\mathcal{X}_t^l$); (2) Multiple labelers $l_1, \cdots, l_M$; (3) Source dataset $\mathcal{X}_s$; and (4) Number of queries allowed by the labelers ($Budget$)
Output: Labeled instance set $\mathcal{L}$, Parameters $\Omega$
Perform transfer learning to calculate $\mathbf{x}'_i$ for each instance $\mathbf{x}_i \in \mathcal{X}_t$ (Eq. (7));
Train an initial model with labeled target data and initialize $\Omega$;
$q \leftarrow 0$;
while $q \leq Budget$ do
    Given instance $\mathbf{x}_i$, we have its new feature representation $\mathbf{x}'_i$;
    $i^* \leftarrow$ the most informative instance from unlabeled pool $\mathcal{X}_t^u$ of target data (Eq. (12));
    $j^* \leftarrow$ most confident labeler for instance $\mathbf{x}_{i^*}$ (Eq. (15));
    $(\mathbf{x}_{i^*}, y_{i^*,j^*}) \leftarrow$ query instance $\mathbf{x}_{i^*}$'s label from labeler $l_{j^*}$;
    $\mathcal{L} \leftarrow \mathcal{L} \cup (\mathbf{x}_{i^*}, y_{i^*,j^*})$;
    $\Omega \leftarrow$ retrain the model using the updated labeled data;
    $q \leftarrow q + 1$;
end while
```



(a) *polarity*

(b) *focus*

(c) *evidence*

Figure 2: Accuracy comparison of different algorithms on text data for the *polarity*, *focus* and *evidence* labels.

Experiments

To validate the effectiveness of our proposed algorithm we conducted experiments on two real-world datasets annotated by multiple labelers: the first is a corpus of scientific texts (Rzhetsky, Shatkay, and Wilbur 2009), and the other is an image dataset collected via the AMT (Welinder et al. 2010). The inconsistency between multiple labelers makes these datasets an ideal test-bed for evaluating the proposed algorithm, referred to as **ALM+TrU**. Four other algorithms are compared:

- **RD+MV** is a baseline method that randomly selects an instance to query, and uses a majority vote to generate the label for the queried instance;
- **AL+raw** is a state-of-the-art method for multi-labeler active learning that uses the raw features of the instances to calculate labelers' reliability (Yan et al. 2011);
- **AL+gmm** models a labeler's reliability using a Gaussian mixture model (GMM) with respect to some concepts, as proposed in the previous work (Fang et al. 2012);
- **ALM** learns the same probabilistic model as ALM+TrU but only using target data; it does not use transfer learning to improve expertise estimation and classification;
- **ALM+MV** uses the same transfer learning strategy as ALM+TrU to learn new bases but only uses new representations of instances for classification. Since it does not estimate labelers' expertise it relies on a majority vote to generate the label for the queried instance.

In our experiments we report average accuracies based on 10-fold cross-validation. For each run, we initially started with a small set of labeled data (30% of training data) before making queries for different active learning algorithms. We used logistic regression as the base classifier for classification and evaluated different algorithms by comparing the accuracies on the same test data.

Results Using a Text Dataset

We first carried out experiments on a publicly available corpus of scientific texts annotated by multiple annotators (Rzhetsky, Shatkay, and Wilbur 2009). This corpus has two parts: the first part, comprising 1,000 sentences annotated by five imperfect labelers, was used as the target data; and the second part, comprising 10,000 sentences annotated by eight labelers, was used as the source data for transfer learning. For labeling, each expert broke a sentence into a number of fragments and provided a label to each fragment.

For the active learning task we used the *focus*, *evidence*, and *polarity* labels in the target data and independently considered a binary classification problem for each label. The fragments were set as the instances and their annotations were treated as labels. Only the fragments that were segmented in the same way by all five labelers were retained, and the fragments with less than 10 characters were removed. The stopwords were removed from them. As a result, 504 instances containing 3,828 features (words) were available in the target data. The source data was processed in the same way but labeling information was not collected.

Label	Basis	Labelers
Scientific	cell, markers, progenitors	L1,L2,L3
	transcripts, express, neurobiology	L1,L4
Generic	express, human, cell	L1,L4,L5
	demonstrate, indeed, examined	L3,L4

Table 1: Examples of learned text bases. Left: class label; Middle: most active bases containing largest magnitude words with the class; Right: Labelers who have high expertise with respect to the bases ($e > 0.7$).

Figure 2 shows the classification accuracies of different active learning algorithms in terms of the number of queries. Our ALM+TrU algorithm consistently outperforms the other algorithms and achieves the highest accuracy. In particular, its accuracy is much higher than others at the beginning of querying, indicating that when there are a limited number of labeled data, transferring knowledge from a related domain boosts the accuracy of active learning. ALM performs slightly better than AL+raw and AL+gmm even though their performances are similar. ALM+MV is better than RD+MV showing that the learned bases are good for classification. ALM is inferior to ALM+TrU, which is intuitive because it utilizes majority vote to aggregate the labels but does not consider the reliability of different labelers.

Table 1 shows examples of learned text bases for the *focus* classification and labelers' expertise. This illustrates that the learned bases can discover the word relations for the class and that the labelers' expertise is correlated with the example bases. For example, labelers L1 and L4 have expertise on different example bases while both of them also have high expertise on the second and third example bases. This bag-of-words representation are more useful for modeling the expertise of a labeler than using raw features.

Results Using an Image Dataset

Experiments were also performed on an image dataset comprising 6,033 bird images collected via the AMT (Welinder et al. 2010). Labels were collected for two different bird species, the Black-chinned Hummingbird and the Reddish Egret, and three labels per image were obtained by three different individuals. In all, 511 (219 vs. 292) images were collected for classification and the rest of the images were used as source data for learning the bases.

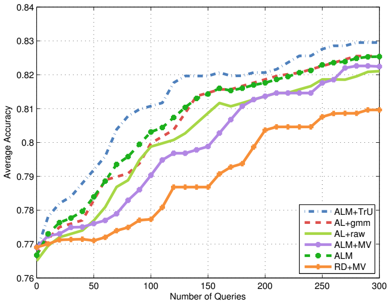


Figure 3: Comparison of classification accuracy with respect to different numbers of queries.

A comparison of the accuracies of different algorithms with respect to different numbers of queries is shown in Figure 3. Our ALM+TrU algorithm is superior to the other baselines, while RD+MV performs the worst. ALM and AL+gmm achieve higher accuracy than AL+raw, while AL+gmm is similar to ALM. This is because ALM+TrU, ALM, and AL+gmm attempt to model the expertise of labelers in terms of abstract patterns or as a multi-dimensional distribution, which better reveals the labelers' knowledge. However, since there is an underlying assumption with AL+gmm that the expertise model follows a GMM distribution its performance is limited when used with complex data. Furthermore, unlike ALM, by using unlabeled data from a related domain our ALM+TrU algorithm yields the highest classification accuracy in the active learning process.

Two example images from the dataset and some of the learned bases are presented in Figure 4. The learned bases are shown as "edges" in a small square, which is a higher-level representation and more abstract than pixels. Each labeler's expertise is shown in Figure 5. It marks three labelers' top ten highest reliability bases using different symbols □, △, and ○, respectively. It shows that labelers are more sensitive to clearer "edges" than unclear ones. For example, labeler L1 has higher reliability of the pattern at 1st row and 7th column than the one at 1st row and 6th column. Most unclear patterns are not recognized by labelers. Labelers also share some expertise of the same patterns for images. For example, labelers L1 and L2 both have high reliability of the pattern at 4th row and 2nd column, which is also a clear "edge". These abstract patterns somehow can be used to represent the labelers' expertise.

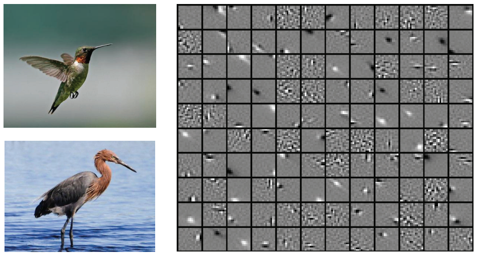


Figure 4: Left: Example images. Right: Example SC bases learned from the unlabeled source domain: 120 bases and each 20 × 20 pixels (small square).

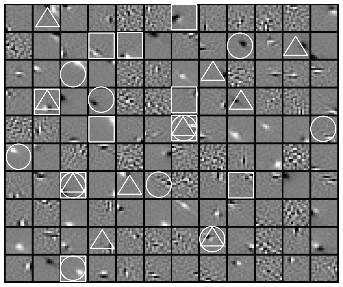


Figure 5: Comparison of expertise of three different labelers: L1-□, L2-△, L3-○. The marked basis belongs to the corresponding labeler's expertise.

Conclusion

Unlabeled data are often easy to obtain, while collecting labeling data can be expensive or time-consuming. Active learning is used to select the most informative instances to label in order to improve accuracy with as few training data as possible. However, in the multi-labeler scenario, the ground truth is unknown and even at the beginning of querying very few labeled data are available. This poses the challenge of how to model labelers' expertise with only limited labeled data and how to select both labeler and instance to make a query. To address these problems, here we have proposed a new probabilistic model for active learning, in which multiple labelers are engaged. The proposed model explicitly represents the expertise of labelers as a distribution over a higher-level representation learned by sparse coding, and unlabeled data are exploited from a related domain to help estimate the labelers' expertise and boost active learning. Based on this model our active learning algorithm can simultaneously select an optimal instance-labeler to query. Experiments on two real-world datasets demonstrate that our proposed method is more effective than existing multiple-labeler active learning methods. Transferring knowledge from a related domain can help model the expertise of labelers more accurately and improve active learning.

Acknowledgments

This work is supported by Australian Research Council Projects (FT-130101457 and DP-140102164) as well as a supplementary postgraduate scholarship from CSIRO.

References

Crammer, K.; Kerns, M.; and Wortman, J. 2008. Learning from multiple sources. *Journal of Machine Learning Research* 9:1757–1774.

Dekel, O.; Gentile, C.; and Sridharan, K. 2012. Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research* 13(1):2655–2697.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 1–38.

Donmez, P., and Carbonell, J. 2008. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of CIKM*, 619–628. ACM.

Fang, M.; Zhu, X.; Li, B.; Ding, W.; and Wu, X. 2012. Self-taught active learning from crowds. In *Proceedings of ICDM*, 858–863. IEEE.

Fang, M.; Yin, J.; and Zhu, X. 2013. Knowledge transfer for multi-labeler active learning. In *Proceedings of ECML/PKDD*. Springer. 273–288.

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

Lee, H.; Battle, A.; Raina, R.; and Ng, A. Y. 2007. Efficient sparse coding algorithms. In *Proceedings of NIPS*, 801–808.

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

Melville, P., and Mooney, R. 2004. Diverse ensembles for active learning. In *Proceedings of ICML*, 584–591. ACM.

Ng, A. Y. 2004. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In *Proceedings of ICML*, 78. ACM.

Nocedal, J., and Wright, S. 1999. *Numerical optimization*. Springer verlag.

Olshausen, B. A., et al. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607–609.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359.

Raina, R.; Battle, A.; Lee, H.; Packer, B.; and Ng, A. Y. 2007. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of ICML*, 759–766. ACM.

Raykar, V.; Yu, S.; Zhao, L.; Jerebko, A.; Florin, C.; Valadez, G.; Bogoni, L.; and Moy, L. 2009. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of ICML*, 889–896. ACM.

Raykar, V.; Yu, S.; Zhao, L.; Valadez, G.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* 11:1297–1322.

Roy, N., and McCallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of ICML*, 441–448. ACM.

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

Saha, A.; Rai, P.; Daumé III, H.; Venkatasubramanian, S.; and DuVall, S. 2011. Active supervised domain adaptation. In *Proceedings of ECML/PKDD*, 97–112. Springer.

Sheng, V.; Provost, F.; and Ipeirotis, P. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of SIGKDD*, 614–622. ACM.

Shi, X.; Fan, W.; and Ren, J. 2008. Actively transfer domain knowledge. In *Proceedings of ECML/PKDD*, 342–357. Springer.

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

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-ucsd birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Yan, Y.; Rosales, R.; Fung, G.; Schmidt, M.; Hermosillo, G.; Bogoni, L.; Moy, L.; Dy, J.; and Malvern, P. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of AISTATS*, volume 9, 932–939.

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