Knowledge Transfer for Multi-Labeler Active Learning

Meng Fang[1], Jie Yin[2], and Xingquan Zhu[1]

[1] QCIS, University of Technology, Sydney
[2] CSIRO ICT Centre, Australia
Meng.Fang@student.uts.edu.au, Jie.Yin@csiro.au,
Xingquan.Zhu@uts.edu.au

Abstract. In this paper, we address multi-labeler active learning, where data labels can be acquired from multiple labelers with various levels of expertise. Because obtaining labels for data instances can be very costly and time-consuming, it is highly desirable to model each labeler's expertise and only to query an instance's label from the labeler with the best expertise. However, in an active learning scenario, it is very difficult to accurately model labelers' expertise, because the quantity of instances labeled by all participating labelers is rather small. To solve this problem, we propose a new probabilistic model that transfers knowledge from a rich set of labeled instances in some auxiliary domains to help model labelers' expertise for active learning. Based on this model, we present an active learning algorithm to simultaneously select the most informative instance and its most reliable labeler to query. Experiments demonstrate that transferring knowledge across related domains can help select the labeler with the best expertise and thus significantly boost the active learning performance.

Keywords: Active Learning, Transfer Learning, Multi-Labeler

1 Introduction

Active learning is an effective tool for reducing the labeling costs by choosing the most informative instance to label for supervised classification. Traditional active learning research has primarily relied on a single omniscient labeler to provide a correct label for each queried instance. This is particularly true for applications involving a handful of well-trained professional labelers. Recent advances in Web 2.0 technology have fostered a new active learning paradigm [16, 21], which involves multiple (non-experts) labelers, aiming to label collections of large-scale and complex data. For example, crowdsourcing services (*i.e.*, Amazon Mechanical Turk[3]) allow a large number of labelers around the world to collaborate on annotation tasks at low cost. In such settings, data can be accessed by different labelers, who annotate the instances based on their own expertise and knowledge. Given multiple (possibly noisy) labels, majority vote is a simple but

[3] https://www.mturk.com/

popular approach widely used by crowdsourcing services to generate the most reliable label for each instance.

In multi-labeler scenarios, labelers tend to have different but hidden competence for a given task, depending on their background knowledge and expertise. Therefore, it is unlikely that all labelers are able to provide accurate labels for all instances, and labels provided by less competent labelers might be more error-prone. As a result, taking majority vote without considering the reliability of different labelers would deteriorate the classification models. More importantly, active learning starts with a small amount of labeled instances, with very few annotations from each labeler. The limited number of labeled data gives very little information to model labelers' expertise, which may incur incorrect labels and degrade classification accuracy. Therefore, accurately modeling labelers' expertise using a limited number of data poses a main challenge for active learning.

While labeled data is either costly to obtain, or easy to be outdated in a given domain, there often exists some labeled data from a different but related domain. This is often the case when the labeled data is out-of-date, but new data continuously arrives from fast evolving sources. For example, there may often be very few Blog documents annotated for certain Blog types, but there may be a lot of newsgroup documents labeled by numerous information sources. The newsgroup and Blog documents are in two different domains, but share common features (*i.e.* topics). Another example is text classification in online mainstream news. The model trained from old news articles may easily become outdated, and its classification accuracy would decrease dramatically over time. It would be very time-consuming to obtain annotations for new documents. Therefore, one important question is, how can we transfer useful knowledge from related domains to accurately model labelers' expertise in order to boost the active learning performance?

In this paper, we propose a novel probabilistic model to address the multi-labeler active learning problem. The proposed model can transfer knowledge from a related domain to help model labelers' expertise for active learning. We use a multi-dimensional topic distribution to represent a labeler's knowledge, which determines the labeler's reliability in labeling an instance. This approach provides a high-level abstraction of the labeled data in a low dimensional space, which reveals the labelers' hidden areas of expertise. More importantly, our model opens opportunity to find "good" latent topics shared by two related domains and further transfer such knowledge for improving the estimation of the labelers' expertise in a unified probabilistic framework. Based on this probabilistic model, we present a new active learning algorithm that simultaneously decides which instance should be labeled next and which labeler should be queried to maximally benefit the active learning performance. Compared with existing multi-labeler active learning methods, the advantage of our proposed method is that it can accurately model the labelers' expertise via transferring knowledge from related domains, and can thus select the labeler with the best expertise to label a queried instance. This advantage eventually leads to a higher classification accuracy for active learning.

2 Related Work

According to the query strategies, existing active learning techniques can be roughly categorized into three categories: 1) uncertainty sampling [8, 18], which focuses on selecting the instances that the current classifier is most uncertain about; 2) query by committee [6, 9], which considers the most informative instance to be the one that a committee of classifiers disagree most; 3) expected error reduction [13], which aims to query instances which can maximally reduce the model loss reduction of the current classifier once labeled. Most of existing works have mainly focused on a single domain and assumed that an omniscient oracle exists to provide an accurate label for each query.

Recently, learning from crowds has drawn a lot of research attention in the presence of multiple labelers [12]. Different from conventional supervised learning in which the annotations for data instances are provided by a single omniscient labeler, a given learning task seeks to collect labels from multiple labelers via crowdsourcing services at low cost, *e.g.*, Amazon Mechanical Turk. Since labelers have different knowledge or expertise, the resultant labels are inherently subjective (possibly noisy) with substantial variations among different annotators. Majority vote is one simple but popular way for integrating multiple noisy labels from crowdsourcing systems. Some research works have attempted to improve the overall quality of labeling from noisy labels. Sheng *et al.* [16] proposed to use repeated labeling strategies to improve the label quality inferred via majority vote. Donmez and Carbonell [4] introduced different costs to the labelers and solved a utility optimization problem to select an optimal labeler-instance pair subject to a budget constraint, in which expensive labelers are assumed to provide high-quality labels. Wallace *et al.* [19] furthered this work and proposed instance allocation strategies to better balance the workload between novice and stronger experts. These works have assumed that the labelers' levels of expertise are known through available domain information such as associated costs or expert salaries. However, the challenge of explicitly estimating each labeler's reliability has not been properly addressed.

In multi-labeler settings, active learning has focused on intelligently selecting the most reliable labelers to reduce the labeling costs. One line of research has tried to build a classifier for each labeler and approximate the labelers' expertise using confidence scores [2, 11]. Other works have proposed to estimate the reliability of labelers based on a small sample of instances labeled by all participating experts. Yan *et al.* [22] directly used raw features of instances to represent the labelers' expertise. Fang *et al.* [5] modeled the reliability of the labelers via a Gaussian mixture model with respect to some concepts. However, these methods have relied on a small set of labeled data to estimate the dependency between labelers' reliability and original instances. Instead, in our work, we model the expertise of a labeler by using a multi-dimensional topic distribution, which, at an abstract level, better represents the labeler's expertise, thus enabling each queried instance to be labeled by a labeler with the best knowledge.

Transfer learning is another learning paradigm designed to save the labeling cost for supervised classification. Given an oracle and a lot of labeled data from a source domain, some researchers have proposed to combine transfer learning and active learning to train an accurate classifier for a target domain. Saha *et al.* [15] proposed to use the source domain classifier as one free oracle, which answers the target domain queries that appear similar to the source domain data. Similarly, Shi *et al.* [17] used the classifier from the source domain to answer the queries as often as possible, and the target-domain labelers are queried only when necessary. These methods assume that the target and source domains have the exactly same labeling problem, that is, the oracle/classifier in the source domain shares a same set of labels with the target domain. Different from these works, we do not require the labeling problems in the two domains to be the same, and there is also no need to involve the source domain oracles/labelers in the active learning process. More importantly, we consider multiple labelers in the target domain and focus on transferring knowledge from the labeled source data to help estimate the expertise of labelers. To the best of our knowledge, our work is the first to leverage transfer learning to help model labelers' expertise for multi-labeler active learning problem.

3 Problem Definition & Framework

We consider active learning in a multiple labeler setting with a target data set $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ and a source data set $\mathcal{X}_s = \{\mathbf{x}_{s_1}, \cdots, \mathbf{x}_{s_{N_s}}\}$. In the target domain, there are a total of M labelers $(l_1, \cdots, l_M)$ to provide labeling information for instances $\mathcal{X}$. For any selected instance $\mathbf{x}_i$, we denote the label provided by labeler l_j as $y_{i,j}$, and its ground truth (unknown) label as z_i. In the source domain, each instance $\mathbf{x}_{s_i}$ is annotated with a label $c_i \in \{c_1, \ldots, c_D\}$ by one or multiple labelers. In this paper, we assume that the labeling problems in the source and target domain can be different. Once the data in the source domain are all labeled, there is no need to involve source domain labelers in the active learning process.

To characterize a labeler's labeling capability, we assume that each labeler's reliability of labeling an instance $\mathbf{x}_i$ is determined by whether the labeler has the expertise with respect to the latent topics, which the instance $\mathbf{x}_i$ belongs to. Formally, we give specific definitions as follows.

Definition 1 Topic: A topic t represents the semantic categorization of a set of instances. Each instance is then modeled as an infinite mixture over a set of latent topics. For example, *sports* is a common topic of a set of documents (*i.e.* instances) related to sports. A document contains words, such as “win”, “games”, “stars”, which may belong to multiple topics, such as *sports* and *music.*

Definition 2 Expertise: The expertise of a labeler l_j, denoted by $\mathbf{e}_j$, is represented as a multinomial distribution over a set of topics $\mathcal{T}$. For example, a labeler may have expertise on two topics $\{t_1 = \textit{sports}, t_2 = \textit{music}\}$, with probabilities 0.8 and 0.6, respectively.

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

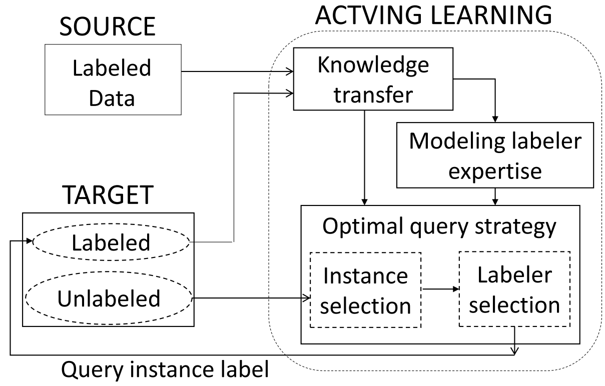


Fig. 1. An overview of the proposed framework. "Knowledge transfer module" uses source data to help model each labeler's expertise in the target domain. During active learning process, the most information instance is selected to be labeled by a labeler with the best expertise.

Proposed Framework The overview of the proposed framework is shown in Figure 1. Our goal is to select the most informative instances and find the labelers with the best expertise to label the instances. Because labeled instances are rather limited and insufficient to characterize the labelers, we leverage the data from some source domains to strengthen the active learning process. In the following, we first describe the modeling of multiple labelers by using knowledge transfer in Section 4, and then detail the active learning algorithm in Section 5.

4 Modeling Expertise of Multiple Labelers

This section details our proposed model for modeling multiple labelers and describes transfer learning techniques used to estimate labelers' expertise.

4.1 Probabilistic Model

The main aim of modeling multiple labelers is to enable the selection of a labeler with the best expertise to label a queried instance. Given an instance $\mathbf{x}$ selected for labeling and a number of labelers, each having his/her own expertise, we assume that the label y provided by each labeler to instance $\mathbf{x}$ is subject to labeler's expertise with respect to some latent topics and the ground truth label z of $\mathbf{x}$. Therefore, we propose a probabilistic graphical model, as shown in Figure 2. The two variables X, where $\mathbf{x}_i \in \mathcal{X}$, represents an instance, and Y, where $y_{i,j}$ denotes the label provided by labeler l_j to instance $\mathbf{x}_i$, are directly observable. All other variables – the topic distribution $\mathbf{t}_i$ of an instance $\mathbf{x}_i$, the ground truth label z_i, and a labeler's expertise $\mathbf{e}_j$ – are hidden, so their values must be inferred from observed variables.

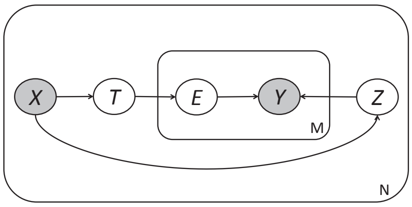


Fig. 2. Probabilistic graphical model for modeling multiple labelers with different expertise. The gray nodes X and Y are two observable random variables denoting instances and their labels, respectively. All other nodes are unobservable. For instances X, the latent topics T of instances and the expertise E of the labelers determine X's labels Y provided by the labelers, which are assumed to be an offset, subject to a Gaussian distribution, with respect to X's genuine labels Z.

This probabilistic graphical model can be represented using the joint probability distribution as follows

$$p = \prod_{i}^{N} p(z_i|\mathbf{x}_i)p(\mathbf{t}_i|\mathbf{x}_i) \prod_{j}^{M} p(e_{i,j}|\mathbf{t}_i)p(y_{i,j}|z_i, e_{i,j}). \tag{1}$$

In our model, we allow different labelers to have varying levels of expertise. That is, the expertise of a labeler depends on the topic distribution $\mathbf{t}_i$ of the instance $\mathbf{x}_i$. Because an instance can belong to one or multiple latent topics, we use $p(t_k|\mathbf{x}_i)$ to represent $\mathbf{x}_i$'s membership probability of belonging to topic t_k. Given an instance's topic distribution $\mathbf{t}_i = \{p(t_1|\mathbf{x}_i), \cdots, p(t_k|\mathbf{x}_i)\}$, we use logistic regression to define the expertise of labeler l_j with respect to $\mathbf{t}_i$ as a probability distribution given by

$$p(e_{i,j}|\mathbf{t}_i) = (1 + \exp(-\sum_{k=1}^{K} e_j^k p(t_k|\mathbf{x}_i) - \nu_j))^{-1}. \tag{2}$$

Our model assumes that the ground truth label z_i of instance $\mathbf{x}_i$ is solely dependent on the instance itself. To capture the relationships between $\mathbf{x}_i$ and z_i, any probabilistic model can be used. For simplicity, we use a logistic regression model to compute the conditional probability $p(z_i|\mathbf{x}_i)$ as

$$p(z_i|\mathbf{x}_i) = (1 + \exp(-\gamma^T \mathbf{x}_i - \lambda))^{-1}. \tag{3}$$

For an instance $\mathbf{x}_i$, the actual label $y_{i,j}$ provided by the labeler l_j is assumed to depend on 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 $y_{i,j}$ provided by the labeler and the instance's genuine label z as a Gaussian distribution

$$p(y_{i,j}|e_{i,j}, z_i) = N(z_i, e_{i,j}^{-1}). \tag{4}$$

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

So far, we have discussed the calculation of probabilities $p(z_i|\mathbf{x}_i)$, $p(e_{i,j}|\mathbf{t}_i)$, and $p(y_{i,j}|z_i, e_{i,j})$ in Eq.(1). We now focus on estimating the distribution of latent topics of the instances, i.e., $p(\mathbf{t}|\mathbf{x})$. Given a set of instances $\mathcal{X}$, the information about the latent topics is usually unavailable. A simple approach would be conducting latent semantic analysis on an initial set of labeled data. However, since the number of initially labeled data for active learning is very small, the accuracy of the induced model is largely limited. Therefore, we resort to leveraging labeled data from a related domain, which is detailed in the next subsection.

4.2 Transferring Knowledge

Given labeled data from a related domain, the basic idea is to exploit transfer learning to help discover latent topics of the instances in the target domain. That is, we aim to find common "good" latent topics to minimize the divergence of the two domains, through which we can estimate a more accurate topic distribution $p(\mathbf{t}|\mathbf{x})$, thus improving the accuracy in estimating labelers' expertise, as defined in Eq. (2). Formally, given a source data $\mathcal{X}_s$, where each instance $\mathbf{x}_s$ is annotated with a corresponding label $c \in \{c_1, \ldots, c_D\}$, our objective is to estimate a topic distribution $p(\mathbf{t}|\mathbf{x})$ for the target data $\mathcal{X}$.

For this task, we employ probabilistic latent semantic analysis (PLSA) to model the instances (*i.e.* documents) in the two domains [7]. PLSA aims to map the high-dimensional feature vectors of documents into a low dimensional representation in a latent semantic space. This abstraction offers an ideal way to represent labelers' expertise with respect to latent topics. Following the assumption that two related domains share similar topics from the terms in [20], we bridge the two domains through common latent topics, denoted by random variable T, as illustrated in Figure 3.

Specifically, we perform PLSA on the two domains. Thus, we have

$$p(\mathbf{x}_s|w) = \sum_t p(\mathbf{x}_s|t)p(t|w), \tag{5}$$

for the source data set, and

$$p(\mathbf{x}|w) = \sum_t p(\mathbf{x}|t)p(t|w), \tag{6}$$

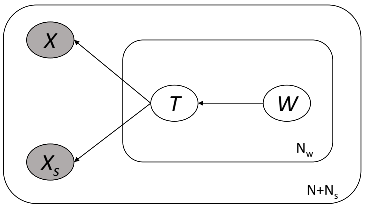


Fig. 3. PLSA model for bridging two related domains. The target data $\mathcal{X}$ and source data $\mathcal{X}_s$ are linked through latent variables T (topics) and W (terms). By transferring knowledge from the source data, a more accurate topic distribution can be obtained, which further improves the estimation of the labelers' expertise.

for the target data set. In the above equations, both decompositions share the same term-specific mixing part $p(t|w)$ and relate the conditional probabilities for the two domains; each topic has a different probability of generating a document, $p(\mathbf{x}_s|t)$, in the source domain, and $p(\mathbf{x}|t)$, in the target domain, respectively.

To fully make use of the label information in the source domain, we also enforce must-link constrains and cannot-link constrains used in semi-supervised clustering [1]. For two instances having the same label, we define the must-link constraint as

$$\text{same}(\mathbf{x}_{s_i}, \mathbf{x}_{s_j}) = \log \sum p(\mathbf{x}_{s_i}|t)p(\mathbf{x}_{s_j}|t), \tag{7}$$

and for any two instances having different labels, we define the cannot-link constraint as

$$\text{diff}(\mathbf{x}_{s_i}, \mathbf{x}_{s_j}) = \log \sum_{t_i \neq t_j} p(\mathbf{x}_{s_i}|t_i)p(\mathbf{x}_{s_j}|t_j) \tag{8}$$

Therefore, we define our objective function to maximize the log-likelihood of the data with two penalty terms:

$$\begin{aligned} L = \sum_w \Big\{ \sum_{\mathbf{x}} \log \sum_t p(\mathbf{x}|t)p(t|w) \\ + \sum_{\mathbf{x}_s} \log \sum_t p(\mathbf{x}_s|t)p(t|w) \Big\} \\ + \beta_1 \text{diff}(\mathbf{x}_{s_i}, \mathbf{x}_{s_j}) + \beta_2 \text{same}(\mathbf{x}_{s_i}, \mathbf{x}_{s_j}), \end{aligned} \tag{9}$$

where β_1 and β_2 are two hyper-parameters that control the weights of the must-link and cannot-link constrains and the question of how they would affected the accuracy of active learning will be empirically investigated.

To solve this optimization problem, we adopt a standard EM algorithm detailed as follows.

- E-step:

$$p(t|\mathbf{x}_s, w) = \frac{p(\mathbf{x}_s|t)p(t|w)}{p(\mathbf{x}_s|w)} \tag{10}$$

$$p(t|\mathbf{x}, w) = \frac{p(\mathbf{x}|t)P(t|w)}{p(\mathbf{x}|w)} \tag{11}$$

– M-step:

$$p(\mathbf{x}|t) \propto \sum_{w} n(w, \mathbf{x}) p(t|\mathbf{x}, w) \tag{12}$$

$$\begin{aligned} p(\mathbf{x}_s|t) \propto & \sum_{w} n(w, \mathbf{x}_s) p(t|\mathbf{x}_s, w) \\ & + \beta_1 \sum_{\mathbf{x}_s, c_i = c_j} \frac{p(\mathbf{x}_{s_i}|t) p(\mathbf{x}_{s_j}|t)}{\sum_t p(\mathbf{x}_{s_i}|t) p(\mathbf{x}_{s_j}|t)} \\ & + \beta_2 \sum_{\mathbf{x}_s, c_i \neq c_j} \frac{p(\mathbf{x}_{s_i}|t) p(\mathbf{x}_{s_j}|t_j)}{\sum_{t_j \neq t} p(\mathbf{x}_{s_i}|t) p(\mathbf{x}_{s_j}|t_j)} \end{aligned} \tag{13}$$

$$p(t|w) \propto \sum_{\mathbf{x}_s} n(w, \mathbf{x}_s) p(t|\mathbf{x}_s, w) + \sum_{\mathbf{x}} n(w, \mathbf{x}) p(t|\mathbf{x}, w) \tag{14}$$

Finally, for the target domain, we can calculate the latent topic distribution $p(\mathbf{t}|\mathbf{x})$ using Eq. (11).

4.3 Parameter Estimation

Now we discuss the learning process to estimate the parameters of our proposed graphical model. Given observed variables – instances, their labels provided by labelers, and topic distribution of instances estimated via transfer learning (described in Section 4.2), 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 by using a Bayesian style of EM algorithm [3].

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

$$\hat{p}(z_i) = p(z_i|\mathbf{x}_i, \mathbf{t}_i, \mathbf{e}_i, \mathbf{y}_i) \propto p(z_i, \mathbf{t}_i, \mathbf{e}_i, \mathbf{y}_i|\mathbf{x}_i), \tag{15}$$

where

$$p(z_i, \mathbf{t}_i, \mathbf{e}_i, \mathbf{y}_i|\mathbf{x}_i) = p(z_i|\mathbf{x}_i) p(\mathbf{t}_i|\mathbf{x}_i) \prod_{j}^{M} p(e_{i,j}|\mathbf{x}_i) p(y_{i,j}|z_i, e_{i,j}). \tag{16}$$

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 E-step:

$$\Omega^* = \underset{\Omega}{argmax}\, \mathcal{Q}(\Omega, \hat{\Omega}), \tag{17}$$

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

$$\begin{aligned}\mathcal{Q}(\Omega,\hat{\Omega}) &= \mathbb{E}_{\hat{p}(z_i)}\left[\sum_i \log p(\mathbf{x}_i,\mathbf{t}_i,\mathbf{y}_i|\ z_i)\right]\\ &= \sum_{i,j}\mathbb{E}_{\hat{p}(z_i)}[\log p(e_{i,j}|\mathbf{x}_i)+\log p(y_{i,j}|z_i,e_{i,j})\\ &+\log p(z_i|\mathbf{x}_i)+\log p(\mathbf{t}_i|\mathbf{x}_i)]. \qquad (18)\end{aligned}$$

To solve the above optimization problem, we compute the updated parameters by using the L-BFGS quasi-Newton method [10].

5 Knowledge Transfer for Active Learning

Based on our probabilistic model, multi-labeler active learning seeks to select the most informative instance and the most appropriate labeler, with respect to the selected instance, to query for its label.

Instance Selection The goal of active learning is to learn the most accurate classifier with the least number of labeled instances. We thus employ a commonly used uncertainty sampling strategy, by using the posteriori probability $p(z|\mathbf{x})$ from our graphical model, to select the most informative instance:

$$\mathbf{x}^* = \underset{\mathbf{x}_i\in\mathcal{X}}{argmax}\, H(z_i|\mathbf{x}_i), \qquad (19)$$

where

$$H(z_i|\mathbf{x}_i) = -\sum_{z_i} p(z_i|\mathbf{x}_i)\log{(z_i|\mathbf{x}_i)}. \qquad (20)$$

Since the calculation of the posteriori probability $p(z|\mathbf{x})$ takes multiple labelers and their expertise into consideration, the instance selected using Eq.(19) represents the most informative instance from all labelers' perspectives.

Labeler Selection Given an instance selected using Eq.(19), labeler selection aims to identify the labeler who can provide the most accurate label for the queried instance. For each selected instance $\mathbf{x}_i$, we first calculate the latent topic distribution $p(\mathbf{t}_i|\mathbf{x}_i)$ using Eq. (11), and then compute the confidence of each labeler as follows:

$$e_{i,j}(\mathbf{x}_i) = \sum_{k=1}^{K} e_j^k p(t_k|\mathbf{x}_i) + \nu_j. \qquad (21)$$

Accordingly, we rank the confidence values from Eq.(21) and select the labeler with the highest confidence score to label the selected instance

$$j^* = \underset{j\in M}{argmax}\, e_{i,j}(\mathbf{x}_i). \qquad (22)$$

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

Algorithm 1 Knowledge Transfer for Active Learning

```
Input: (1) Target data set $\mathcal{X}$; (2) Multiple labelers $l_1, \cdots, l_M$; (3) Source data set $\mathcal{X}_s$;
    and (4) Labeling budget: $budget$
Output: Labeled instance set $\mathcal{L}$, Parameters $\Omega$
Train an initial model with the labeled target data $\mathcal{L}$ and source data $\mathcal{X}_s$;
Perform transfer learning from $\mathcal{X}_s$ to calculate topic distribution $p(\mathbf{t}|\mathbf{x})$ for each
    instance $\mathbf{x}$ (Eq.(11));
$numQueries \leftarrow 0$;
while $numQueries \leq budget$ do
    $\mathbf{x}^* \leftarrow$ the most informative instance from pool $\mathcal{X}$ (Eq.(19));
    $j^* \leftarrow$ the most reliable labeler for instance $\mathbf{x}^*$ (Eq.(22));
    $(\mathbf{x}^*, y_{\mathbf{x}^*,j*}) \leftarrow$ query instance $\mathbf{x}^*$'s label from labeler $l_{j^*}$;
    $\mathcal{L} \leftarrow \mathcal{L} \cup (\mathbf{x}^*, y_{\mathbf{x}^*,j*})$;
    $\Omega \leftarrow$ retrain the model using the updated labeled data;
    $numQueries \leftarrow numQueries + 1$.
end while
```

6 Experiments

To validate the effectiveness of our proposed algorithm, we conduct experiments on both synthetic data and real-world data. Our proposed algorithm is referred to as **AL+kTrM**. For comparison, we use five other algorithms as baselines:

- **RD+MV** is a baseline method that randomly selects an instance to query. It collects all labels provided by multiple labelers and then uses majority vote to generate the label for the queried instance.
- **AL+MV** uses the same strategy as our proposed AL+kTrM algorithm to select an instance but it relies on majority vote to generate the label for the queried instance.
- **AL+rM** is a state-of-the-art multi-labeler active learning method [21]. It uses raw features of the instances to calculate the reliability of labelers.
- **AL+gM** models a labeler's reliability using a Gaussian mixture model (GMM) with respect to some concepts, as proposed by [5].
- **AL+kM** uses the same probabilistic model as our proposed AL+kTrM algorithm, but does not utilize transfer learning to improve the estimation of labelers' expertise. By comparing with this baseline, we can validate whether the transfer learning module in our AL+kTrM algorithm can help improve active learning to achieve a higher accuracy.

In our experiments, all results are based on 10-fold cross-validation. In each round, we initially started with a small labeled data set (3% of train data), and then made queries by using different active learning strategies. The reported results are averaged over 10 rounds. We used logistic regression as the base classifier for classification, and evaluated algorithms by comparing their accuracies on the same test data. For our proposed AL+kTrM algorithm, the two parameters in Eq.(9) were set as $\beta_1 = 60$, $\beta_2 = 40$, and their impact on the classification accuracy will be empirically studied in Section 6.3.

6.1 Results on Real-World Data

We carried out experiments on a real-world data set, which is a publicly available corpus of scientific texts annotated by multiple annotators [14]. The inconsistency between multiple labelers makes this data set an ideal test-bed for evaluating the proposed algorithm. This corpus consists of two parts; we used the first part of 1,000 sentences annotated by five labelers as the target data, and the second part of 10,000 sentences annotated by eight labelers as source data for transfer learning. During the original labeling process, each expert broke a sentence into a number of fragments and provided a label to each fragment.

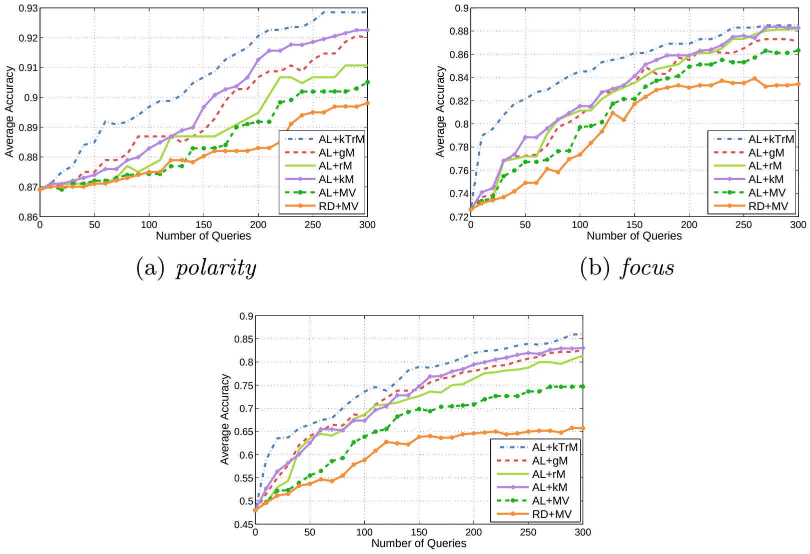
For the target data, we used the *focus*, *evidence*, *polarity* labels and considered a binary classification problem for each label. We set the fragments as the instances and their annotations were treated as labels. We removed the fragments whose number of characters is less than 10 and only kept the fragments segmented by all five labelers in the same way. The fragments were pre-processed by removing stopwords. As a result, we had 504 instances containing 3828 features (words) in the target data set. In the source data, the labels, including *generic*, *methodology*, and *science* were used to form the constraints in Eq.(9).

Figure 4 compares the classification accuracy of different algorithms with respect to the number of queries. Figure 4(a)-4(c) clearly show that AL+kTrM outperforms other baselines and achieves the highest accuracy. Particularly, at the beginning of querying, its accuracy is much higher than others. This indicates that, when there is a limited number of labeled data, transferring knowledge from a related domain boosts the accuracy of active learning. AL+kM and AL+gM perform slightly better than AL+rM, although their performance is close to each other in Figure 4(b). AL+MV and RD+MV perform worst, because they use majority vote to aggregate the labels but do not consider the expertise and reliability of different labelers. AL+MV performs better than RD+MV because it selects the most informative instance to query. Overall, AL+kTrM achieves the highest classification accuracy during the active learning process.

6.2 Results on Synthetic Data

Since real-world data does not have the ground truth information about labelers' expertise, we also evaluated the effectiveness of our algorithm using synthetic data in which we can construct different expertise domains for labelers and explicitly evaluate the accuracy of labeler selection. The synthetic data we used is based on the 20 Newsgroups[4]. This data set contains 16,242 postings that are tagged by four high-level domains: *comp*, *rec*, *sci* and *talk*. To simulate the labelers, we assume that each labeler knows the ground truth labels of two tagged sub data sets, and gives a random guess for the rest of the data. In this way, we constructed five labelers of different expertise and formulated a binary classification problem. For each domain, we selected 150 instances as the target data, and used the rest as the source data for transfer learning. We started

[4] `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html`



(a) *polarity*

(b) *focus*

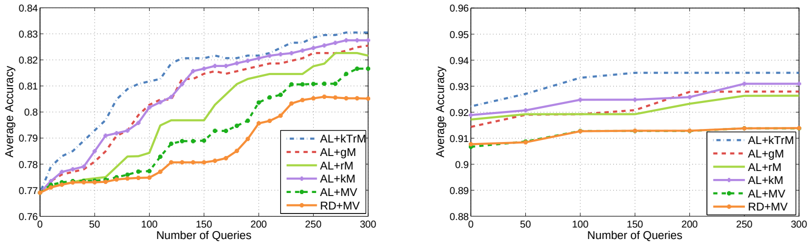
(c) *evidence*

Fig. 4. Accuracy comparison of different algorithms on scientific text data for the *polarity*, *focus* and *evidence* labels.

all active learning algorithms with an initially labeled set and made queries to improve the classification accuracy.

Figure 5(a) compares the accuracy of different algorithms with respect to the number of queries. We can see that, AL+kTrM is superior to other baselines, and RD+MV performs worst. RD+MV randomly selects the instances thus leading to the worst performance. AL+MV improves RD+MV because it selects the most informative instance to label. However, the two methods, RD+MV and AL+MV, rely on majority vote to aggregate the labels for instances without considering the reliability of labelers. AL+gM and AL+kM achieve higher accuracy than AL+rM, while AL+kM performs slightly better than AL+gM. This is because, both AL+gM and AL+kM model the expertise of labelers in terms of some topics at an abstract level, which better reveals the labelers' areas of knowledge. However, since AL+gM has a strong assumption that the expertise model follows a GMM distribution, its performance is limited in complex text data. Furthermore, by utilizing labeled data from a related domain, our AL+kTrM algorithm yields the highest classification accuracy in the active learning process.

In order to better understand how AL+kTrM models the expertise of labelers, Table 1 shows the correlation between the top two latent topics discovered by our algorithm and the domain of expertise of two labelers we constructed: Labeler 1 which is simulated to have expertise in *comp* and *rec* domain, and Labeler 2 to have expertise in *sci* and *talk* domain. The results clearly show that for Labeler 1, Topic 2 is related to *rec.sport* domain, and Topic 4 is related to *comp.sys*



(a) Classification accuracy (b) Accuracy of labeler selection

Fig. 5. Performance comparison of different algorithms

Table 1. Correlation between the labelers' expertise and the latent topics discovered by our AL+kTrM algorithm

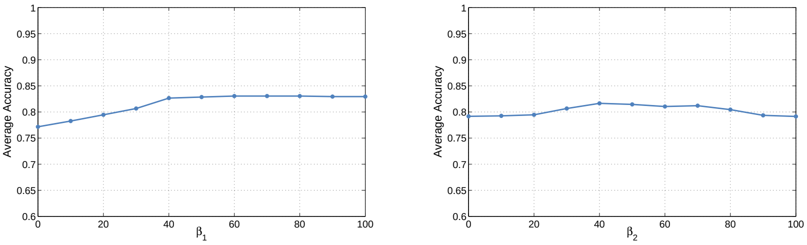
Latent topic	Top correlated words
Labeler 1 (*comp*, *rec*)	
Topic 2	win,team,games,players,baseball,season
Topic 4	drive,data,card,technology,video,driver
Labeler 2 (*sci*, *talk*)	
Topic 1	world,law,children,jews,religion,fact,war
Topic 10	question,state,research,earth,space,orbit

domain; for Labeler 2, Topic 1 and Topic 10 are correlated to *talk.religion* and *sci.space* domain, respectively. This well explains our motivation that discovering latent topics can better reveal labelers' areas of expertise.

To further demonstrate the advantage of our AL+kTrM algorithm, we explicitly compared different algorithms with respect to their abilities to select the best labelers to label the queried instances. Figure 5(b) reports the accuracy of labeler selection in terms of different numbers of queries. We can observe that, AL+MV and RD+MV performs much worse than other methods. This is because they use majority vote to aggregate labels without considering the reliability of labelers. In contrast, by modeling the labelers' expertise, multi-labeler active learning methods significantly improve majority vote. Among them, our AL+kTrM algorithm can be observed to yield the highest accuracy for selecting the best labelers to label the queried instances.

6.3 Study on the Impact of β_1 and β_2

Now we study the impact of the two parameters β_1 and β_2 in our AL+kTrM algorithm on classification accuracy. Parameters β_1 and β_2 are two coefficients in the knowledge transfer module that controls the contribution of the must-link and cannot-link constrains, as defined in Eq.(9). Specifically, we fixed the value of one parameter at 50, and studied the impact of the other parameter by varying its value from 0 to 100. We analyzed the impact of β_1 and β_2 on both synthetic data and real data. Due to the space limit, we used the synthetic data as a case study, because similar observations are obtained for real data.



(a) Different β_1 values

(b) Different β_2 values

Fig. 6. Classification accuracy with different β_1 and β_2 values

Figure 6 shows the classification accuracy by varying the values of β_1 and β_2, respectively. From Figure 6(a), we can see that, as the value of β_1 increases, AL+kTrM gradually achieves higher accuracy. When β_1 reaches the value of 60, the accuracy becomes relatively saturated. From Figure 6(b), we can observe that, the accuracy is not very sensitive to different values of β_2. Overall, the impact of the must-link constraint (controlled by β_1) seems to be larger than that of the cannot-link constraint (controlled by β_2).

7 Conclusion

This paper proposed a new probabilistic model to address active learning involving multiple labelers. We argued that when labelers have different levels of expertise, it is important to properly characterize the knowledge of each labeler to ensure the label quality. In active learning scenarios, the quantity of instances labeled by all participating labelers is very small, which raises a challenge to model each labeler's strength and weakness. So we proposed to utilize data from a related domain to help estimate labelers' expertise. Using the proposed model, our active learning algorithm can always select the most informative instance and query its label using a single labeler with the best expertise with respect to the queried instance. Experiments demonstrated that our method significantly outperforms existing multi-labeler active learning methods, and transferring knowledge from a related domain can indeed help improve active learning.

Acknowledgments: This research is sponsored by an Australian Research Council (ARC) Future Fellowship under grant No. FT100100971 and Discovery Project under grant No. DP130102748, as well as a supplementary postgraduate scholarship from CSIRO.

References

1. D. Cohn, R. Caruana, and A. McCallum. Semi-supervised clustering with user feedback. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, 4(1):17, 2003.
2. K. Crammer, M. Kerns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9:1757–1774, 2008.
3. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.
4. P. Donmez and J.G. Carbonell. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proc. of CIKM*, pages 619–628, 2008.
5. M. Fang, X. Zhu, B. Li, W. Ding, and X. Wu. Self-taught active learning from crowds. In *Proc. of ICDM*, pages 858–863. IEEE, 2012.
6. Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168, 1997.
7. T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.
8. D.J.C. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.
9. P. Melville and R. Mooney. Diverse ensembles for active learning. In *Proc. of ICML*, pages 584–591, 2004.
10. J. Nocedal and S.J. Wright. *Numerical optimization*. Springer verlag, 1999.
11. V.C. Raykar, S. Yu, L.H. Zhao, A. Jerebko, C. Florin, G.H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proc. of ICML*, pages 889–896, 2009.
12. V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.
13. Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. of ICML*, pages 441–448, 2001.
14. A. Rzhetsky, H. Shatkay, and W.J. Wilbur. How to get the most out of your curation effort. *PLoS computational biology*, 5(5):e1000391, 2009.
15. A. Saha, P. Rai, H. Daumé III, S. Venkatasubramanian, and S.L. DuVall. Active supervised domain adaptation. In *Proc. of ECML/PKDD*, pages 97–112, 2011.
16. V.S. Sheng, F. Provost, and P.G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proc. of SIGKDD*, pages 614–622. ACM, 2008.
17. X. Shi, W. Fan, and J. Ren. Actively transfer domain knowledge. In *Proc. of ECML/PKDD*, pages 342–357. ACM, 2008.
18. S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2002.
19. Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Who should label what? instance allocation in multiple expert active learning. In *Proc. of SDM*, 2011.
20. G.R. Xue, W. Dai, Q. Yang, and Y. Yu. Topic-bridged plsa for cross-domain text classification. In *Proc. of SIGIR*, pages 627–634. ACM, 2008.
21. Y. Yan, R. Rosales, G. Fung, and J. Dy. Active learning from crowds. In *Proc. of ICML*, pages 1161–1168, 2011.
22. Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, J. Dy, and PA Malvern. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proc. of AISTATS*, volume 9, pages 932–939, 2010.