Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Transfer active learning by querying committee*

Hao SHAO[1], Feng TAO[2], Rui XU[3]

([1]*School of WTO Research & Education, Shanghai University of International Business and Economics, Shanghai 200336, China*)

([2]*School of Business, East China University of Science and Technology, Shanghai 200237, China*)

([3]*School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China*)

E-mail: shaohao@suibe.edu.cn; ftao@ecust.edu.cn; rxu@ustc.edu.cn

Received June 20, 2013; Revision accepted Nov. 9, 2013; Crosschecked Jan. 15, 2014

Abstract: In real applications of inductive learning for classification, labeled instances are often deficient, and labeling them by an oracle is often expensive and time-consuming. Active learning on a single task aims to select only informative unlabeled instances for querying to improve the classification accuracy while decreasing the querying cost. However, an inevitable problem in active learning is that the informative measures for selecting queries are commonly based on the initial hypotheses sampled from only a few labeled instances. In such a circumstance, the initial hypotheses are not reliable and may deviate from the true distribution underlying the target task. Consequently, the informative measures will possibly select irrelevant instances. A promising way to compensate this problem is to borrow useful knowledge from other sources with abundant labeled information, which is called transfer learning. However, a significant challenge in transfer learning is how to measure the similarity between the source and the target tasks. One needs to be aware of different distributions or label assignments from unrelated source tasks; otherwise, they will lead to degenerated performance while transferring. Also, how to design an effective strategy to avoid selecting irrelevant samples to query is still an open question. To tackle these issues, we propose a hybrid algorithm for active learning with the help of transfer learning by adopting a divergence measure to alleviate the negative transfer caused by distribution differences. To avoid querying irrelevant instances, we also present an adaptive strategy which could eliminate unnecessary instances in the input space and models in the model space. Extensive experiments on both the synthetic and the real data sets show that the proposed algorithm is able to query fewer instances with a higher accuracy and that it converges faster than the state-of-the-art methods.

Key words: Active learning, Transfer learning, Classification
doi:10.1631/jzus.C1300167 **Document code:** A **CLC number:** TP3

1 Introduction

Nowadays, a challenging issue in nosology is that when people encounter a new epidemic, it is crucial to classify the patients as early as possible with a high accuracy. However, even though there exist thousands of suspected cases, only a few of them are labeled (diagnosed) and labeling an unlabeled instance by experts is often expensive and time-consuming. Active learning (AL) (Settles, 2010) provides a solution by selecting unlabeled examples to query, with the objective of obtaining a satisfactory classifier using as few instances as possible where the labeling cost is high. Another typical example is document classification, which requires users to label a certain number of documents but which turns out to be tedious and even unnecessary if thousands of instances are required to be labeled. In this learning scenario, a significant challenge is how to select the

* Project supported by the Humanity and Social Science Youth Foundation of Ministry of Education of China (No. 13YJC630126), the 085 Foundation of SUIBE (Nos. Z085YYJ13014 and 085LXPT13020), the Fundamental Research Funds for the Central Universities (No. WK0110000032), the National Natural Science Foundation of China (Nos. 71171184, 71201059, 71201151, 71090401, and 71090400), and the Funds for the Creative Research Group of China (No. 70821001)

most informative instance in each query from the large portion of the unlabeled pool in the target task. There have been a substantial amount of works on active learning (Settles, 2010). One representative approach is the query by committee (QBC) strategy, which assumes a correct Bayesian prior on the set of hypotheses, and the committee members are all trained on the current labeled data set (Seung *et al.*, 1992). However, as pointed out in Balcan *et al.* (2006) and Settles (2010), given only a few labeled instances in the data set, the initial hypotheses sampled from these instances are not reliable since they may deviate from the optimal hypothesis with respect to the input distribution in the end of the classification procedure. Consequently, the informative measure that is based on the initial hypotheses will possibly select irrelevant instances.

To help tackle the problem of the lack of labeled information in the target task, transfer learning (TL) techniques aim to borrow the strength of existing data and models. In the TL setting, a target data set is assumed to have only a small number of labeled instances, while abundant useful information is available in the source data sets that can be obtained with little cost. However, an essential problem in TL is that, when the distributions of the source and the target domains are different, directly transferring knowledge could hurt the performance on the target task, which is also known as 'negative transfer' (Rosenstein *et al.*, 2005). It is likely to occur if we underestimate the side effects resulting from the distribution differences of multiple source tasks, which is common in real applications. Due to the high cost of querying experts in AL, we believe that it is possible to borrow the strength of TL to effectively reduce the number of queries. An informative instance can be selected with the help of transferred information from the source domain, and therefore we can make AL more effective with the advantages of TL. However, the discrepancy between the source and the target domains should seriously be considered; otherwise, the negative transfer will lead to undesirable results. There exist several research works concerning integrating AL and multi-task learning (Reichart *et al.*, 2008; Harpale and Yang, 2010; Zhang, 2010; Zhu *et al.*, 2011; Li *et al.*, 2012) with the same assumption that all tasks are similar and related. However, for our problem with the existence of irrelevant domains, which is common in practice, this assumption may not hold. The reason is that most multi-task learning strategies concern only distribution differences in data; in such a case, the performance will possibly degenerate when we encounter data sets such as the two that have the same distribution but reversed label assignments (Figs. 2a and 2e in Section 6). Moreover, informative instances are selected to improve the overall performance on multiple tasks but without a guarantee on every single task. Consequently, a trade-off must be made between a specific task and a set of tasks, which contradicts the TL setting in which the performance of the target task is most concerned. In this study, we aim to improve the classification performance in a specific target domain. If these methods are directly applied to our problem, the classification accuracy on the specific domain may not be improved and may even degenerate. Recently, only a few approaches have been proposed to explore the feasibility of improving AL given the TL framework (Shi *et al.*, 2008; Li *et al.*, 2010). McCallum and Nigam (1998) proposed a hybrid method by combining the discriminative method and the generative method with a support vector machine (SVM), with both methods being assigned to a certain number of queries depending on their respective weights. However, without enough query samples on both positive and negative cases, the algorithm would fail to calculate the corresponding weights because the ratio of the positive instances is taken into the calculation form. Therefore, in the algorithm, the number of queries is fixed to a large value, and it turns out to be costly in real applications. The active transfer learning method in Shi *et al.* (2008) tries to use instances from the source domain to label the ones in the target domain by firstly adopting an existing AL method to induce the initial hypothesis, and then the decision function for labeling informative instances which relies on it is used to decide the instance to query. However, as discussed above, the initial hypothesis is not reliable; it may deviate from the true underlying distribution, and consequently impedes the performance on the target task. Also, there does not exist any effective strategy in these methods to avoid querying unnecessary instances, which may increase the querying cost.

As mentioned before, the QBC framework maintains a committee of models that consider the consensus probability based on some disagreement measure instead of an individual model. To utilize the

useful information from the source domain, unlike conventional QBC methods which consider only sampled models from the target task and assigned the same weights to all members, we extend the QBC active learning framework in cooperation with the transfer learning (ALTL) setting with each member as a model from the source domain, to avoid training the initial hypotheses by using the re-sampled data. Models in the source tasks are assigned with different weights related to the similarities to the target task, and the weights are updated during each of the iterations. By exploring the advantages of Kullback Leibler (KL) divergence in measuring similarities between models, which has proved to be useful in TL, we propose a weighted KL divergence measure to update the weights of the initial hypotheses that can decrease the negative effects of inferior hypotheses. We also develop an adaptive strategy which could eliminate unnecessary instances in the input space and models in the model space. In such a way, the proposed algorithm aims to query only instances of interests to avoid the negative transfer problem.

2 Related works

In this section, we review research works relevant to ours, including different methods for inductive TL that deal with negative transfer problems, related works on AL, and the hybrid methods that consider both AL and TL. ALTL belongs to the supervised inductive TL where both the source and the target tasks contain labeled data (Caruana, 1997; Shao and Suzuki, 2011). From the perspective of AL, the proposed method is the pool-based AL algorithm (Settles, 2010).

To tackle the problem of negative transfer, current works focus mainly on finding the similarities among tasks or instances. Dai *et al.* (2007) extended the AdaBoost algorithm (Freund and Schapire, 1997) which aims at improving the accuracy of a weak learner by adjusting the weights of the instances in the training sets. Their TrAdaBoost algorithm could evaluate the instances from a large amount of data in the source domain and assign weights based on the similarities to the target task to boost the accuracy of the classifier. Shi *et al.* (2009) proposed a semi-supervised learning method by extending the co-training method, which deals with the same problem as Dai *et al.* (2007). Instances that can be put into the target task are obtained by re-weighting those in the source task. Argyriou *et al.* (2008) tried to find a common representation among different groups of tasks, which can be regarded as the transferred information. However, the method is restricted to linear classification functions and the number of instances in the target task is not considered to be much smaller than that in the source tasks. The basic probabilistic latent semantic analysis (PLSA) was extended in Zhuang *et al.* (2010), to simultaneously capture both the domain distinctions and commonality among multiple domains. In Cao *et al.* (2010), a new kernel function was designed by exploiting the Gaussian process to evaluate the negative similarity between two instances, but it was designed for only one single source task and without taking multiple source tasks into the framework. Shao *et al.* (2011) proposed a compact coding method for hyperplane classifiers (CCHC) under a two-level framework for inductive TL. In particular, the degree of similarity is represented by the relevant code length of the class boundary of each source task with respect to the target task. In addition, informative parts of the source tasks are adaptively selected to make the choice of the specific source task more accurate.

AL aims to find the most informative instance and query it to an oracle. Rajan *et al.* (2006) provided an active learner that can identify the data points that change the current belief in the class distributions the most. In Muslea *et al.* (2002), AL had been applied in the multi-view setting. In the multi-view problem, features can be partitioned into subsets, each of which is sufficient for learning the mapping from the input to the output space.

Although there exist several studies concerning integrating AL and multi-task learning (Reichart *et al.*, 2008; Harpale and Yang, 2010; Zhang, 2010; Zhu *et al.*, 2011; Li *et al.*, 2012), with the objective to improve the overall performance rather than a single task, they cannot be applied to our problem where the performance of the target task is emphasized. Some research work has been done by combining AL and TL together to improve the classification accuracy on the target task (Shi *et al.*, 2008; Li *et al.*, 2010; Luo *et al.*, 2012; Chattopadhyay *et al.*, 2013; Yang *et al.*, 2013); however, the negative transfer problem was not considered explicitly. For example, Li *et al.* (2010) provided a hybrid method which combines the discriminative method and the gener-

ative method with SVM, but the query number is fixed to a large number and there is not an effective strategy to avoid selecting irrelevant instances. An AL approach, error reduction sampling (ERS) (Roy and McCallum, 2001), was integrated into TL (Shi *et al.*, 2008) with a heuristic similarity function. Shi *et al.* (2008) pointed out that experts are heavily relied on, as the possibility to query an expert is set to be higher than 50%. Moreover, the decision function that relies on the initial hypothesis of an existing AL method is not reliable due to the initial hypothesis problem. Our proposal places emphasis on not only the combination of the two learning scenarios, but also the alleviation of the problem of negative transfer. By adopting a strategy to eliminate unnecessary models and instances during the learning process, the proposed algorithm has proven to be more robust against negative transfer problems.

3 Problem statement and preliminaries

In this study, we deal with the classification problem on a target data set where several source data sets are available. There is a task set S from the source and the target domains which contain $K+1$ data sets S_i $(i=1,2,...,K,K+1)$, where the first K data sets are from the source domain and the $(K+1)$th data set is from the target domain. In the target data set, labeled data is regarded to be insufficient, denoted by L, while abundant unlabeled data is available, denoted by U. Y is the set of possible labels for the instances with d dimensions and the jth class is denoted by y_j, where $y_j \in Y=\{y_1,y_2,...,y_d\}$. The objective is to obtain the class label y_j for each instance x in the unlabeled data using a d-dimensional parametric model $P_\theta(y_j|x)$ on the target data set, and we write P_θ for convenience. Note that the model for each of the source data set S_i $(i=1,2,...,K)$ is denoted by P_{θ_i}.

AL aims to achieve satisfactory results by querying only a small number of instances from the unlabeled data set. A typical approach is to select the most informative unlabeled instance by some information measurements. It can be divided into two main categories, pool-based sampling (Dagan and Engelson, 1995) and stream-based sampling (Lewis and Gale, 1994). Stream-based AL scans the data sequentially and makes query decisions individually, while pool-based AL makes one decision each time from the entire collection of the unlabeled pool. In this study, we put our emphasis on pool-based AL. The key issue is how to measure the 'usefulness' of an unlabeled instance using only a small amount of available information. Assuming that the current model from the labeled data is denoted by $\hat{p}=\hat{p}(Y|x)$, where $x\in U$, the most informative instance is the one that maximizes the value of information (VOI) (Krause and Guestrin, 2009):

$$\mathrm{VOI}(Y,x)=\sum_y P(Y=y|x)R(\hat{p},Y=y,x). \quad (1)$$

In Eq. (1), the VOI of a labeling request (Y,x) is the sum of the reward of each possible labeling outcome $Y{=}y$ for the current model $\hat{p}$, represented by $R(\hat{p},Y{=}y,x)$, weighted by the probability of this outcome $P(Y{=}y|x)$ (Zhang, 2010). In real applications, however, the true label probability $P(Y{=}y|x)$ is generally unknown and most existing methods replace it by $\hat{p}(Y{=}y|x)$. Therefore, the VOI function could be written as

$$\mathrm{VOI}(Y,x)=\sum_y \hat{p}(Y=y|x)R(\hat{p},Y=y,x), \quad (2)$$

and there exist several heuristic methods for the reward function $R(\hat{p},Y=y,x)$. A simple but direct way is the 0/1 reward (Roy and McCallum, 2001), in which an impossible outcome has an infinite value and an already known outcome has a zero value. Another way is the log reward given as follows (Zhang, 2010):

$$R(\hat{p},Y=y,x)=-\log_2\hat{p}(Y=y|x). \quad (3)$$

An inevitable problem for this kind of method is that, if only a few labeled instances are given, the initial hypothesis for $\hat{p}$ is probably inaccurate and may lead to unsatisfactory consequences. To solve this problem, a more theoretically-motivated framework, QBC, was proposed (Seung *et al.*, 1992). In the framework, a committee of t members are maintained, all trained using re-sampled data from the labeled data L. Each member will vote on the candidates and decide the most informative instance to query. The key issue is to design an effective information measurement, which can represent the 'disagreement' of every committee member. In Dagan and Engelson (1995), the most informative instance

was shown to be the one with the maximum vote entropy, given as follows:

$$x^*_{\mathrm{VE}} = \arg\max_x -\sum_i \frac{V(y_i)}{C}\log\frac{V(y_i)}{C}, \tag{4}$$

where y_i denotes all possible labels, $V(y_i)$ is the number of votes that a label receives from the committee members' predictions, and C is the committee size.

However, one common problem in AL is that due to the lack of labeled information, it is difficult for us to obtain satisfactory models in the initial stage and useful information may be neglected due to the incompatibility with the initial models. Therefore, we consider the need to borrow the strength of TL in which the useful information in the source domain could be used in the target domain to help the query selection process.

4 Framework of ALTL

In the conventional framework of QBC, a committee of members are maintained, all trained using re-sampled data from L in the target task (Settles, 2010), and the query is chosen according to the principle of maximum disagreement. However, without solid prior knowledge of the target task, the initial model induced from the few labeled instances is possibly inappropriate and may deviate from the model in the end of the learning process (Fig. 1). In such a circumstance, the distributions of the committee members sampled from these few labeled instances will have a large discrepancy towards the distribution underlying the data set. We are motivated to borrow the strength of TL, and let each model be induced from the source tasks and the target task be a committee member, to avoid obtaining inferior initial hypotheses. However, an inevitable problem is that, if the discrepancy between the distributions of the source and the target domains is too large, a negative transfer is more likely to occur (Rosenstein *et al.*, 2005). Therefore, given the objective to find the most informative instance, we propose a novel weighted disagreement measure to deal with the 'inferior' models in the learning stage. In our framework, each model P_{θ_i} in one of the $K+1$ tasks is regarded as a committee member. Therefore, the labeled information from multiple source data sets can be directly transferred to the target task. During the learning process, the most informative instance is selected based on the disagreement on the class label of all committee members. In the conventional QBC setting (McCallum and Nigam, 1998), all members are treated equally important. However, in a TL setting, some models may have similar distributions to the target task, while some models have totally different distributions. This situation is also shown later in Fig. 2 in Section 6. Therefore, instead of assigning equal weights to every member, a more flexible way is to use different weights based on the similarities of distributions with the target task. Therefore, we first consider measuring the similarities among tasks.

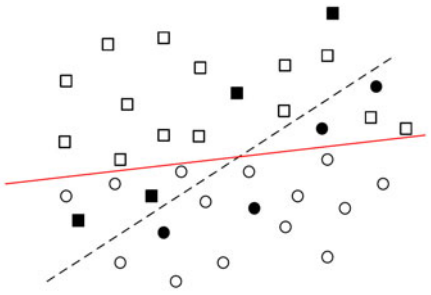

Fig. 1 Illustration of a best model given instance. If all the instances are labeled, the solid line is the optimal hyperplane. If only a few labeled instances are available, the dashed line is likely to be induced, which may deviate from the optimal one. The squares and the circles denote different classes, where solid ones are in L and empty ones are in U, respectively

KL divergence was successfully implemented in AL to measure the committee disagreement (McCallum and Nigam, 1998). It is an information-theoretic measure which captures the expected number of extra 'bits of information' required to code samples from one distribution when using a code based on the other. Therefore, we can evaluate the relevance between the two models without calculating the real bits needed for coding. In the proposed method, we add a weight w_i for each model P_{θ_i} based on the KL divergence, to measure the discrepancies between the model and the current best model P^* with the lowest error rate, given as follows:

$$w_i = \exp\left[-K(P_{\theta_i}(Y|X_L)||P^*(Y|X_L))\right], \tag{5}$$

where X_L denotes the labeled set of instances, $P_{\theta_i}(Y|X_L)$ is the class distribution of X_L of a committee member, $P^*(Y|X_L)$ is the best one among $P_{\theta_i}(Y|X_L)$, and $K(Q_1||Q_2)$ denotes the K directed

divergence (Lin, 1991) between two probability distributions Q_1 and Q_2 of a discrete random variable, defined as

$$\begin{aligned} K(Q_1||Q_2) =& \mathrm{KL}\left(Q_1||\frac{Q_1+Q_2}{2}\right) \\ =& \sum_j [Q_1(j)\log Q_1(j) \\ & - Q_1(j)\log((Q_1(j)+Q_2(j))/2)]. \end{aligned}$$

We adopt K directed divergence instead of KL divergence for the weights in our framework based on the following reason: for Q_1 and Q_2, $\mathrm{KL}(Q_1||Q_2)$ is not defined when $Q_2(j)=0$ but $Q_1(j)>0$. In our proposal, the same technical difficulty is encountered when we try to calculate $\mathrm{KL}(P_{\theta_i}(Y|X_L)||P^*(Y|X_L))$. There are basically two techniques to deal with this kind of problem. The first way is to smooth the distributions in some way such as that introduced by Church and Gale (1991), for instance, with a Bayesian prior or taking the convex combination of the observations with some valid (nonzero) distribution. The second way is to employ heuristics to discard those zero frequencies. However, as pointed out by Pereira *et al.* (1993), these methods violate the nature of the true distributions, which may lead to unsatisfactory results. Therefore, we sidestep this problem using the K directed divergence, which inherits the good properties of KL divergence while avoiding the zero frequency problems.

Note that $w_i = 1$ only when $P_{\theta_i}(Y|X_L) = P^*(Y|X_L)$. For some y satisfying $P^*(y|X_L) = 0$, it is easy to prove that the divergence function is still meaningful. Note that $w_i \leq 1$ all the time; thus, the exponential function converts the number of 'bits of information' into a scalar distance between 0 and 1.

In our framework, the most informative instance selected by the committee is given as follows:

$$x^* = \arg\max_x \sum_{i=1}^{K+1} w_i \mathrm{KL}(P_{\theta_i}(Y|x)||P_C(Y|x)), \quad (6)$$

where

$$\begin{aligned} & \mathrm{KL}(P_{\theta_i}(Y|x)||P_C(Y|x)) \\ = & \sum_j P_{\theta_i}(y_j|x)\log\frac{P_{\theta_i}(y_j|x)}{P_C(y_j|x)}, \end{aligned} \quad (7)$$

$$P_C(y_j|x) = \frac{\sum_{i=1}^{K+1} w_i P_{\theta_i}(y_j|x)}{\sum_{i=1}^{K+1} w_i}, \quad (8)$$

where $P_{\theta_i}(Y|x)$ is the class distribution of a committee member, and $P_C(y_j|x)$ is the weighted 'consensus' probability that y_j is the correct label.

The argmax function in Eq. (6) denotes that, the instance with the maximum divergence to the weighted 'consensus' probability will be chosen to query. Different from the classical QBC method in the symmetric setting which measures only the disagreement between models, the proposed method is able to balance the divergence between a single model to the current best model, and the divergence between this model to the consensus model.

5 ALTL algorithm and analysis

5.1 Procedure of ALTL

In pool-based AL algorithms, one query is selected from the large pool of unlabeled data U in each iteration. Therefore, it is necessary to reduce the region of the instance space in U. In the proposed algorithm, before selecting the most informative instance using the uncertainty measure, the region of the instance space is reduced by eliminating the instances with labels on which all committee members agree. It means that in the eliminated region, there does not exist an instance x that, for any two committee members P_{θ_i} and P_{θ_j}, the labels predicted by the two models are identical. This is reasonable because, if all members agree on an instance x, there is no necessity to query the label of this instance as its entropy is equal to zero.

The main flow of our algorithm is summarized in Algorithm 1, and the termination condition (11) in the pseudo code will be explained in Section 5.2. Note that for each iteration, when updating the committee, we add a new committee member by building a model from the labeled data L and the instance queried. Cohn *et al.* (1994) tried to reduce the model space by eliminating those models that disagree with a query in iteration i. However, due to the lack of the knowledge of the underlying distribution of the target task, a model that performs unsatisfactorily might possibly be the best model given enough instances, as shown by the solid line in Fig. 1. As

pointed out by Balcan *et al.* (2006), it is not reasonable to get rid of a model that performs badly in the current stage. In the proposed method, we keep all the models in the committee, and the final model is determined at the end of the iterations. Each model is assigned a weight, and the weights of those models deviated from the class distribution of the current best model tend to become lower. In such a way, we try to decrease the impacts of inferior models in each iteration instead of eliminating them from further consideration.

Algorithm 1 ALTL algorithm

Input: S_i $(i=1,2,...,K,K+1)$; parametric models for the source tasks $P(y|x,\theta_i)$ $(i=1,2,...,K)$

Output: Parametric model $P(y|x,\theta)$, $x\in U$

```
Build an initial model P^0_{θ_{K+1}} from the labeled data L of S_{K+1}, P* = P^0_{θ_{K+1}}
Create a committee C consisting of P_{θ_i} (i = 1,2,...,K) and P^0_{θ_{K+1}}
Set t = 1
while U ≠ ∅ and the termination condition (11) is not satisfied do
    Eliminate the instances in U with labels on which all committee members agree
    Select one instance to query by the uncertainty measure of Eq. (6)
    Update the current best model P* of the model with the smallest ε(P) on L ∪ {x*}
    L = L ∪ {x*}, U = U\{x*}
    Build a new model P^t_{θ_{K+1}} from L
    Update the committee C by C ∪ P^t_{θ_{K+1}}
    t = t + 1
end while
Output P*
```

5.2 Termination condition and analysis

A simple way adopted in most existing methods to terminate the loop of querying unlabeled instances is to set a number N as the maximum number of iterations (McCallum and Nigam, 1998; Shi *et al.*, 2008). Although in such a way, it is easy to control the number of queries in the learning procedure, we cannot guarantee the performance of the algorithm. To learn a classifier with an error less than ϵ in a noise-free environment, passive learning requires $O(\frac{1}{\epsilon})$ labels while binary search needs $O(\ln\frac{1}{\epsilon})$ labels (Balcan *et al.*, 2006). This means N should not be larger than $\frac{1}{\epsilon}$ or it will be meaningless to adopt AL. Since ϵ is not available in the initial stage of learning, it is not appropriate to fix the number of iterations.

We adopt a flexible way to terminate the learning process. We first define the hypothesis space $\mathcal{H}_t$ in iteration t that consists of the models with error rates between the lower bound LB and the upper bound UB:

$$\mathcal{H}_t=\{P_{\theta_i}:\mathrm{LB}\le\epsilon(P_{\theta_i})\le\mathrm{UB}\},\tag{9}$$

where

$$\begin{aligned}\mathrm{LB} &= \min\epsilon(P_{\theta_i}), i\in(1,2,...,K+t),\\ \mathrm{UB} &= \frac{|U|+|L|\epsilon_{\min}(P_{\theta_i})}{|S_{K+1}|},\end{aligned}$$

where $|\cdot|$ denotes the number of instances in the corresponding set.

For the termination condition, we define the volume of the current region of uncertainty (VOU) similar to Balcan *et al.* (2006) as

$$\begin{aligned}&\mathrm{VOU}(\mathcal{H}_t)\\ =&\Pr_{x\in U}[\exists P_{\theta_i},P_{\theta_j}\in\mathcal{H}_t:P_{\theta_i}(y|x)\neq P_{\theta_j}(y|x)].\end{aligned}\tag{10}$$

Therefore, the termination condition can be written as

$$\mathrm{VOU}(\mathcal{H}_t)\le\epsilon_0,\tag{11}$$

where ϵ_0 denotes the allowed error rate.

Our model is able to distinguish from the source domain those models that deviate from the distribution underlying the target domain. If all the models are different from the target task, the weights of these models become zero and the TL problem can thereafter be regarded as a single task learning. Only if the weights of all the models are equal to 1, is the asymmetric problem regarded as the symmetric multi-task learning problem, where all the tasks are treated as equally important. Therefore, the proposed algorithm can adaptively fit to the learning scenario.

6 Experiments

We perform experiments on both the synthetic data sets (Shi *et al.*, 2008) with specific structures to demonstrate the performance of our algorithm, and two real data sets, which are the 20 newsgroups data set (http://people.

csail.mit.edu/jrennie/20Newsgroups) and the four-university data set (http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/). The proposed algorithm is compared with the basic QBC model (Dagan and Engelson, 1995), the state-of-the-art AL method, as well as the random query. We also compare the proposed method with active transfer (Shi *et al.*, 2008), which is a hybrid method combining AL and TL methods. If the baseline method is chosen as SVM, we can then use the C-support vector classification (C-SVC) with the polynomial kernel. We follow the values of the parameters in the original papers (Chang and Lin, 2001). Ten labeled instances are chosen in the beginning for each experiment. In the following figures, we use 'ALTL' to denote the proposed algorithm, 'QBC' for the QBC algorithm, 'RQ' for the random query learning method, and 'AT' as the active transfer algorithm. All the experiments are performed 10 times, and we report the average results.

6.1 Results on synthetic data sets

For the synthetic data sets, we use the same two-dimensional data sets as in Shi *et al.* (2008). Note that, except for data sets A and B (Figs. 2a and 2b), others have different distributions. Data set C (Fig. 2c) shares some similarities with data sets A and B, but for data set D (Fig. 2d), the underlying distribution is dramatically different. Data set E (Fig. 2e) has the reversed distribution from data sets A and B.

The objective of the proposed method is to obtain a high accuracy in classification on the unlabeled data sets, while querying as few instances as possible. We set each of the data sets in Fig. 2 as a single target task, while the others are treated as the source tasks. The results are presented in Fig. 3. Generally, the proposed method is the best among all the methods especially when the number of queries becomes larger than five, and it converges more quickly than the others.

For Figs. 3a and 3b, in most circumstances, the performances of all the methods are better than those on the other data sets. For example, in Fig. 3a, the error rate for the baseline method is about 0.12, while the result of the proposed method becomes stable and converges to 0 after about 20 queries. For RQ and QBC, the convergence is not as good as that of the proposed method. The reason may be that, for data set A as the target task, there exists a similar source task (data set B) with the same underlying distribution and therefore TL could bring greater improvements to the accuracy. The same situation occurs in Fig. 3b, in which the proposed method converges more quickly after about 10 queries.

The distribution for the target task data set C is similar but not identical to those for the underlying data sets A and B. However, the results shown in Fig. 3c illustrate that the performance of the proposed method is still satisfactory due to TL. One advantage of TL is that it can borrow strength from similar tasks to help improve the performance on the target task. Even if the distributions among tasks are not identical, we are still able to obtain lower error rates after 10 queries.

The challenge tasks are situations in which data sets D and E are acting as the target tasks, respectively. Fig. 3d shows that the performances of all the algorithms under the target task data set D fluctuate and the proposed algorithm becomes much more stable after 20 queries. In Fig. 3e, the reversed distribution with data sets A and B helps improve the results and the proposed method obtains an error rate nearly equal to 0 after 15 queries. The possible reason is that, in TL, the negative information is sometimes helpful to improve the performance (Fig. 2e) when the target task has a totally different distribution with others but is spatially similar. However, when the target task is selected as data set D, the useful information that can be transferred from the source domain to the target domain is inadequate. We also notice that, in the initial stage of learning, the proposed algorithm sometimes performs better than the baseline method. We attribute these results to the transferred knowledge from similar tasks which helps improve the performance of the target task in the beginning.

6.2 Results on real data sets

In this section, we provide the experimental results on the two real data sets. The 20 newsgroups data set is a collection of approximately 20 000 newsgroup documents (Table 1). The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related, while others are highly unrelated. We follow the splitting strategy on the 20 newsgroups data set as in Dai *et al.* (2007), to generate different

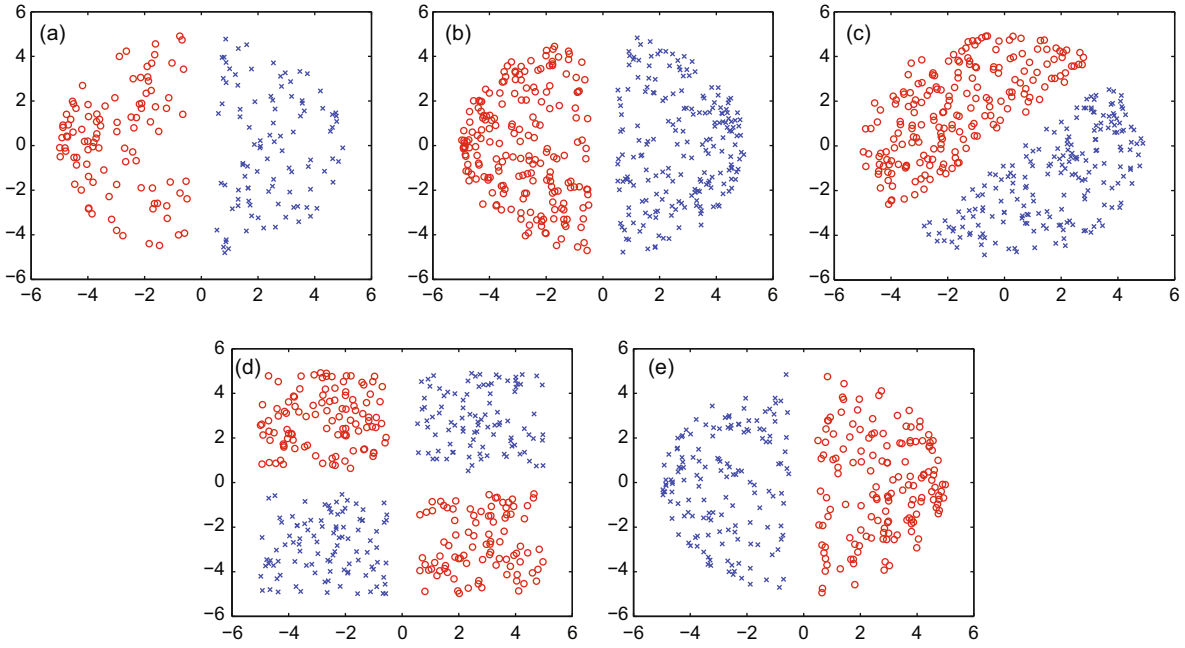

Fig. 2 Structures of the synthetic data sets *A*–*E* (resp. (a)–(e)). The circles and crosses denote the positive and negative instances, respectively

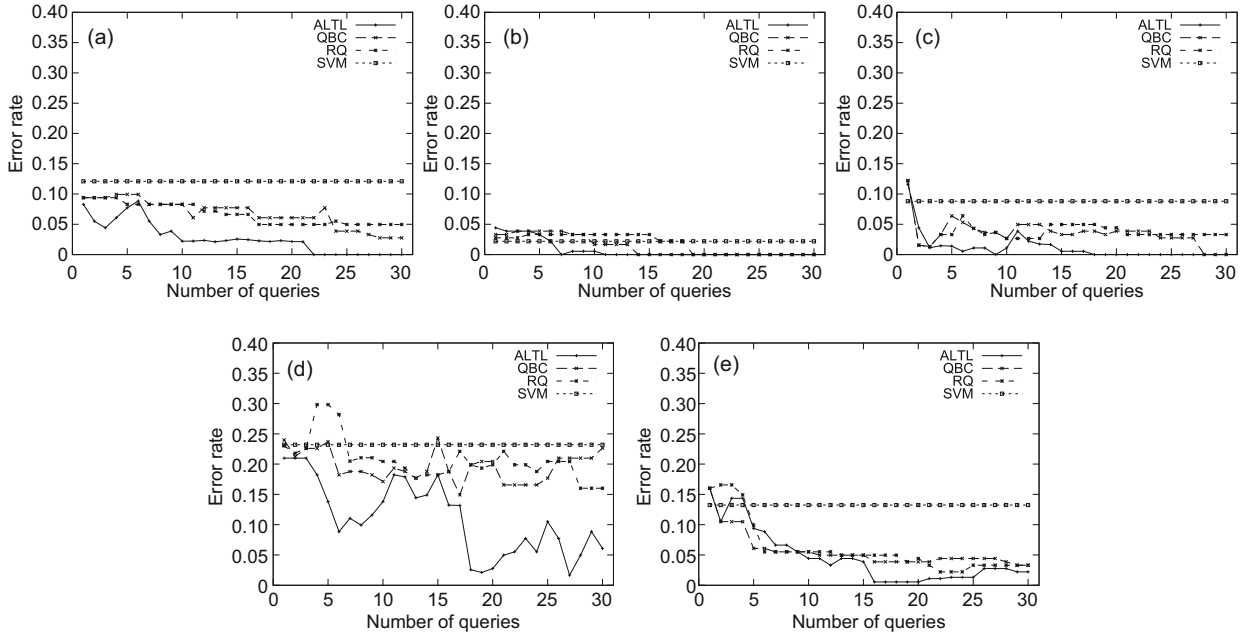

Fig. 3 Error rates on synthetic data sets for different numbers of queries under different target tasks. Data sets *A*–*E* act as the single target task ((a)–(e)), respectively

Table 1 Description of the 20 newsgroups data set

Category	Positive instance	Negative instance	Size
rec vs. sci (1)	rec.autos rec.motorcycles	sci.med sci.space	3961
rec vs. sci (2)	rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics	3965
rec vs. talk (1)	rec.autos rec.motorcycles	talk.politics.guns talk.politics.misc	3669
rec vs. talk (2)	rec.sport.baseball rec.sport.hockey	talk.politics.mideast talk.religion.misc	3561
sci vs. talk (1)	sci.med sci.electronics	talk.religion.misc talk.politics.misc	3374
sci vs. talk (2)	sci.space sci.crypt	talk.politics.mideast talk.politics.guns	3828

tasks for TL. Three categories are chosen, i.e., rec vs. talk, rec vs. sci, and sci vs. talk. For example, in data set rec vs. talk, all the positive instances are from category rec, while the negative ones are from category talk. Different tasks are selected based on the subcategories. In the experiments, each task is chosen as the single target task, and the others are all chosen as the source tasks.

The four-university data set contains web pages collected from computer science departments at various universities. The 8282 pages are manually classified into several categories including student, faculty, course, etc. For each class the data set contains pages from the four universities, which are Cornell, Texas, Washington, and Wisconsin. We perform preprocessing on the data sets and select students and faculty as the positive and negative class labels, respectively. Therefore, we have four data sets for the four universities. In the experiments, each university is treated as the target task, and the other three are regarded as the source tasks.

First, we show the results for the 20 newsgroups data set (Fig. 4). Generally, the proposed method outperforms others in most circumstances especially after 10 queries. For example, in Fig. 4a (rec vs. sci), the performance of the proposed algorithm becomes stable after about 15 queries. However, in the initial stage, the error rate for ALTL fluctuates and sometimes is higher than others. We believe that, in the initial stage, some committee members overfit the target task and therefore they gain higher weights, but after several iterations, the proposed algorithm could decrease their weights when more robust committee members start to dominate the learning process. Therefore, after several queries, the performance becomes stable and the informative instances could be adaptively selected.

For the results of the other categories, the proposed ALTL is able to achieve lower error rates than the other methods. It outperforms the basic QBC algorithm by more than 10% in most circumstances,

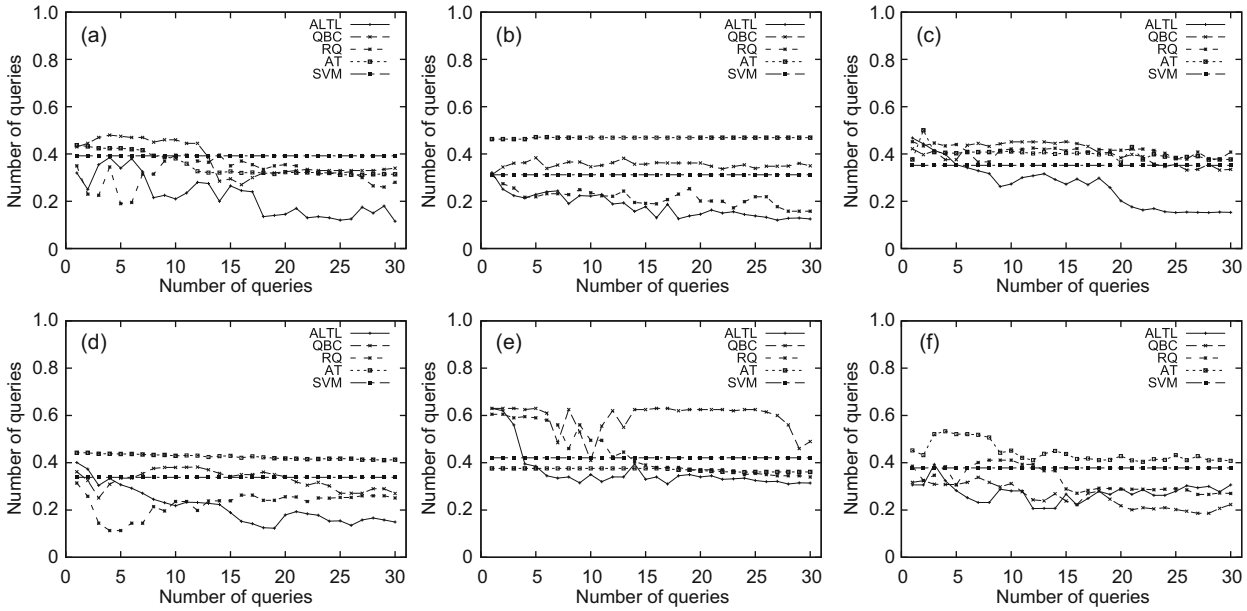

Fig. 4 Error rates for the 20 newsgroups data set for different numbers of queries under different categories: (a) rec vs. sci (1); (b) rec vs. sci (2); (c) rec vs. talk (1); (d) rec vs. talk (2); (e) sci vs. talk (1); (f) sci vs. talk (2)

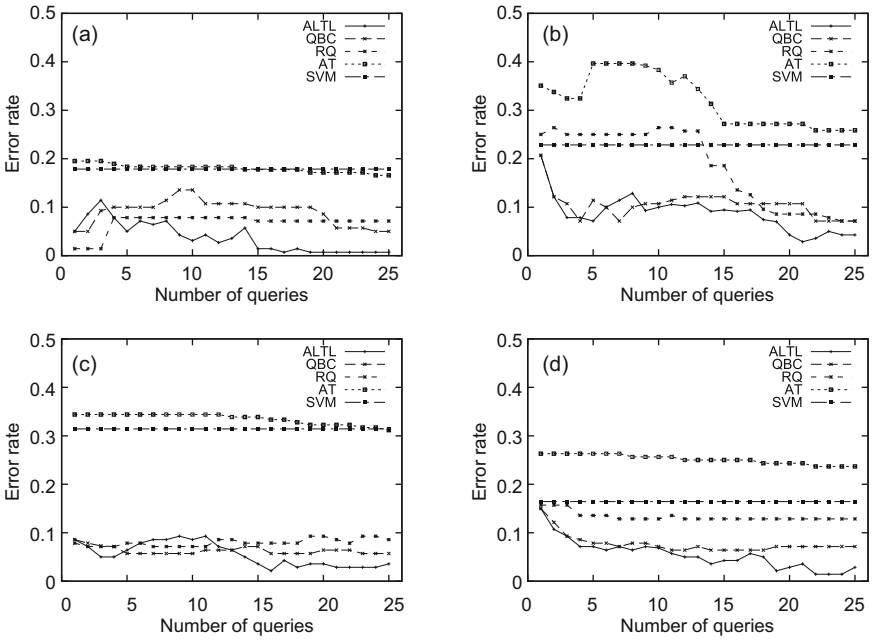

Fig. 5 Error rates for the four-university data set for different numbers of queries: (a) Cornell; (b) Texas; (c) Washington; (d) Wisconsin

and converges more quickly than others. The results could serve as evidence of the effectiveness of the proposed algorithm.

As illustrated in Fig. 5, for the four-university data set, the proposed method is still the best one among all the methods, especially after 10 queries. Note that, the error rates in Fig. 5a (Cornell) are lower than those in the other figures, while in Fig. 5b (Texas), the discrepancy is much larger. Even in these kinds of circumstances, the proposed algorithm could obtain an error rate as low as 5%, which is much lower than the error rates of the other methods. It is obvious that ALTL converges more quickly than the state-of-the-art algorithms. In each iteration, the most informative instance could be selected effectively, which leads to the overall improvement of the classification accuracy.

7 Conclusions

In this study, we propose a novel AL framework by extending the basic QBC algorithm in the target domain, but with the help of useful information from source domains. A weighted KL divergence measurement is adopted to evaluate the similarities between committee members to decrease the negative effects of inferior models. An adaptive strategy is designed to get rid of those unnecessary instances and models, to avoid querying irrelevant instances. By incorporating the advantages of both AL and TL, the proposed method is able to obtain high classification accuracy with fewer queries while converging faster than the state-of-the-art methods. For future research directions, we believe that a detailed mathematical analysis of the lower bound and the upper bound is necessary to develop more robust algorithms for transfer active learning.

References

Argyriou, A., Maurer, A., Pontil, M., 2008. An algorithm for transfer learning in a heterogeneous environment. Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases, p.71-85. [doi:10.1007/978-3-540-87479-9_23]

Balcan, M.F., Beygelzimer, A., Langford, J., 2006. Agnostic active learning. Proc. 23rd Int. Conf. on Machine Learning, p.65-72. [doi:10.1145/1143844.1143853]

Cao, B., Pan, S.J., Zhang, Y., *et al.*, 2010. Adaptive transfer learning. Proc. 24th AAAI Conf. on Artificial Intelligence, p.407-412.

Caruana, R., 1997. Multitask learning. *Mach. Learn.*, **28**(1):41-75. [doi:10.1023/A:1007379606734]

Chang, C.C., Lin, C.J., 2001. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**(3):27. [doi:10.1145/1961189.1961199]

Chattopadhyay, R., Fan, W., Davidson, I., *et al.*, 2013. Joint

transfer and batch-mode active learning. Proc. 30th Int. Conf. on Machine Learning, p.253-261.

Church, K.W., Gale, W.A., 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Comput. Speech Lang.*, **5**(1):19-54. [doi:10.1016/0885-2308(91)90016-J]

Cohn, D., Atlas, L., Ladner, R., 1994. Improving generalization with active learning. *Mach. Learn.*, **15**(2):201-221. [doi:10.1007/BF00993277]

Dagan, I., Engelson, S.P., 1995. Committee-based sampling for training probabilistic classifiers. Proc. 12th Int. Conf. on Machine Learning, p.150-157.

Dai, W., Yang, Q., Xue, G., *et al.*, 2007. Boosting for transfer learning. Proc. 24th Int. Conf. on Machine Learning, p.193-200. [doi:10.1145/1273496.1273521]

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**(1):119-139. [doi:10.1006/jcss.1997.1504]

Harpale, A., Yang, Y., 2010. Active learning for multi-task adaptive filtering. Proc. 27th Int. Conf. on Machine Learning, p.431-438.

Krause, A., Guestrin, C., 2009. Optimal value of information in graphical models. *J. Artif. Intell.*, **35**:557-591.

Lewis, D.D., Gale, W.A., 1994. A sequential algorithm for training text classifiers. Proc. 17th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.3-12.

Li, H., Shi, Y., Chen, M.Y., *et al.*, 2010. Hybrid active learning for cross-domain video concept detection. Proc. Int. Conf. on Multimedia, p.1003-1006. [doi:10.1145/1873951.1874135]

Li, L., Jin, X., Pan, S., *et al.*, 2012. Multi-domain active learning for text classification. Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.1086-1094. [doi:10.1145/2339530.2339701]

Lin, J.H., 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory*, **37**(1):145-151. [doi:10.1109/18.61115]

Luo, C.Y., Ji, Y.S., Dai, X.Y., *et al.*, 2012. Active learning with transfer learning. Proc. ACL Student Research Workshop, p.13-18.

McCallum, A.K., Nigam, K., 1998. Employing EM and pool-based active learning for text classification. Proc. 15th Int. Conf. on Machine Learning, p.350-358.

Muslea, I., Minton, S., Knoblock, C.A., 2002. Active+semi-supervised learning = robust multi-view learning. Proc. 19th Int. Conf. on Machine Learning, p.435-442.

Pereira, F., Tishby, N., Lee, L., 1993. Distributional clustering of English words. Proc. 31st Annual Meeting of Association for Computational Linguistics, p.183-190. [doi:10.3115/981574.981598]

Rajan, S., Ghosh, J., Crawford, M.M., 2006. An active learning approach to knowledge transfer for hyperspectral data analysis. Proc. IEEE Int. Conf. on Geoscience and Remote Sensing Symp., p.541-544. [doi:10.1109/IGARSS.2006.143]

Reichart, R., Tomanek, K., Hahn, U., *et al.*, 2008. Multi-task active learning for linguistic annotations. Proc. Annual Meeting of Association for Computational Linguistics, p.861-869.

Rosenstein, M.T., Marx, Z., Kaelbling, L.P., *et al.*, 2005. To transfer or not to transfer. Proc. NIPS Workshop on Inductive Transfer: 10 Years Later.

Roy, N., McCallum, A., 2001. Toward optimal active learning through sampling estimation of error reduction. Proc. 18th Int. Conf. on Machine Learning, p.441-448.

Settles, B., 2010. Active Learning Literature Survey. Technical Report No. 1648, University of Wisconsin, Madison.

Seung, H.S., Opper, M., Sompolinsky, H., 1992. Query by committee. Proc. 5th Annual Workshop on Computational Learning Theory, p.287-294. [doi:10.1145/130385.130417]

Shao, H., Suzuki, E., 2011. Feature-based inductive transfer learning through minimum encoding. Proc. SIAM Int. Conf. on Data Mining, p.259-270.

Shao, H., Tong, B., Suzuki, E., 2011. Compact coding for hyperplane classifiers in heterogeneous environment. Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases, p.207-222. [doi:10.1007/978-3-642-23808-6_14]

Shi, X.X., Fan, W., Ren, J.T., 2008. Actively transfer domain knowledge. Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases, p.342-357. [doi:10.1007/978-3-540-87481-2_23]

Shi, Y., Lan, Z.Z., Liu, W., *et al.*, 2009. Extending semi-supervised learning methods for inductive transfer learning. Proc. 9th IEEE Int. Conf. on Data Mining, p.483-492. [doi:10.1109/ICDM.2009.75]

Yang, L., Hanneke, S., Carbonell, J., 2013. A theory of transfer learning with applications to active learning. *Mach. Learn.*, **90**(2):161-189. [doi:10.1007/s10994-012-5310-y]

Zhang, Y., 2010. Multi-task active learning with output constraints. Proc. 24th AAAI Conf. on Artificial Intelligence, p.667-672.

Zhu, Z., Zhu, X., Ye, Y., *et al.*, 2011. Transfer active learning. Proc. 20th ACM Int. Conf. on Information and Knowledge Management, p.2169-2172. [doi:10.1145/2063576.2063918]

Zhuang, F., Luo, P., Shen, Z., *et al.*, 2010. Collaborative Dual-PLSA: mining distinction and commonality across multiple domains for text classification. Proc. 19th ACM Int. Conf. on Information and Knowledge Management, p.359-368. [doi:10.1145/1871437.1871486]