On Gleaning Knowledge from Multiple Domains for Active Learning

Zengmao Wang[1], Bo Du[1*], Lefei Zhang[1], Liangpei Zhang[2], Ruimin Hu[1,3], Dacheng Tao[4]
[1] School of Computer, Wuhan University
[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing
[3]National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University
[4] UBTech Sydney AI Institute, The School of Information Technologies, The University of Sydney
wangzengmao@whu.edu.cn, gunspace@163.com, {zhanglefei,zlp62}@whu.edu.cn,
hrm1964@163.com, dacheng.tao@sydney.edu.au

Abstract

How can a doctor diagnose new diseases with little historical knowledge, which are emerging over time? Active learning is a promising way to address the problem by querying the most informative samples. Since the diagnosed cases for new disease are very limited, gleaning knowledge from other domains (classical prescriptions) to prevent the bias of active leaning would be vital for accurate diagnosis.

In this paper, a framework that attempts to glean knowledge from multiple domains for active learning by querying the most uncertain and representative samples from the target domain and calculating the important weights for re-weighting the source data in a single unified formulation is proposed. The weights are optimized by both a supervised classifier and distribution matching between the source domain and target domain with maximum mean discrepancy. Besides, a multiple domains active learning method is designed based on the proposed framework as an example. The proposed method is verified with newsgroups and handwritten digits data recognition tasks, where it outperforms the state-of-the-art methods.

1 Introduction

Given the symptoms of a patient, how can a doctor from historical diagnoses determines whether he/she gets a new disease, which has never or hardly happened before? It is indeed the same challenge in machine learning field, named the limited labeled samples problem. Actually, in many real-world applications, it is usually very expensive to collect training data and manually annotate them by experts, especially in text and visual data classification. [Tan *et al.*, 2015; Long *et al.*, 2014]. Active learning is a promising approaches that can be used to address this challenge by querying the most informative unlabeled samples iteratively for human experts' labeling [Zhang *et al.*, 2014].

Specifically, there are two main inspirations to design practical active learning algorithms [Huang *et al.*, 2014]. The first one is to find the samples that can train a classifier with a low generalization error. These samples can help the classifier find the optimal decision boundary with the uncertainty information. In these methods, it is assumed that the labeled data and the test data are independent and identically distributed (i.i.d.) [Chaudhuri *et al.*, 2015]. However, in the era of big data, the available labeled data may not fit the same underlying distribution with the test data, since there are usually much more unlabeled samples than the labeled samples, and the limited labeled samples cannot represent all the unlabeled data. To overcome this disadvantage, the second kind of active learning approaches which are based on selecting the most representative samples is developed [Chattopadhyay *et al.*, 2012]. Although the second type of active learning methods has a strong generalization ability on unknown data, the efficiency of active learning is decreased, as the uncertainty of the labeled data is not fully used and the representative samples are usually far away from the decision boundary. Since both kinds of active learning methods have flaws, some works have been proposed to query both the uncertain and representative samples [Huang *et al.*, 2014; Wang and Ye, 2015].

The need for domain adaptation: Since there are very limited labeled samples at the beginning of active learning, domain adaptation, which is a technique related to transfer learning, can help compensate for the lack of labeled data in a target domain by exploring a related source domain [Tianyi *et al.*, 2016]. If many labeled samples are available at the beginning of the active learning process, only a small quantity of samples will need to be queried in the whole active learning process. However, if there are very limited labeled samples at the beginning of active learning, the introduction of domain adaptation would be critical to enhance active learning performance. Indeed, many researchers have proved that the use of additional labeled data from a related source domain can increase the reliability of the classifier used in active learning [Chattopadhyay *et al.*, 2013; Kale *et al.*, 2015]. Meanwhile, the availability of the informative labeled data in the target domain can enable efficient transfer of knowledge from the source domain to the target domain [Chattopadhyay *et al.*, 2013; Shi *et al.*, 2008].

Motivating example: Both BBC[1] and CNN[2], two popular

*Corresponding author

[1]http://www.bbc.com/; [2]http://edition.cnn.com/

news websites in the world, are used to illustrate our motivation. A simple example is presented in Figure 1. In Figure 1, there are many items on BBC, such as sports, travel, earth, and so on. However, movie is also a popular item for the young people in the world, evolving over time. If a new section about movie news on BBC is required to satisfy young people's need, we can build a supervised model to recognize the large amount of movie news to achieve this goal. However, the labeled movie news on BBC is very limited, leading to not enough prior knowledge to build a reliable movie classifier. To overcome this problem, we can glean knowledge from the screen items on CNN to movie news on BBC with domain adaptation. With the knowledge from multiple domains, the performance of active learning can be improved with cold start problem. Meanwhile, the labeled information in multiple domains are also more and more suitable to the target tasks in the active learning process.

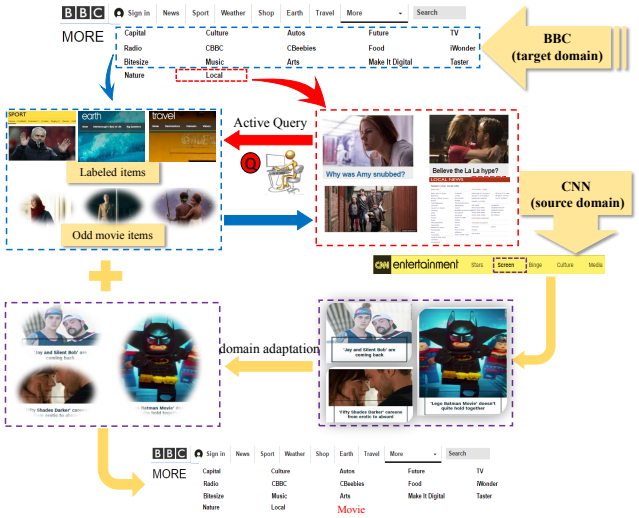

Figure 1: Problem illustration

In this paper, we develop a novel active learning framework by gleaning knowledge from multiple domains to address the problem with the limited labeled samples. In the proposed framework, we re-weight the data in the source domain as additional labeled data to measure the uncertain and representative information in the target domain. Meanwhile, maximum mean discrepancy (MMD) [Ren and Luo, 2016] is adopted to measure the distribution difference between the labeled data (including the labeled data in the target domain and the labeled data in the source domain) and the unlabeled data from the target domain. Moreover, the uncertainty measurement is derived by minimizing the loss in the classifier training procedure with the labeled data in both the target domain and the source domain. To the best of our knowledge, this is the first work that unifies multiple domains for active learning in a formulation by simultaneously considering the uncertain and representative information.

Based on the proposed active learning framework, a practical algorithm is formulated as an example, where a batch of samples which can efficiently improve the quality of the knowledge from multiple domains for the recognition tasks are selected with both high uncertainty and representativeness. Meanwhile, since MMD measures the representative information without label information, the proposed method can handle the cold start problem with no initial labeled target data. We tested the proposed method on 20 tasks in newsgroup and handwritten digit recognition. Significant improvements in classification accuracy were achieved by the proposed method with respect to state-of-the-art methods.

2 Related Works

Active learning and transfer learning are two promising approaches to reduce the labeling cost for classification tasks in the computer vision and data mining fields, and many efforts have been made to develop the two kinds of approaches [Zhang *et al.*, 2014; Guo and Schuurmans, 2007; Fang and Zhang, 2016].

Active learning is an approach querying the most informative samples iteratively and mainly focusing on representative and uncertain information in the unlabeled data [Huang *et al.*, 2014]. [Chakraborty *et al.*, 2015] combined the uncertain and representative into a convex framework to perform active learning loops. [Huang *et al.*, 2014] queried the informative and representative samples based on a min-max framework. [Wang and Ye, 2015] put the discriminativeness and representativeness together via a trade-off parameter to query the i.i.d samples. However, the labeled samples are limited, and we have to query many samples to achieve satisfactory performance for active learning. Transfer learning is another approaches to improve the quantity of labeled samples in target domain by matching distribution between the target domain and a related domain. [Long *et al.*, 2014] aimed to extract common latent factors for knowledge transfer by preserving the statistical property across domains, and simultaneously, refined the latent factors to alleviate negative transfer by preserving the geometric structure in each domain. [Segev *et al.*, 2016] focused on the setting of model transfer whereby the adaptation of a given source model to a target domain relies on a relatively small training set from the target. Although transfer learning improves the quantity of labeled samples, the quality of domain adaptation usually depends on the labeled data in the target domain with supervised techniques [Bruzzone and Marconcini, 2010].

Hence, to make full use of the available data, the combination of active learning and transfer learning has become a hot topic. The method in [Shi *et al.*, 2008] introduced a transfer learning based active learning algorithm. In this method, two classifiers are trained based on the source domain and the target domain, respectively. The confidence of the unlabeled data in the target domain are then predicted with the two classifiers. The samples with the lowest confidence are then labeled by human experts to enhance the classification performances. Using a similar idea, [Li *et al.*, 2013] proposed a new transfer learning based active learning algorithm with a different active learning strategy. [Rai *et al.*, 2010] proposed another method by harnessing the data in the source domain to learn the most possible initializer hypothesis, and using the domain divergence information to perform active learning in the tar-

get domain. [Chattopadhyay *et al.*, 2013] developed a new approach by combining active learning and transfer learning in a single unified formulation to measure the marginal probability distribution difference between the labeled data in both the source domain and target domain, and that between the unlabeled data in the target domain using MMD. [Kale *et al.*, 2015] proposed a hierarchical framework to exploit the cluster structure shared across different domains, which is further utilized for both imputing labels for the unlabeled data and selecting active queries in the target domain. To further reduce the labeling cost, [Hunag and Chen, 2016] queried the samples from the source domain by distribution matching based on MMD.

In the methods stated above, the knowledge of different domains are usually not fully used, such as only with uncertainty, only with representativeness or only with one domain, leading to the use of different knowledge bias. In this paper, to further improve the performance of active learning, an active learning framework with multiple domains knowledge is proposed. It gleans knowledge from multiple domains to query the most uncertain and representative samples, which can make the learner effectively utilize the knowledge of multiple domains, under a unified formulation.

3 Multiple Domains Active Learning

Suppose given a target data set $T = \{(x_1, y_1), ..., (x_n, y_n)\}$ and a source data set $S = \{(x_1, y_1), (x_2, y_2), ..., (x_s, y_s)\}$, initially, we label l samples from the target data set T as the labeled data set $L = \{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\}$, and the remaining $u = n - l$ samples in target data set T as the unlabeled data $U = \{(x_{l+1}, y_{l+1}), (x_{l+2}, y_{l+2}), ..., (x_n, y_n)\}$ in active learning, with $y_i \in \{1, -1\}$, as we only focus on binary problem. It is assumed that $l \leq u$. Since active learning is an iterative procedure to select a subset of b most informative samples to label, we use Q to denote the query samples set.

3.1 The Framework of MDAL

In this paper, we glean the knowledge of the source data by re-weighting each sample in the source domain to adapt the distribution of the target data, with $\{(\beta_i(x_i)x_i, y_i), x_i \in S\}$, where $\beta_i(x_i)$ is the importance weight for x_i in S. Meanwhile, we use both the labeled target data and the re-weighted source data to lower the generalization error of the classifier. In the methods proposed by [Chattopadhyay *et al.*, 2013] and [Hunag and Chen, 2016], the weights are directly optimized by minimizing the difference between the distributions of the target domain and the source domain. MMD is usually adopted as the criterion to estimate the difference of the distributions. The empirical estimate of MMD between the labeled data (including the labeled data in the target domain and source domain) and the unlabeled data in the target domain can be written as:

$$\left\| \frac{1}{l+s} \left[\sum_{x_i \in L} \phi(x_i) + \sum_{x_i \in S} \beta_i \phi(x_i) \right] - \frac{1}{u} \sum_{x_j \in U} \phi(x_j) \right\|_{\mathcal{H}}^2 \quad (1)$$

where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS), and $\phi(x)$ is a function that maps x to RKHS. However, the samples with a high uncertainty are also very important for a supervised training model[Guo and Schuurmans, 2007]. These samples are more efficient at boosting the learning accuracy with certain labeled data. Hence, we use both the labeled target data and the re-weighted source data to build a supervised model by minimizing the loss function as follows:

$$\min_f \sum_{x_i \in L} \ell(y_i, f(x_i)) + \sum_{x_j \in S} \ell(y_j, f(\beta_j x_j)) + \lambda \|f\|_{\mathcal{H}}^2 \quad (2)$$

where $\ell(.,.)$ is the loss function. Since the labeled data in the target domain are limited, it is difficult to train a high-confidence classifier. Active learning provides an effective way to query a set of the most informative samples Q to improve the confidence of the supervised model. With the informative samples in Q, the objective function to train a supervised model can be given by:

$$\begin{aligned} \min_f \sum_{x_i \in L} \ell(y_i, f(x_i)) + \sum_{x_i \in Q} \ell(\widehat{y_i}, f(x_i)) \\ + \sum_{x_j \in S} \ell(y_j, f(\beta_j x_j)) + \lambda \|f\|_{\mathcal{H}}^2 \end{aligned} \quad (3)$$

Since the set of samples is unknown and we mainly focus on the binary problem, $\hat{y_i}$ is the pseudo-label of sample x_i in Q, and it belongs to 1 or -1. To make sure that the query samples in Q can improve the generalization ability of the supervised model, we integrate the representative information into the supervised model with a tradeoff parameter λ_w, which is used to balance the uncertainty and representativeness, and the objective function can be given as follows:

$$\begin{aligned} &\min_{f, \widehat{y_i} \in \{1-1\}} \sum_{x_i \in L} \ell(y_i, f(x_i)) + \sum_{x_i \in Q} \ell(\widehat{y_i}, f(x_i)) \\ &+ \sum_{x_j \in S} \ell(y_j, f(\beta_j x_j)) + \lambda \|f\|_{\mathcal{H}}^2 + \lambda_w \| \frac{1}{l+b+s} \\ &[\sum_{x_i \in L \cup Q} \phi(x_i) + \sum_{x_i \in S} \beta_i \phi(x_i)] - \frac{1}{u-b} \sum_{x_j \in U/Q} \phi(x_j) \|_{\mathcal{H}}^2 \end{aligned} \quad (4)$$

Q is a subset from a large amount of unlabeled data U, and it is unknown in the objective function. Mathematically, we search for it with an indicator vector. A binary indicator vector α is introduced with length u. If x_i is a sample that should belong to Q, α_i is equal to 1; otherwise, α_i is equal to 0. The objective function can then be given by:

$$\begin{aligned} &\min_{\substack{f, \alpha^T 1 = b; \\ \alpha_i \in \{0,1\}, \beta}} \sum_{x_i \in L} \ell(y_i, f(x_i)) + \sum_{x_i \in U} \alpha_i \ell(\widehat{y_i}, f(x_i)) \\ &+ \sum_{x_j \in S} \ell(y_j, f(\beta_j x_j)) + \lambda \|f\|_{\mathcal{H}}^2 + \lambda_w \| \frac{1}{l+b+s} \\ &\left[\sum_{x_i \in L} \phi(x_i) + \sum_{x_i \in U} \alpha_i \phi(x_i) + \sum_{x_i \in S} \beta_i \phi(x_i) \right] \\ &- \frac{1}{u-b} \sum_{x_j \in U} (1 - \alpha_j) \phi(x_j) \|_{\mathcal{H}}^2 \end{aligned} \quad (5)$$

By solving the above objective function, we select the most informative samples according to α, and utilize the source data by domain adaptation with the weights β.

3.2 A Practical Example

With the framework described in Section 3.1, a practical multiple domains active learning (MDAL) method is designed. Defining $f(x_i) = \omega^T \phi(x_i)$ and using the least-squares error as the loss function, we obtain the following formulation:

$$\begin{aligned}&\min_{\omega,\alpha^T 1=b;\alpha_i\in\{0,1\}} \sum_{x_i\in L}\left(y_i-\omega^T\phi\left(x_i\right)\right)^2+\\&\sum_{x_i\in U}\alpha_i\left(\widehat{y_i}-\omega^T\phi\left(x_i\right)\right)^2+\sum_{x_j\in S}\left(y_j-\omega^T\phi\left(\beta_j x_j\right)\right)^2\\&+\lambda\left\|\omega\right\|_{\mathcal{H}}^2+\lambda_w(\alpha^{\mathrm{T}}K_{u,u}\alpha+\beta^T K_{s,s}\beta+\beta^T K_{s,u}\alpha\\&-k_{u,u}^T\alpha-k_{s,u}^T\beta+k_{u,l}^T\alpha+k_{s,l}^T\beta)\end{aligned}\tag{6}$$

To allow the above expression to be clearly understood, we denote K_{UU} as the kernel matrix of the unlabeled data U, K_{SS} as the kernel matrix of the source data S, K_{UL} as the kernel matrix between the unlabeled data U and the labeled data L in the target domain, and K_{SU} and K_{SL} as the kernel matrices between the source data S and the unlabeled target data U and the labeled target data L, respectively. Using $p = (l+s+u)/(u-b)$, $q = (l+s+b)/(u-b)$, the terms can then be represented as

$$\begin{aligned}&K_s=\tfrac{1}{2p^2}K_{SS};K_u=\tfrac{1}{2}K_{UU};K_{s,u}=\tfrac{1}{p}K_{SU}\\&K_{u,u}\left(i\right)=\tfrac{2q}{p^2}\sum_{x_j\in U}K_u\left(i,j\right),K_{s,u}\left(i\right)=\tfrac{q}{p^2}\sum_{x_j\in U}K_{s,u}\left(i,j\right)\\&K_{s,l}\left(i\right)=\tfrac{1}{p^2}\sum_{x_j\in L}K_{SL}\left(i,j\right),K_{u,l}\left(i\right)=\tfrac{1}{p}\sum_{x_j\in L}K_{UL}\left(i,j\right)\end{aligned}$$

In the above formulation, α, β, $\hat{y_i}$, and ω are the main parameters we want to obtain. To simplify the above formulations, we minimize the worst-case scenario introduced by the pseudo-labels of the b query samples in Q. In this case, the loss between $\hat{y_i}$ and $f(x_i)$ is given by $(|\hat{y_i}|+|f(x_i)|)^2$. Since $\hat{y_i}$ is equal to 1 or -1, the worst-case loss can be represented by $(1+|f(x_i)|)^2$. Hence, the objective function becomes:

$$\begin{aligned}&\min_{\substack{\omega,\alpha^T 1=b;\\ \alpha_i\in\{0,1\}}}\sum_{x_i\in L}\left(y_i-\omega^T\phi\left(x_i\right)\right)^2+\sum_{x_i\in U}\alpha_i[\left(\omega^T\phi\left(x_i\right)\right)^2\\&+2\left|\omega^T\phi\left(x_i\right)\right|]+\sum_{x_j\in S}\left(y_j-\omega^T\phi\left(\beta_j x_j\right)\right)^2+\lambda\left\|\omega\right\|_{\mathcal{H}}^2\\&+\lambda_w(\alpha^{\mathrm{T}}K_{u,u}\alpha+\beta^T K_{s,s}\beta+\beta^T K_{s,u}\alpha\\&-k_{u,u}^T\alpha-k_{s,u}^T\beta+k_{u,l}^T\alpha+k_{s,l}^T\beta)\end{aligned}\tag{7}$$

Obviously, the objective function is not convex *w.r.t* α, β, and ω, respectively. However, Eq.(7) can be solved by employing an alternating optimization strategy [Wang and Ye, 2015]. Thus, we iteratively optimize the above objective function by the following two alternating steps:

- Keep α and β fixed, and update ω. The objective function becomes:

$$\begin{aligned}&\min_{\omega}\sum_{x_i\in L}\left(y_i-\omega^T\phi\left(x_i\right)\right)^2+\sum_{x_i\in Q}[\left(\omega^T\phi\left(x_i\right)\right)^2+2\\&\left|\omega^T\phi\left(x_i\right)\right|]+\sum_{x_j\in S}\left(y_j-\omega^T\phi\left(\beta_j x_j\right)\right)^2+\lambda\left\|\omega\right\|_{\mathcal{H}}^2\end{aligned}\tag{8}$$

  (8) can be solved with augmented Lagrange method by setting ω as a kernel form $\omega=\sum_{x_i\in L}\theta_i\phi(x_i)$.

- Keep ω fixed, and define P = [α;β]. The objective function becomes:

$$\begin{aligned}&\min_{P:P_i\in[0,1],P^TV=b}\frac{1}{2}\lambda_w P^T H P+h^T P\\&H=\begin{pmatrix}K_u & K_{u,s}\\ K_{u,s}^T & K_s+F_S\end{pmatrix},\\&F_S\in R^{s\times s}is\text{ a diagonal matrix with}\\&F_S\left(i,i\right)=\left(\omega^T\phi\left(x_i\right)\right)\left(\omega^T\phi\left(x_i\right)\right)^T,x_i\in S\\&h=\begin{pmatrix}\lambda_w(K_{u,l}-K_{u,u})+a\\ \lambda_w(K_{s,l}-K_{s,u})+c\end{pmatrix}\\&a_i=\left(\omega^T\phi\left(x_i\right)\right)^2+2\left|\omega^T\phi\left(x_i\right)\right|,x_i\in U\\&c_i=y_i\left(\omega^T\phi\left(x_i\right)\right),x_i\in S\\&V=\left[\ I;O\ \right],I=1_{u\times 1},O=0_{s\times 1}.\end{aligned}\tag{9}$$

  (9) can be solved as a quadratic program.

4 Experiments

To demonstrate the effectiveness of the proposed MDAL method, we evaluated the proposed method on newsgroups and handwritten digits recognition tasks based on the 20 Newsgroups data set and the USPS and MNIST handwritten digit data sets [Long *et al.*, 2014]. The 20 Newsgroups data set consists of a collection of approximately 20,000 newsgroup documents, partitioned into 20 different categories.

Table 1: The target domain and source domain combinations for the 20 Newsgroups data set

Target domain	Source domain
comp.sys.ibm	comp.sys.mac
comp.sys.mac	comp.sys.ibm
comp.windows.x	comp.os.ms-windows.misc
rec.autos	rec.motorcycles
rec.motorcycles	rec.autos
rec.sport.baseball	rec.sport.hockey
sci.electronics	sci.med
sci.med	sci.electronics
sci.space	sci.crypt
talk.politics.guns	talk.politics.mideast
talk.politics.mideast	talk.politics.guns
talk.religion.misc	talk.politics.misc

Each document is represented by a 200-dimension feature vector, as in [Chattopadhyay *et al.*, 2013]. For the 20 Newsgroups data set, we selected the top four categories-comp, rec, sci, and talk-each containing four sub-categories. Hence, 16 classes were used in the experiments. Since the proposed method is based on binary classes, we generated several binary tasks to distinguish each class from a set of negative classes. We selected the two most similar sub-categories as the target domain and source domain alternately. The target domains and source domains of the 20 Newsgroups data set are listed in Table 1. In each task, the positive class of the source and target domain data consisted of 200 documents randomly sampled from the corresponding categories, as in [Chattopadhyay *et al.*, 2013]. The negative classes consisted of a random mixture of 400 samples from the other categories. The USPS and MNIST handwritten digit data sets

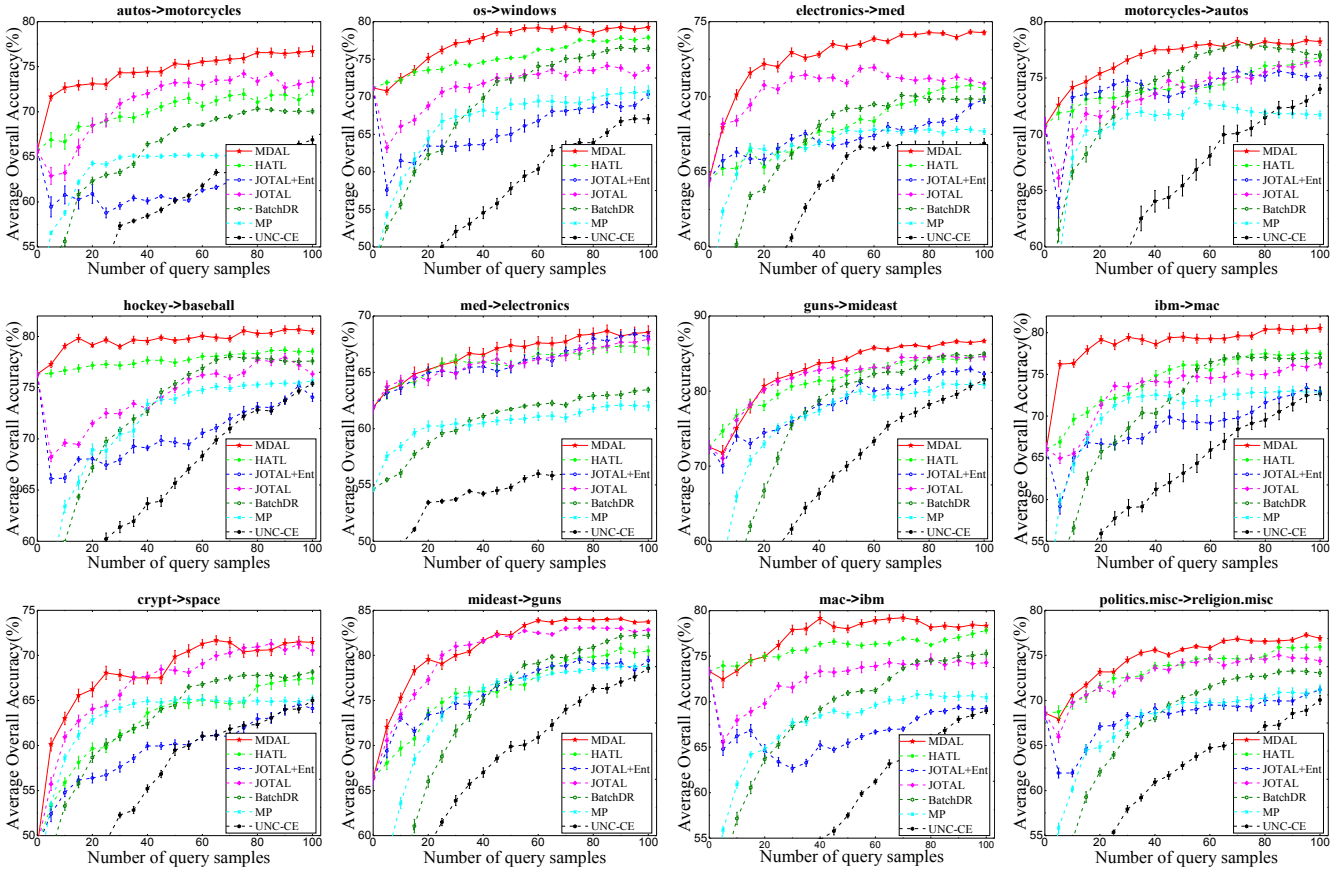

Figure 2: Comparison between the different methods on the 20 Newsgroups data set with different target domains and source domains. The curves show the average overall accuracy over the queries and standard deviation results. Each curve represents the average result of 10 runs.

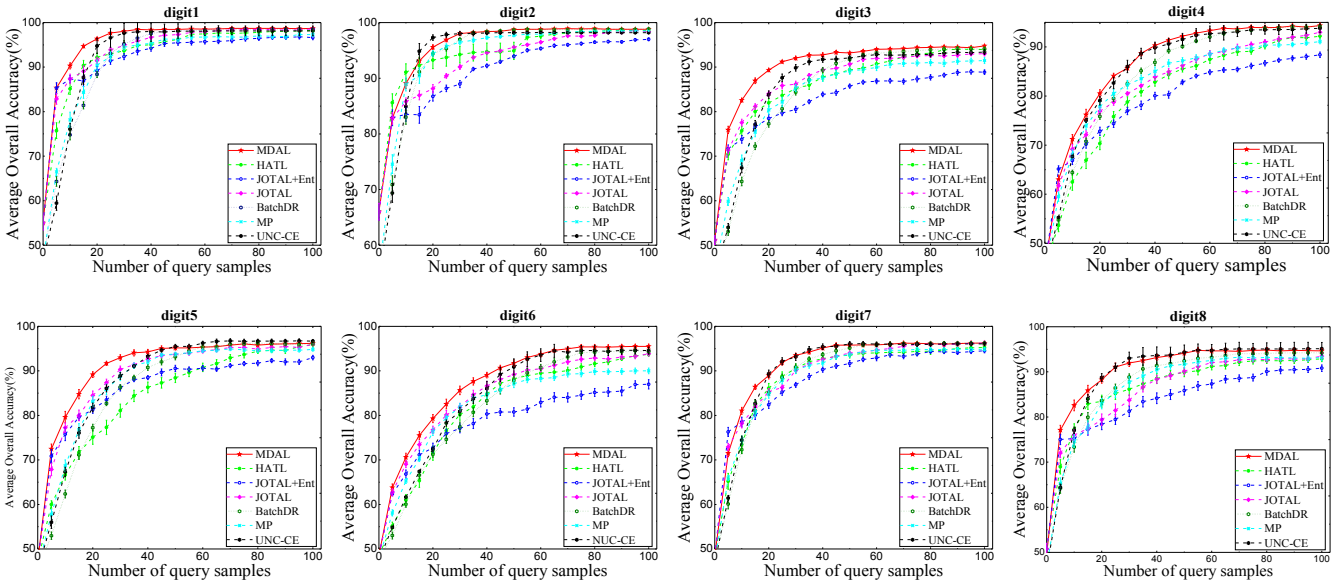

Figure 3: Comparison between the different methods with the USPS and MNIST handwritten digit data sets, with USPS as the source domain and MNIST as the target domain. The curves show the average overall accuracy over the queries and standard deviation results. Each curve represents the average result of 10 runs.

[Long *et al.*, 2014] represent the various fonts of each digit from 1 to 10 using 256-dimension features normalized to the range [0, 1]. In our experiments, we treated digits 1 to 8 in USPS as the source domain, and those in MNIST as the target domain. For the task of digit recognition, we provide an example for illustration. During the task of recognizing digit 1, we treated digit 1 in USPS as the source domain and digit 1 in MNIST as the target domain. The negative samples in this task were 400 samples randomly selected from the other digits in MNIST. The tasks for other digits were constructed in the same way as the digit 1 task. Meanwhile, several state-of-the-art and baseline methods were compared in the experiments: **HATL** [Kale *et al.*, 2015]; **JOTAL+Ent** [Chattopadhyay *et al.*, 2013]; **JOTAL** [Chattopadhyay *et al.*, 2013]; **BatchDR** [Wang and Ye, 2015] **MP** [Chattopadhyay *et al.*, 2012]; **UNC-CE** [Sharma and Bilgic, 2017] based on support vector machine.

In the experiments, we randomly divided each target domain as follows: For the positive samples in each task, 50% for testing, one sample as the initial labeled data, and the other near 50% as the unlabeled data for the active learning. For the negative samples in each task, we also randomly divided them into three parts: 20% for testing, 60% as the initial labeled data, and the other 20% as the unlabeled data for the active learning. For the classifier, without loss of generality, support vector machine (SVM) with a Gaussian kernel was adopted with the LibSVM tool [Chang and Lin, 2011]. There are two important parameters in the SVM classifier: the kernel width parameter g and the penalty parameter C. For convenience, we set the two parameters with empirical values of C = 100 and g = 0.05. For a fair comparison, we adopted the same kernel parameter in all the methods. For the methods with a tradeoff parameter, we fixed it as 10, as in [Hunag and Chen, 2016]. At each iteration, five samples were selected for labeling, and we stopped the iteration loop when 20 iterations were reached. The experiments were repeated 10 times for each task with different sets of unlabeled data and test data. The average results and the standard deviation results are reported. To make the results in figures are clear, we use a large error bar scale from 0 to 200 with the right line in each result to represent the standard deviation results.

4.1 Comparative Study: 20 Newsgroups Data Set

The performances of the proposed MDAL method and the other methods on the 20 Newsgroups data set with different sub-categories are shown in Figure 2. From these results, the proposed method outperforms the other methods in the following aspects. Firstly, the proposed method achieves the best performances at most cases. No compared method performs better than all the other compared methods at all the time. Secondly, the uncertainty and representativeness of multiple domains in a unified formulation can boost the performance of active learning. Among the compared methods, JOTAL+Ent is an active transfer learning method that measures the uncertainty and representativeness with an ad hoc design, while JOTAL is an active transfer learning method that just measures the representativeness. However, JOTAL + Ent always performs worse than JOTAL. This may be because, with only a few labeled samples in the target domain, the measurement of the uncertainty is biased, since the negative samples in the target domain are much more than the positive samples, leading to the distribution measurement bias. Meanwhile, MDAL performs much better than BatchDR, M-P, and UNC-CE, which are active learning methods. This strongly demonstrates that gleaning knowledge from multiple domains is greatly beneficial to active learning.

4.2 Comparative Study: Handwritten Digit Data

Figure 3 shows the results obtained with the handwritten digit data sets. Nine tasks were constructed based on digits 1 to 8. From all the results obtained with the handwritten digit data, we can observe that, in most of the tasks, MDAL performs better than the other compared methods. While UNC-CE also performs well. The reason may be that the distributions of different digits may be very different. Hence, the uncertain information for each task are more important. It just need to query the samples around their boundaries to build relatively reliable boundaries. Meanwhile, BatchDR also performs much better than the rest four compared methods. Since the representative information may be not important on digits data, UNC-CE performs better than BatchDR. This demonstrates that the proposed method can keep the quality of the labeled target data when the information from the domain is of bad quality.

5 Conclusion

In this paper, a framework that integrates the uncertainty and representativeness of multiple domains in a single unified formulation for active learning has been proposed. With the source domain, the proposed method is a superior framework of active learning by gleaning multiple domains. Without the source domain, it becomes a traditional active learning framework. To the best of our knowledge, this is the first work to integrate the uncertainty and representativeness of multiple domains for active learning. Based on the proposed framework, a practical MDAL algorithm is proposed, which can query the most uncertain and representative samples from the target domain to make the knowledge of multiple domains be suitable for the recognition of the target data. The proposed method was verified with the 20 Newsgroups data set and two handwritten digit data sets. The results confirm the superior performance of the proposed method, which is clear evidence that the gleaning knowledge from multiple domains is very promising for active learning.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61471274, 41431175, 61231015, 61671336, the Natural Science Foundation of Hubei Province under Grants 2014CFB193, the Fundamental Research Funds for the Central Universities under Grants 2042016kf0152, National High Technology Research and Development Program of China (863 Program) No. 2015AA016306, EU FP7 QUICK project under Grant Agreement No. PIRSES-GA-2013-612652* and Australian Research Council Projects FT-130101457, DP-140102164, LP-150100671.

References

[Bruzzone and Marconcini, 2010] L. Bruzzone and M. Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010.

[Chakraborty *et al.*, 2015] S Chakraborty, V Nallure Balasubramanian, Q Sun, S Panchanathan, and J Ye. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):1–1, 2015.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[Chattopadhyay *et al.*, 2012] Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Batch mode active sampling based on marginal probability distribution matching. *Acm Transactions on Knowledge Discovery from Data*, 7(3):13, 2012.

[Chattopadhyay *et al.*, 2013] R. Chattopadhyay, W. Fan, I. Davidson, S. Panchanathan, and J. Ye. Joint transfer and batch-mode active learning. In *International Conference on Machine Learning*, pages 253–261, 2013.

[Chaudhuri *et al.*, 2015] Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Neural Information Processing Systems*, pages 1090–1098, 2015.

[Fang and Zhang, 2016] Yin J. Zhu X. Fang, M. and C. Zhang. Trgraph: Cross-network transfer learning via common signature subgraphs. In *International Conference on Data Engineering*, pages 1534–1535, 2016.

[Guo and Schuurmans, 2007] Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. In *Neural Information Processing Systems*, pages 593–600, 2007.

[Huang *et al.*, 2014] S. J. Huang, R. Jin, and Z. H. Zhou. Active learning by querying informative and representative examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1936–1949, 2014.

[Hunag and Chen, 2016] Shengjun Hunag and Song Chen. Transfer learning with active queries from source domain. In *International Joint Conference on Artificial Intelligence*, 2016.

[Kale *et al.*, 2015] David Kale, Marjan Ghazvininejad, Anil Ramakrishna, Jingrui He, and Yan Liu. Hierarchical active transfer learning. In *SIAM International Conference on Data Mining*, 2015.

[Li *et al.*, 2013] Shoushan Li, Yunxia Xue, Zhongqing Wang, and Guodong Zhou. Active learning for cross-domain sentiment classification. In *International Joint Conference on Artificial Intelligence*, pages 2127–2133, 2013.

[Long *et al.*, 2014] Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. Transfer learning with graph co-regularization. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1805–1818, 2014.

[Rai *et al.*, 2010] Piyush Rai, Avishek Saha, Iii Hal Daum, and Suresh Venkatasubramanian. Domain adaptation meets active learning. *In Workshop on Active Learning for Natural Language Processing*, pages 27–32, 2010.

[Ren and Luo, 2016] Zhu J. Li J. Ren, Y. and Y. Luo. Conditional generative moment-matching networks. In *Advances in Neural Information Processing Systems*, pages 2928–2936, 2016.

[Segev *et al.*, 2016] Noam Segev, Maayan Harel, Shie Mannor, Koby Crammer, and El Yaniv Ran. Learn on source, refine on target: A model transfer learning framework with random forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2016.

[Sharma and Bilgic, 2017] Manali Sharma and Mustafa Bilgic. Evidence-based uncertainty sampling for active learning. *Data Mining and Knowledge Discovery*, 31(1):164–202, 2017.

[Shi *et al.*, 2008] Xiaoxiao Shi, Wei Fan, and Jiangtao Ren. Actively transfer domain knowledge. In *Machine Learning and Knowledge Discovery in Databases, European Conference*, pages 342–357, 2008.

[Tan *et al.*, 2015] Ben Tan, Yangqiu Song, Erheng Zhong, and Qiang Yang. Transitive transfer learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1155–1164, 2015.

[Tianyi *et al.*, 2016] Joey Tianyi, Sinno Jialin Pan, Ivor W Tsang, and Shen Shyang Ho. Transfer learning for cross-language text categorization through active correspondences construction. In *AAAI Conference on Artificial Intelligence*, 2016.

[Wang and Ye, 2015] Zheng Wang and Jieping Ye. Querying discriminative and representative samples for batch mode active learning. *ACM Transactions on Knowledge Discovery from Data*, 9(3):1–23, 2015.

[Zhang *et al.*, 2014] Luming Zhang, Yue Gao, Rongrong Ji, Yingjie Xia, Qionghai Dai, and Xuelong Li. Actively learning human gaze shifting paths for semantics-aware photo cropping. *IEEE Transactions on Image Processing*, 23(5):2235–2245, 2014.