Knowledge-Based Systems 141 (2018) 178–187

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A Selective Multiple Instance Transfer Learning Method for Text Categorization Problems☆

Bo Liu[a], Yanshan Xiao[b,*], Zhifeng Hao[c]

[a] School of Automation, Guangdong University of Technology, Guangzhou, China
[b] School of Computers, Guangdong University of Technology, Guangzhou, China
[c] School of Mathematics and Big Data, Foshan University, Foshan, China

ARTICLE INFO

Article history:
Received 2 December 2016
Revised 10 November 2017
Accepted 13 November 2017
Available online 14 November 2017

Keywords:
Data Mining
Transfer Learning

ABSTRACT

Multiple instance learning (MIL) is a generalization of supervised learning which attempts to learn a distinctive classifier from bags of instances. This paper addresses the problem of the transfer learning-based multiple instance method for text categorization problem. To provide a safe transfer of knowledge from a source task to a target task, this paper proposes a new approach, called selective multiple instance transfer learning (SMITL), which selects the case that the multiple instance transfer learning will work in step one, and then builds a multiple instance transfer learning classifier in step two. Specifically, in the first step, we measure whether the source task and the target task are related or not by investigating the similarity of the positive features of both tasks. In the second step, we construct a transfer learning-based multiple instance method to transfer knowledge from a source task to a target task if both tasks are found to be related in the first step. Our proposed approach explicitly addresses the problem of safe transfer of knowledge for multiple instance learning on the text classification problem. Extensive experiments have shown that SMITL can determine whether the two tasks are related for most data sets, and outperforms classic multiple instance learning methods.

1. Introduction

Multiple instance learning (MIL) [1,2] is a new paradigm in machine learning that addresses the classification of bags. In MIL, the labels in the training set are associated with sets of instances, which are called bags. A bag is labelled positive if it contains at least one positive instance; otherwise, the bag is labelled negative. The task of MIL is to learn a multiple instance classifier to classify unknown bags as either positive or negative. To date, MIL has been successfully used in the text categorization domain [3–7] and delivers superior performance. In this case, a document is considered as a bag and each part of the document is an instance. A document is classified as positive if it contains at least one part which is related to the subject of interest.

Depending on the nature of the principle models, previous approaches to MIL can be broadly classified into two categories: (1) bag-level approaches [4,8–10], in which each bag is considered as a whole and operations are directly conducted on the bags to find their labels in the training phase; (2) instance-level approaches [3,11–13], which first attempt to infer the hidden instance label, and calculate the labels of the bags from the labels of their instances.

Despite much progress made on multiple instance learning, most of the previous work considers the MIL problem as a single learning task in the training. However, in many real-world applications, we expect to reduce the labeling effort of a new task (referred to as target task) by transferring knowledge from one or more related tasks (source tasks), which is called transfer learning [14–17]. For example, we may have plenty of user's previous labeled webpages, which indicate the user's interest; as time goes on, user's interest may gradually drift; however, we may not have too much user's current labeled webpages, since labeling plenty of webpages timely may be impossible for the user. In this case, we expect the user's previously labeled webpages transfer knowledge to help build a multiple instance classifier for prediction. Therefore, it is necessary to explore a transfer learning-based multiple instance method for the text categorization problem. To address this, we have the following two challenges.

- How to judge the similarity of the source and target tasks. In transfer learning, when two tasks are unrelated, the knowledge extracted from a source task may not help, and may even hurt,

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.

E-mail addresses: csboliu@163.com (B. Liu), xiaoyanshan@189.cn (Y. Xiao), zfhao@fosu.edu.cn (Z. Hao).

https://doi.org/10.1016/j.knosys.2017.11.019
0950-7051/© 2017 Elsevier B.V. All rights reserved.

the performance of a target task, which can be referred to as negative transfer [18,19]. To avoid negative transfer in multiple instance learning for text categorization, user's interest of subject in the source task should be similar to that in the target task.
- How to build a transfer learning-based classifier for a multiple instance problem. In the process, we want to use the previous task to help learn a more accurate multiple instance classifier for the target task. This classifier based on the target task is then used for prediction.

In this paper, we address the problem of transfer learning for multiple instance learning on text categorization. In order to provide a safe transfer from a source task to a target task, this paper proposes a new approach, termed as selective multiple instance transfer learning (SMITL), which first evaluates the similarity of the source and target tasks and then builds a transfer multiple instance learning classifier for the target task in two steps. In all, the main contributions of the paper are summarized as follows.

- In the first step, we extract positive features from positive bags for both tasks to evaluate the similarity of their positive features such that we know whether the user's interest in the two tasks is similar. Specially, we put forward the similarity evaluation method to measure the similarity of the positive features from both tasks, which can investigate weather both tasks are related or not.
- In the second step, if the two tasks are found to be similar in the first step, we then propose a new multiple instance transfer learning classifier to transfer knowledge from a source task to a target task by extending our previous work for single multiple instance learning in [13]. We then present an alternative framework to deliver a multiple instance transfer learning classifier.
- Extensive experiment has conducted to investigate the performance of our proposed SMITL method. The results have shown that SMITL performs better than classic multiple instance learning methods.

The rest of the paper is organized as follows. Section 2 discusses previous work. Section 3 introduces the preliminary of our method, Section 4 proposes our selective multiple instance learning method for textual classification. Experiments are conducted in Section 5. The conclusion is presented in Section 6.

2. Related Work

2.1. Transfer Learning

Transfer learning [16,17,20–22] has been recognized as an important topic in machine learning and data mining. In contrast to multi-task learning [23–25], transfer learning focuses on transferring knowledge from the source task to the target task, rather than ensuring the performance of each task.

Some of the previous transfer learning methods are usually based on certain assumptions. For example, the work in [25–28] assume that source and target tasks should share some parameters in the learning model. By discovering the shared parameters, the knowledge can be transferred from the source task to the target task. However, these work always assume the distribution of the data to be specified as a priori, which makes them inapplicable to many real-world applications. Other algorithms such as [29,30] assume that some instances or features can be used as a bridge for knowledge transfer.

In addition, Melih [31] introduces a Gaussian process based Bayesian model for asymmetric transfer learning by adopting a two-layer feed-forward deep Gaussian process. The work in [32] exploits four kinds of concepts including the identical concepts, the synonymous concepts, the different concepts and the ambiguous concepts simultaneously, for cross-domain classification. Xiao [33] includes the transfer learning to handle the one-class classification problem with the uncertain data. Furthermore, researchers propose the approaches to learn dictionaries for robust action recognition across views by learning a set of view-specific dictionaries [34]. Zhao [35] introduces the transfer learning from different data distributions of the multi domains. Thereafter, multi-bridge transfer learning was proposed to learn the distributions in the different latent spaces together [36].

Most of the previous work has not explicitly addressed the problem of multiple instance transfer learning for text categorization with safe knowledge transfer from the source task to the target task. This paper put forward the selective multiple instance transfer learning method, which first evaluates the similarity of the positive features between the source and target tasks to judge whether both tasks are related, and then builds our proposed multiple instance transfer learning classifier for text categorization.

2.2. Multiple Instance Learning

Since too many work has been done on MIL, we briefly review some of the relevant work as follows.

Previous work on MIL can be broadly classified into two categories: bag-level and instance-level approaches. For bag-level approaches, each bag is considered as a whole in the training. In this category, representative algorithms including Diverse Density (DD) method [8], EM-DD [9], DD-SVM [10] and the MILES [4] method. For example, DD-SVM [10] selects a set of prototypes using DD function, and then an SVM was trained based on the bag features summarized by these selected prototypes.

For instance-level approaches [3,11–13,37], they attempt to infer the hidden instance label, and calculate the labels of the bags using the label information of the instances. For example, mi-SVM [3] alternatively builds an SVM-based classifier and identifies the positive instances from positive bags such that the final classifier can accurately classify the unknown bags. SMILE [13] method introduces the similarity between each instance and positive bag into the learning and outperforms other multiple instance learning method.

Some work focus on selecting a subset of instance from positive bags to learn the classifier. For example, EM-DD [38] chooses one instance that is most consisitent with the current hypothesis in each positive bag to predict an unknown bag. MI-SVM [39] adopts an iterative framework to learn an SVM classifier. Wu et al. [40] learn a deep multiple instance learning classifier for image classification and auto-annotation problem. Furthermore, Carbonneau et al. [41] use random subspace instance selection into ensemble of multiple instance learning classifier. The work in [42] introduces the MITL (multiple instance transfer learning), which aims at multi-task problems for multiple instance learning problem, however, this method does not determine the similarity of the tasks. Wang et al. [43] design the knowledge transfer in multiple instance learning for the image data classification. Furthermore, the multi-label multi-instance transfer learning [44] was proposed for multiple human signaling pathways data in bio-informatics domain. sub-space-based approach [45] was proposed for multi-instance data.

In addition, Wu et al. [46] propose an efficient and novel Markov chain-based multi-instance multi-label (Markov-Miml) learning algorithm to evaluate the importance of a set of labels associated with objects of multiple instances. Wu et al. [47] propose a co-transfer learning framework that can perform learning simultaneously by co-transferring knowledge across different feature spaces. Xu et al. [48] propose exploit the intrinsic geometry of the multi-instance data by using the Mahalanobis distance and utilize the bag importance weights to transfer knowledge from a source domain to target domain.

The difference between the transfer learning setting for single-instance learning and the transfer learning setting for multi-instance learning is that: in the former one, the sample units of the source and target tasks are represented by single instances; while in the transfer learning for multi-instance learning, the sample unit, called a bag, is composed of a set of single instances. Despite much progress made on the MIL problem, most of the previous work has not explicitly dealt with multiple instance transfer learning for text categorization. This paper proposes the SMITL approach which first determines the similarity between the source and target tasks and then builds our proposed transfer learning classifier to transfer knowledge from the source task to the target task.

3. Preliminary

In multi-instance learning, a bag contains a set of instances, the label of a bag is associated with the labels of instances in this bag. A bag is labelled as positive if it contains at least one positive instance, otherwise, the bag is labelled negative. For convenience, we utilize capital letter B_I and Y_I to denote the *Ith* bag and its label. For the instances in a bag, we utilize lower letter x_i and y_i to denote the *ith* instance and the instance label.

Let (B^s_I, Y^s_I), $I = 1, 2, \dots |Y^s|$ and (B^t_J, Y^t_J), $J = 1, 2, \dots, |Y^t|$ denote the training sets for the source task *S* and target task *T*. Here, $|Y^s|$ and $|Y^t|$ are the number of bags for the source task *S* and target task *T* respectively. For (B^s_I, Y^s_I), $I = 1, 2, \dots |Y^s|$, B_I denotes a bag which contains a number of instance, we utilize $\mathbf{x}_i$ (here $\mathbf{x}_i \in B_I$) to denote an instance of B_I; Y_I denotes the bag label: if $Y_I = 1$, then at least one positive instance exists in B_I, if $Y_I = -1$, then each instance in B_I is negative. For (B^t_J, Y^t_J), $J = 1, 2, \dots, |Y^t|$, we have the same explanation.

In MIL, as we do not know the label of the instances in the positive bags, we represent an instance $\mathbf{x}$ from a positive bag using the following data model:

$$\{\mathbf{x}, m^+(\mathbf{x}), m^-(\mathbf{x})\}, \tag{1}$$

where $m^+(\mathbf{x})$ and $m^-(\mathbf{x})$ represent the similarity of $\mathbf{x}$ towards the positive and negative classes, respectively. We have $0 \le m^+(\mathbf{x}) \le 1$ and $0 \le m^-(\mathbf{x}) \le 1$. $\{\mathbf{x}, 1, 0\}$ means that $\mathbf{x}$ belongs to the positive class, while $\{\mathbf{x}, 0, 1\}$ indicates that $\mathbf{x}$ is negative. For $\{\mathbf{x}, m^+(\mathbf{x}), m^-(\mathbf{x})\}$, where $0 < m^+(\mathbf{x}) < 1$ and $0 < m^-(\mathbf{x}) < 1$, it implies that the similarity of $\mathbf{x}$ towards the positive and negative classes are both considered.

Using the above data model, we can convert a multiple instance transfer learning problem into a single instance learning problem. This makes it possible for the supervised learning methods adapted to solve the transfer MIL problem.

4. Proposed Approach

In this section, we will introduce our proposed method in detail. Suppose that we have an MIL problem as a source task with a large amount of training data and another MIL problem as a target task with a small amount of training data.

The main task of multiple instance transfer learning is to transfer knowledge from a source task to a target task. However, the two tasks may be not related in reality, such that the transfer may be unsuccessful or may even hurt the target task [19]. To avoid this, this paper proposes a selective multiple instance transfer learning for text categorization. Our proposed method works in two steps:

1. In the first step, we evaluate the similarity of the source and target tasks by investigating the similarity of their positive features from positive bags.
2. In the second step, we construct a selective multiple instance transfer learning classifier for the target task if the source and target tasks are found related in the first step.

In the following, we exhibit the two steps in detail.

4.1. Step 1: Similarity Evaluation of the Source and Target Tasks

In MIL for text classification, the positive bag instances contain user's interests, and the instances in the negative bags indicate user's unfavorite subjects. Since each instance is text-based data, in the process of converting an instance into features for learning, the user's interested subject contents are converted into some features of the instances in the positive bags which are called positive features; therefore, we can evaluate the similarity of both tasks by investigating the positive features of two tasks. If their positive features are similar, it is believed that user's interests won't drift much; otherwise, user's interest may change a lot. So, we extract positive features from the positive bags for both tasks and evaluate their similarity as follows.

4.1.1. Positive Feature Extraction

We first extract positive features from the positive bags for the source and target tasks respectively.

First of all, for the source task, we put the instances from the positive bags into $S^+_s = \{\mathbf{x}_1, \dots, \mathbf{x}_{|S^+_s|}\}$ and store the instances of the negative bags into $S^-_s = \{\mathbf{x}_1, \dots, \mathbf{x}_{|S^-_s|}\}$. Similarly, for the target task, we have S^+_t and S^-_t to store positive bag instances and negative bag instances respectively. For similarity, we introduce the operations on the source task as follows.

For the instances in S^+_s, it is believed that they possess some features contained in the core vocabulary of S^+_s [49][1]. The set of positive features for S^+_s is denoted by VP_s. Following the operation in [49], we extract the positive features which appear frequently in positive bags (S^+_s) while occur less frequently in negative bags (S^-_s) as follows.

First of all, feature strength is utilized to indicate the important of a feature toward the classification. In general, a feature with a higher feature strength indicates that it is more effective in terms of classification. Techniques for feature strength can be found from probabilistic theory [50]. Unfortunately, we may do not have any prior knowledge about the feature distribution between the positive examples and negative examples over the entire document domain. For S^+_s, we calculate the *feature strength* of a particular feature f_k by measuring the differences of the normalized examples frequency between S^+_s and S^-_s

$$H_s(f_k) = \frac{n_{S^+_s}(f_k) - min_{S^+_s}}{max_{S^+_s} - min_{S^+_s}} - \frac{n_{S^-_s}(f_k) - min_{S^-_s}}{max_{S^-_s} - min_{S^-_s}}, \tag{2}$$

Above, $n_{S^+_s}(f_k)$ and $n_{S^-_s}(f_k)$ represent the number of examples that contain f_k in S^+_s and S^-_s respectively, $max_{S^+_s}$ and $min_{S^+_s}$ are maximum and minimum values of $n_{S^+_s}(f_k)$ for $f_k \in V_s$ where V_s is the feature set of the source task. $max_{S^-_s}$ and $min_{S^-_s}$ are maximum and minimum values of $n_{S^-_s}(f_k)$ for $f_k \in V_s$.

In the multi-instance textual learning, the positive features should appear more frequently in the positive bags while appear less in the negative bags; therefore, positive features should have

[1] The work in [49] addresses the problem of positive and unlabeled learning, the authors found that positive documents always contain positive features which appear more frequently in positive documents than in unlabeled or negative documents.

a relatively large value of $H_s(f_k)$ in (2). Thus, the positive features are identified by extracting f_k such that $H_s(f_k) > \theta_s$, where

$$\theta_s = \tfrac{1}{|V_s|} \textstyle\sum_{f_k \in V_s} H_s(f_k) \quad (3)$$

is set. After that, we extract positive features for the source task and put them into VP_s. We then let $G_s = \frac{1}{|VP_s|} \sum_{f_k \in VP_s} |H_s(f_k)|$.

Similarly, for the target task, we calculate feature strength $H_t(f_k)$, positive feature set VP_t and G_t.

4.1.2. Similarity Evaluation of the Positive Features

We compare the similarity of the *positive features* between two tasks to identify whether they are related or not. Specifically, for the sets VP_s and VP_t, the *feature strength* of feature f_k ($f_k \in VP_s \cup VP_t$) between S_s^+ and S_t^+, denoted as $H_c(f_k)$, is calculated in the same way as Equation (2).

$$H_c(f_k) = \frac{n_{S_s^+}(f_k) - min_{S_s^+}}{max_{S_s^+} - min_{S_s^+}} - \frac{n_{S_t^+}(f_k) - min_{S_t^+}}{max_{S_t^+} - min_{S_t^+}}. \quad (4)$$

Let

$$G_c = \frac{1}{|VP_s \cup VP_t|} \sum_{f_k \in VP_s \bigcup VP_t} |H_c(f_k)|. \quad (5)$$

We then calculate

$$\eta = \frac{G_c}{min(G_s, G_t)}. \quad (6)$$

In general, the smaller the η, the more likely both tasks are related. If η is larger than 1, we believe both tasks are not related.

In this step, we utilize the method in [49] to extract positive features from the positive bags, and then put forward the similarity evaluation method to investigate whether both tasks are related or not. In our method, we utilize η to measure the similarity of both tasks.

4.2. Step 2: Selective Multiple Instance Transfer Learning Classifier Construction

If the source task and target task are identified as related in step one, we then construct an SVM-based classifier to transfer knowledge from the source task to the target task. Since a transfer learning classifier is always built on the top of the classifier for single task learning [16], we extend our previous SMILE method [13,37] to propose a transfer multiple instance learning classifier, which has been shown to be more accurate than others.

Suppose we train SVM on (B_I^s, Y_I^s) for the source task (Task 1) and on (B_J^t, Y_J^t) for the task task (Task 2):

$$\mathbf{w}_1 = \mathbf{w}_o + \mathbf{v}_1 \quad and \quad \mathbf{w}_2 = \mathbf{w}_o + \mathbf{v}_2, \quad (7)$$

where $\mathbf{w}_o$ is a common parameter and $\mathbf{v}_1$ and $\mathbf{v}_2$ are specific parameters. We assume $f_1 = \mathbf{w}_1 \cdot \phi(\mathbf{x}) + b_1$ and $f_2 = \mathbf{w}_2 \cdot \phi(\mathbf{x}) + b_2$ are two hyperplanes for both tasks.

Algorithm 1: Optimization heuristics for Selective transfer multiple instance learning.

```
For source task,initialize positive candidate for each positive bag:
    $arg \max_{\mathbf{x} \in B_I : Y_I = 1} \frac{1}{2}(R(\mathbf{x}, S_s^+) + 1 - R(x, S_s^-))$    (13)
Put the initialized positive candidates in $S_p^{s+}$. For thetarget task, similarly, initialize positive candidates and put them in $S_p^{t+}$ ;
$l = 0$;
repeat
    $l = l + 1$;
    Obtain $S_p^{s+}$, $S_a^{s+}$, $S_n^{s-}$ andcalculate memberships for $S_a^{s+}$ based on Equations (8) and (9);
    Obtain $S_p^{t+}$, $S_a^{t+}$, $S_n^{t-}$ and memberships for $S_a^{t+}$ in a similar way;
    Solve Problem (11) to obtain $\alpha$ for bothtasks ;
    Let $\alpha^{(l)} = \alpha$ ;
    for (each positive bag $B_I : Y_I = 1$ in source task) do
        for (each instance $\boldsymbol{x}_i$ in$B_I$) do
            Let $\mathbf{x}_i$ be positive candidate and update $S_p^{s+}$, $S_a^{s+}$;
            Using $\alpha^{(l)}$, calculate the valueof $F$, denoted as $F(\mathbf{x}_i)$, based on (11);
        end
        Obtain new positivecandidate which returns $arg \max_{\mathbf{x} \in B_I} F(\mathbf{x}_i)$;
    end
    Similarly, we have new positive candidate foreach positive in target task ;
    Replace $S_p^{s+}$ and $S_p^{t+}$ with the new positive candidates, also update $S_a^{s+}$ and $S_a^{t+}$;
    Set $F^{(l)} = F(\alpha^{(l)})$ ;
until $|F^l - F^{(l-1)}| \leq \epsilon F^l$;
Output:($\mathbf{w}_0$, $\mathbf{v}_1$, $\mathbf{v}_2$, $b_1$, $b_2$)
```

For source task S, let S_p^{s+}, S_a^{s+}, S_n^{s-} to store positive candidates in the positive bags, the remaining instances except for the positive candidates in the positive bags, and the instances from the negative bags respectively. The initialized positive candidates are the instances which have higher similarity with the positive bags which share lower similarity with the negative bag, which is detailed in the beginning of Algorithm 1 . For the instances in S_p^{s+} or S_n^{s-}, their memberships are set to positive one or negative one, respectively. For the instances in S_a^{s+}, they have bi-memberships $(m_s^+(\mathbf{x}_i), m_s^-(\mathbf{x}_i))$ towards the positive and negative classes respectively. Let us define

$$R(x, S) = \frac{1}{2} \sum_{\mathbf{x}_i \in S} e^{-\|\phi(\mathbf{x}) - \phi(\mathbf{x}_i)\|},$$

then, both memberships are calculated as follows

$$m_s^+(\mathbf{x}_i) = \frac{1}{2}(R(\mathbf{x}, S_p^{s+}) + 1 - R(\mathbf{x}, S_n^{s-})), \quad (8)$$

$$m_s^-(\mathbf{x}_i) = \frac{1}{2}(R(\mathbf{x}, S_n^{s-}) + 1 - R(\mathbf{x}, S_p^{s+})) \quad (9)$$

Based on the above definitions, if an instance is close to the positive candidates while far from the negative instances, it has large membership towards the positive while low membership towards the negative. We further let $S^{s'} = S_p^{s+} \cup S_a^{s+}$ and $S^{s''} = S_a^{s+} \cup S_n^{s-}$.

Similarly, for the target task, we have S_p^{t+}, S_a^{t+}, S_n^{t-} based on the above operation for the source task, and let $S^{t'} = S_p^{t+} \cup S_a^{t+}$ and $S^{t''} = S_a^{t+} \cup S_n^{t-}$. Based on these, the following multiple instance transfer learning classifier is constructed to transfer knowledge from the source task to the target task.

$$\begin{aligned}
\min_{\mathbf{w}_0, \mathbf{v}_t, \xi} \quad & \frac{1}{2}\|\mathbf{w}_0\|^2 + C_1\|\mathbf{v}_1\|^2 + C_2\|\mathbf{v}_2\|^2 + \\
& C(\sum_{S^{s'}} m_s^+(\mathbf{x}_i)\xi_i + \sum_{S^{s''}} m_s^+(\mathbf{x}_j)\xi_j + \\
& \sum_{S^{t'}} m_t^+(\mathbf{x}_k)\xi_k + \sum_{S^{t''}} m_t^-(\mathbf{x}_h)\xi_h) \\
s.t. \quad & (\mathbf{w}_0 + \mathbf{v}_1) \cdot \phi(\mathbf{x}_i)) + b_1 \geq 1 - \xi_i, \\
& (\mathbf{w}_0 + \mathbf{v}_1) \cdot \phi(\mathbf{x}_j)) + b_1 \leq -1 + \xi_j,
\end{aligned}$$

$$\begin{aligned}&(\mathbf{w}_0+\mathbf{v}_2)\cdot\phi(\mathbf{x}_k))+b_2\le 1-\xi_k,\\&(\mathbf{w}_0+\mathbf{v}_2)\cdot\phi(\mathbf{x}_h))+b_2\le -1+\xi_h,\\&\xi_i\ge 0,\quad \xi_j\ge 0,\quad \xi_k\ge 0\quad \xi_h\ge 0.\end{aligned}\tag{10}$$

For the above model, we explain each term:

- Parameters C_1 and C_2 control the preference of the two tasks. If $C_1 > C_2$, Task 1 is preferred to Task 2; otherwise, Task 2 is preferred to Task 1. Parameter C is a parameter to balance the margin and errors and ξ are slack variables.
- In the source task, for instance $\mathbf{x}_i$ in S_p^{s+}, its membership is $m_s^+(\mathbf{x}_i)=1$ towards positive class. Then, each instance in $S^{s'}$ has $m_s^+(\mathbf{x}_i)$ towards the positive class. We set $m_s^-(\mathbf{x}_j)=-1$ for each instance in S_n^{s-}, thus each instance in $S^{s''}$ has $m_s^-(\mathbf{x}_j)$ towards the negative class. Similarly, we have $m_t^+(\mathbf{x}_k)$, $m_t^-(\mathbf{x}_h)$ for the target task.

Theorem 1 By introducing Lagrangian multipliers α_i^{s+}, α_j^{s-}, α_k^{t+} and α_h^{t-} for the instances in $S^{s'}$, $S^{s''}$, $S^{t'}$, $S^{t''}$, the solution of Problem (10) is to resolve the dual problem:

$$\begin{aligned}min\ \ F(\alpha)=&\frac{1}{2}\|\mathbf{w}_0\|^2+C_1\|\mathbf{v}_1\|^2+C_2\|\mathbf{v}_2\|^2-\sum_{S^{s'}}\\&\alpha_i^{s+}(\mathbf{w}_0+\mathbf{v}_1)\phi(\mathbf{x}_i)+\sum_{S^{s''}}\alpha_j^{s-}(\mathbf{w}_0+\mathbf{v}_1)\phi(\mathbf{x}_j)\\&-\sum_{S^{t'}}\alpha_k^{t+}(\mathbf{w}_0+\mathbf{v}_2)\cdot\phi(\mathbf{x}_k)+\sum_{S^{t''}}\alpha_h^{t-}\\&(\mathbf{w}_0+\mathbf{v}_2)\cdot\phi(\mathbf{x}_h)+\sum_{S^{s'}}\alpha_i^{s+}+\sum_{S^{s''}}\alpha_j^{s-}\\&+\sum_{S^{t'}}\alpha_k^{t+}+\sum_{S^{t''}}\alpha_h^{t-}\end{aligned}\tag{11}$$

$$\begin{aligned}s.t\ \ &0\le\alpha_i^{s+}\le Cm_s^+(\mathbf{x}_i),\quad 0\le\alpha_j^{s-}\le Cm_s^-(\mathbf{x}_j)\\&0\le\alpha_k^{t+}\le Cm_t^+(\mathbf{x}_k),\quad 0\le\alpha_h^{t-}\le Cm_t^-(\mathbf{x}_h)\\&\sum_{S^{s'}}\alpha_i^{s+}-\sum_{S^{s''}}\alpha_j^{s-}=0,\quad \sum_{S^{t'}}\alpha_k^{t+}-\sum_{S^{t''}}\alpha_h^{t-}=0,\end{aligned}$$

where

$$\begin{aligned}\mathbf{v}_1&=\frac{1}{2C_1}\Big(\sum_{S^{s'}}\alpha_i^{s+}\cdot\mathbf{x}_i-\sum_{S^{s''}}\alpha_i^{s-}\cdot\mathbf{x}_j\Big)\\\mathbf{v}_2&=\frac{1}{2C_2}\Big(\sum_{S^{t'}}\alpha_k^{t+}\cdot\mathbf{x}_k-\sum_{S^{t''}}\alpha_h^{t-}\cdot\mathbf{x}_h\Big)\\\mathbf{w}_0&=C_1\mathbf{v}_1+C_2\mathbf{v}_2\end{aligned}\tag{12}$$

By substituting $\mathbf{w}_0$, $\mathbf{v}_1$ and $\mathbf{v}_2$ into Problem (11), each item containing $\phi(\mathbf{x})$ can be written as kernel formula $K(,)$.

We use an alternative framework, as suggested in MI-SVM [3] and SMILE [13], to obtain the transfer classifier $f_2(\mathbf{x})=(\mathbf{w}_0+\mathbf{v}_2)\cdot\phi(\mathbf{x})+b_2$, which is listed in Algorithm 1. For a test bag B_J, if there exists at least one instance $\mathbf{x}_j$ in B_J which satisfies

$$f(\mathbf{x}_j)=\mathbf{w}_t\cdot\mathbf{x}_j>0,\tag{14}$$

B_J is labeled as positive; otherwise, it is labeled as negative.

Although we extended SMILE method for multiple instance transfer learning; the optimization problem (10) here is different from the optimization problem (26) in [13]. In our problem (10), for the source task, $S^{s'}$ contains positive candidates and the instances having bi-memberships towards the positive and negative, while $S^{s''}$ consists of the negative instances and the instances having bi-memberships. So do the set $S^{t'}$ and $S^{t'}$ for the target task. In addition, We include the multiple instance learning classifiers for the source and target tasks at the same time; while our previous method in [13] works for a single task, not for transfer learning problem.

Since we utilize alternative framework, the convergence of SMITL is similar to mi-SVM and SMILE methods. The value of $F(.)$ is determined by the Lagarange multipliers α and the positive candidates in subset S_p^{t+}. We alternatively optimize the Lagrange multipliers and positive candidate is an EM-like fashion to minimize the values $F(.)$. Based on this, we have the following relations.

$$F(l)\ge F(l+1)\tag{15}$$

It is seen that the value of $F(.)$ is monotonically decreased during the whole process of optimization. Therefore, the procedure will converge until $|F^l-F^{(l-1)}|\le\epsilon F^l$ satisfies. Here ϵ is a threshold, which is set to be 0.01 in the experiments.

It is noted that, if SMITL determines the source task is related with the target task, SMITL then conducts the transfer learning step to learn a classifier for the target task. On the contrary, the knowledge transfer will degrade the accuracy of classifier; in this case, SMITL won't conduct the transfer learning step, while constructs a classifier only on the target task as SMILE does. In this case, SMITL degenerates into SMILE method.

5. Experiments

5.1. Baselines and Metrics

In this section, we will investigate the performance of our approach SMITL empirically. For comparison, another four MIL methods are used as baselines. The first baseline is the classic mi-SVM [3], which is a typical multiple instance learning baseline. The second baseline is SMILE method [13] which trains a single MIL task by considering the similarity of each sample in the training. The SMILE baseline is to show the improvement of our transfer learning method over a single task multiple instance learning method. The third baseline is SubMIL [45], which is sub-space-based approach to determine discriminative subspaces for multi-instance data. The fourth one is MITL (multiple instance transfer learning) [42], which aims at multi-task problems; however, it does not determine the similarity of a task with the target task.

The performance of an MIL problem is generally evaluated based on accuracy [8,13]. We use accuracy as metrics in the experiments.

5.2. Data Sets and Settings

To evaluate the properties of our framework, we conduct experiments on 20 Newsgroups[2], Reuters-21578[3] and Web-KB[4]. The first two data sets have hierarchical structures. The 20 Newsgroups corpus contains several top categories, and under the top categories, there are 20 sub-categories where each subcategory has 1000 samples. Similarly, Reuters-21578 contains Reuters news wire articles organized into five top categories, and each category includes different sub-categories. Web-KB contains 8,282 web pages and seven categories and each Web page belongs to one category.

The textual data sets we used are not originally designed for evaluating multiple instance learning and transfer learning. Similar to the operations in [51–53], we reorganize the original data in a way for the transfer multiple instance learning problem as follows.

For the 20 Newsgroups, since each news is very short, the same as [51], we use several news to form a bag and each news is considered as an instance. Specifically, we first choose a sub-category (a_1) from a top category (A), and consider this sub-category (a_1) as a positive class. Second, for each positive bag, the instances from the positive class (a_1) and the other categories form a positive bag.

[2] Available at http://people.csail.mit.edu/jrennie/20Newsgroups/.

[3] Available at http://www.daviddlewis.com/resources/testcollections/.

[4] Available at http://www.cs.cmu.edu/webkb/.

Table 1
Description of the data sets. The account number is considered as dataset number, the number of source task bags and target task bags are 1500 and 800 for each dataset.

	Source Task	Target task		Source Task	Target task
1	Com-wind.misc	Com-wind.x	2	Rec-autos	Rec-moto
3	Sci-elec	Sci-med	4	Talk-religion	Talk-christ
5	Com-graphics	Rec-baseball	6	Rec-hockey	Sci-space
7	Sci-crypt	Alt-atheism	8	Talk-forsale	Com-hard
9	Orgs(1)	Orgs(2)	10	People(1)	People(2)
11	People (1)	Places(1)	12	Staff(1)	Staff (2)
13	Faculty (1)	Faculty(2)	14	Student(1)	Project (2)

Third, for each negative bag, the instances are randomly derived from the categories except for sub-category (a_1). Based on this, we generate 1500 bags for the source task and the number of positive and negative bags are similar. For the target task, we choose another sub-category as the positive class and generate 800 bags.

For the Reuters-21578, each top category has many subcategories, for example, “people” has 267 sub-categories and the size of each sub-category is not always large. We organize it as follows. For a top category (A), all of the subcategories are organized into two parts (denoted as $A(1)$ and $A(2)$), and each part is approximately equal in size. In a similar way to [53], if we consider the documents in $A(1)$ category as positive, then each document in $A(1)$ is considered as a positive bag, and each instance corresponds to a text segment enclosed in one sliding window of a size of around 70. For the negative bags, we randomly select documents from the other category except for (A). If we regard the documents in A_2 category as positive, the same as the operations for category A_1, we can generate a data set for multiple instance learning.

For the Web-KB : There are 8,282 web pages and seven categories, i.e., “student”, “faculty”, “staff”, “department”, “course”, “project” and “other”. Each Web page belongs to one category. For this dataset, since it does not have hierarchical structure, we organize it as follows. For one categories (A), we randomly split it into two parts (denoted as $A(1)$ and $A(2)$) to organize the positive bags for the source task and target task respectively. For the negative bags of source and target tasks, we randomly select documents from the other category except for (A). Since the positive bags of source and target tasks are generated from the same category (A), we add random noise to the features of $A(2)$, it is then considered as related tasks. In a similar way, we generate the positive bags of the source and target tasks from different categories (named (A) and (B)), we obtain tasks not related.

Using the above operations, we generate data sets for the source task and target task. Both tasks are listed in the first three columns of Table 1. The first eight pairs of data sets are generated using 20 Newsgroups group; each data set is written as (A-a) format where (A) means top category and (a) is the sub-category under category (A), which is considered as a positive class. The last four pairs of data are generated from Reuters-21578; each is denoted as $A(1)$ or $A(2)$ where (A) is the top category and (1) or (2) means which grouped sub-category is considered as positive.

In the experiment, the linear kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$ is used since it always performs well for text classification [54]. The configurations of the algorithms in the experiments are as follows. In our SMITL method, C_1 and C_2 control the tradeoff between the source task and target task. Since we care about the target task more than the source task, we set $C_2 > C_1$ and C_1, C_2, C is chosen from 1 to 1000, the threshold ϵ is set to be 0.01. For the parameters contained in SMILE [13], SubMIL [45] and MITL [42], we adopt the same settings as their own work. For SMILE [13] method, let $C_1 = C_2$ and $C_3 = C_4$, with each of them being from 2^{-5} to 2^5. For mi-SVM, we let the parameter C chosen from 1 to 1000. For MITL [42] method, let C_1, C_2 search from the exponential 2^{-4} to 2^4. All the experiments are on a laptop with a 2.8 GHz processor and 3GB DRAM under the Windows XP System. The SMITL method is implemented based on LibSVM [55].

The difference between traditional multi-instance learning for single task and the transfer multi-instance learning setting is that: in the former one, we only utilize the target class data to obtain the performance; in the latter one, we incorporate the source task data into the classifier learning of the target class.

5.3. Experimental Results

For the target data set, we randomly choose about 10% to form a training set and the rest is used for testing. This is because transfer learning always assumes that we do not have sufficient training data for the target task, and utilize the data information of the source task to help the learning of the target task. For the corresponding source data, since we are more concerned about the performance of the target task we incorporate them into training. To avoid a sampling bias, we repeat the above process ten times, and calculate the average accuracy values for the twelve datasets.

For the first four, the ninth, tenth, eleventh, twelfth pairs of source and target tasks, the positive classes of each pair belong to the same top category, and therefore they should be considered as related tasks. For these dataset, SMITL can successfully identify that they are related tasks, and then conducts the latter transfer learning step on these datasets. The performance comparison of mi-SVM, SubMIL,SMILE, MITL and our proposed SMITL is shown in the Table 2. From the Figure, we observe that SMITL always performs better than mi-SVM, SMILE and MITL. Further, we find that the MITL method also outperforms mi-SVM and SMILE. This is because SMITL and MITL consider transfer knowledge from the source task, while mi-SVM, SubMIL, SMILE were designed to build a classifier solely on the target task without considering the knowledge transfer.

In addition, we utilize nonparametric statistical Wilcoxon test [56,57] to compare the performance of each previous work with our proposed SMITL method. In general, the properties of the dataset are not assumed to have uniform distributions such as independence, normality, and homoscedasticity, while Wilcoxon test does not require the data distribution having uniform distributions [56,57]. We then utilize Wilcoxon test in the experiments. In general, if a test value is lower than the confidence level 0.05, there is a significant difference between SMITL and the method compared. For each method, the test value between it and SMITL is shown in Table 2. We further report the test values ranges between the method and SMITL over the datasets (the first four, the ninth, tenth, eleventh, twelfth datasets) in Table 3. It can be seen that, the top test value of each method to SMITL is less than the confidence level 0.05, which means that SMITL can obtain a improvement performance in a statistical test view.

After compare the performance of the SMITL over mi-SVM, SubMIL, SMILE and MITL methods, it is necessary to apply the test using the results obtained in all datasets. We utilize Wilcoxon test [56,57] to calculate the test values between SMITL and mi-SVM, SubMIL, SMILE, MITL respectively using the average value. The test values are 0.0078, 0.0078, 0.0141 and 0.0078 respectively. The corresponding R^+ and R^- values between SMITL and other method are 8 and 0 respectively. It is shown that SMITL performs better than mi-SVM, SubMIL, SMILE and MITL in a significant difference.

For the fifth, sixth, seventh, eighth, eleventh and twelfth pairs of the source and target tasks, each pair of tasks is from different top categories and they are therefore considered not to be related. For the not related tasks, it believes that the transfer learning step won’t help the target task to train a classifier, while degrades the performance of target classifier, which is called negative transfer [18,19]. The results of mi-SVM, SubMIL, SMILE, MITL and SMITL are

Table 2
Accuracy obtained by mi-SVM, SubMIL, SMILE, MITL and SMITL on the target task for the related tasks

Data Number	mi-SVM	SubMIL	SMILE	MITL	SMITL
1	0.610 (0.001)	0.592 (< 0.001)	0.625 (0.001)	0.657 (0.011)	0.678
2	0.625 (0.013)	0.632 (0.012)	0.644 (0.018)	0.673 (0.020)	0.694
3	0.803 (0.010)	0.785 (< 0.001)	0.813 (0.022)	0.837 (0.031)	0.863
4	0.585 (0.015)	0.573 (0.012)	0.602 (0.035)	0.635 (0.037)	0.672
9	0.732 (0.021)	0.738 (0.022)	0.754 (0.027)	0.773 (0.031)	0.825
10	0.758 (0.035)	0.765 (0.025)	0.782 (0.040)	0.812 (0.045)	0.858
12	0.632 (0.013)	0.638 (0.020)	0. 685 (0.025)	0.723 (0.029)	0.786
13	0.712 (0.016)	0.723 (0.021)	0. 752 (0.029)	0.803 (0.036)	0.853

Table 3
Test values ranges of each method to SMITL over six datasets (the first four, the ninth and the tenth datasets).

	mi-SVM	SubMIL	SMILE	MITL
p-values range	[0.001, 0.035]	[< 0.001, 0.025]	[0.001, 0.040]	[0.011, 0.045]

Table 4
Accuracy of mi-SVM, SubMIL, SMILE, MITL and SMITL on the target task for the not related tasks. Here, if SMITL determines the tasks are not related, the performance of SMITL is the same as SMILE

Data Number	mi-SVM	SubMIL	SMILE	MITL	SMITL	
	Accuracy	Accuracy	Accuracy	Accuracy	Transfer?	Accuracy
5	0.701	0.693	0.712	0.611	n	0.712
6	0.776	0.778	0.796	0.705	n	0.796
7	0.683	0.686	0.708	0.595	n	0.708
8	0.778	0.768	0.793	0.703	n	0.793
11	0.773	0.779	0.793	0.754	n	0.793
12	0.732	0.748	0.788	0.685	n	0.788

listed in the Table 4. For these datasets, the task similarity criterion in SMITL determines the source and the target tasks are not related except for the twelfth pair of tasks, and won't conduct the transfer learning step. For the datasets which are determined not to be related, SMITL constructs a classifier on the target classifier, as SMILE does, and reports the same performance as SMILE method. However, MITL does not judge whether the source and target tasks are related, and builds a multi-task classifier for the target task; as a result, the data information of the source task won't help the learning of the target classifier. On the contrary, SMILE performs remarkably better than MITL method. In all, we observe that SMITL can determine the similarity of source and target tasks for most cases and shows superior performance compared with others.

In Fig. 1, we present the convergence curves of SMITL on the data sets which were identified as being related by SMITL. It can be seen that SMITL always converges at, or close to the best performance points, where the rates of the convergence are very fast. SMITL converges in less than 10 iterations on most data sets. Therefore, we believe SMITL is efficient for these data sets.

From the experiments, we observe that SMITL can well identify whether the source and target tasks related or not, and deliver an accurate classifier by incorporating the source task data into the learning of the target task. For the related source and target tasks, the classifier built with the help of the source task data significantly outperforms the multi-instance classifier only on the target task.

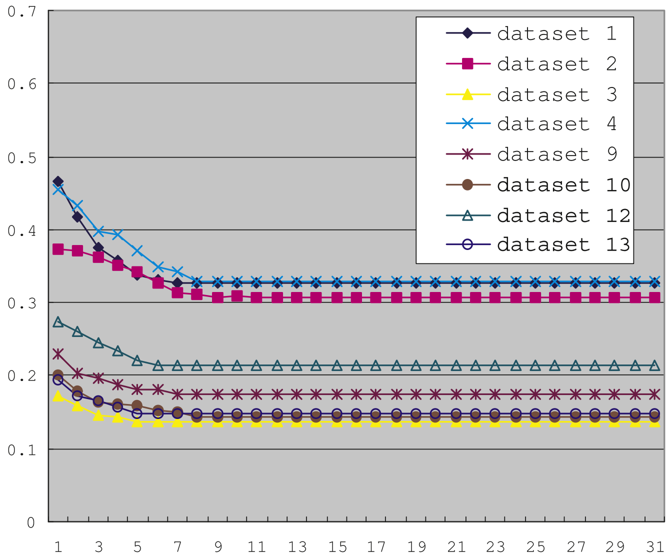

Fig. 1. The error rate curves during iteration on the data sets.

6. Conclusions and Future Work

This paper addresses the problem of multiple instance transfer learning for text categorization. To provide a safe transfer from a source task to a target task, this paper has proposed a selective multiple instance transfer learning (SMITL). Our proposed approach first determines the relation of both tasks by investigating the similarity of the positive features from two tasks, and then builds a multiple instance transfer learning classifier to transfer knowledge from the source task to the target task. Extensive experiments have been conducted to investigate the performance of our proposed approach, and statistical results show that SMITL can determine the similarity of the source and target tasks for most data sets and outperforms classic multiple instance learning methods.

About the computational complexity of the algorithm, since transfer learning models are always based on the models built on the single task model, it believes that they therefore need more computational cost compared with single task models. Therefore, SMITL and MITL methods need around twice time cost compared to SMILE methods. Additionally, some learning methods, such as [58,59] are proposed to speed up SVM, they can be adopted to accelerate the training efficiency of method. In the future, we plan to design a better mechanism to determine the similarity between the source and target tasks by considering the bag data distributions. Second, we plan to design an efficient solution to speed up the select transfer multiple instance learning method for other data types.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 61472090, Grant 61472089 and Grant 61672169, in part by the NSFC-Guangdong Joint Found U1501254, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014133, in part by the Natural Science Foundation of Guangdong under Grant 2015A030313486, in part by the Science and Technology Planning Project of Guangzhou under Grant 201707010492 and Grant 201604016041.

References

[1] W. Pan, Q. Yang, Transfer learning in heterogeneous collaborative filtering domains, Artificial Intelligence 197 (2013) 39–55.
[2] M. Long, J. Wang, Transfer learning with graph co-regularization, IEEE Transactions on Knowledge and Data Engineering 26 (7) (2014) 1805–1818.
[3] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, Advances in Neural Information Processing Systems, 2002.
[4] Y. Chen, J. Bi, J. Wang, Miles: multiple-instance learning via embedded instance selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12) (2006) 1931–1947.
[5] R. Rahmani, S. Goldman, Missl: Multiple instance semi-supervised learning, International Conference on Machine Learning, 2006.
[6] Y. Hang, A.C. Surendran, J.C. Platt, M. Narasimhan, Learning from multi-topic web documents for contextual advertisement, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2008.
[7] Z. Lu, Y. Zhu, S. Pan, Xiang, Y. E. Wang, Q. Yang, Source free transfer learning for text classification, American Association for Artificial Intelligence (2014) 122–128.
[8] O. Maron, T. Lozano-Perez, A framework for multiple-instance learning, Advances in Neural Information Processing Systems, 1998.
[9] Q. Zhang, S. Goldman, Em-dd: An improved multiple-instance learning technique, Advances in Neural Information Processing Systems, 2001.
[10] Y. Chen, J. Wang, Image categorization by learning and reasoning with regions, Journal of Machine Learning Research 5 (2004) 913–939.
[11] J. Wang, J. Zucker, Solving the multiple instance problem: A lazy learning approach, International Conference on Machine Learning, 2000.
[12] Z. Zhou, J. Xu, On the relation between multiinstance learning and semi-supervised learning, International Conference on Machine Learning, 2007.
[13] Y. Xiao, B. Liu, Z. Hao, L. Cao, A similarity-based classification framework for multiple-instance learning, IEEE Transactions on Cybernetics 44 (4) (2014) 500–515.
[14] E. Zhong, W. Fan, Q. Yang, User behavior learning and transfer in composite social networks, ACM Transactions on Knowledge Discovery from Data 8 (1) (2014) 1–32.
[15] B. Tan, Song, E. Y. Zhong, Q. Yang, Transitive transfer learning, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2015) 1155–1164.
[16] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2010) 1345–1359.
[17] H.Y. Wang, Q. Yang, Transfer learning by structural analogy, American Association for Artificial Intelligence, 2011.
[18] T.R. Michael, Z. Marx, L.P. Kaelbling, To transfer or not to transfer, NIPS Workshop, 2005.
[19] B. Cao, J. Pan, Y. Zhang, D.Y. Yeung, Q. Yang, Adaptive transfer learning, American Association for Artificial Intelligence, 2010.
[20] X.X. Shi, Q. Liu, W. Fan, P.S. Yu, R. Zhu, Transfer learning on heterogenous feature spaces via spectral transformation, IEEE International Conference on Data Mining, 2010.
[21] X.X. Shi, Q. Liu, W. Fan, Q. Yang, P.S. Yu, Predictive modeling with heterogeneous sources, SIAM Conference on Data Mining, 2010.
[22] S.J. Pan, X. Ni, J.T. Sun, Q. Yang, Z. Chen, Cross-domain sentiment classification via spectral feature alignment, International World Wide Web Conference, 2010.
[23] T. TEvgeniou, M. Pontil, Regularized multi–task learning, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2004.
[24] K. Yu, V. Tresp, A. Schwaighofer, Learning gaussian processes from multiple tasks, International Conference on Machine Learning, 2005.
[25] A. Schwaighofer, V. Tresp, K. Yu, Learning gaussian process kernels via hierarchical bayes, in: Advances in Neural Information Processing Systems, 2005.
[26] N.D. Lawrence, J.C. Platt, Learning to learn with the informative vector machine, in: International Conference on Machine Learning, 2004.
[27] R. Raina, A.Y. Ng, D. Koller, Constructing informative priors using transfer learning, International Conference on Machine Learning, 2006.
[28] S.I. Lee, V. Chatalbashev, D. Vickrey, D. Koller, Learning a meta-level prior for feature relevance from multiple related tasks, International Conference on Machine Learning, 2007.
[29] W. Dai, Q. Yang, G.-R. Xue, Y. Yu, Boosting for transfer learning, International Conference on Machine Learning, 2007.
[30] R. Raina, A. Battle, H. Lee, B. Packer, A.Y. Ng, Self-taught learning: transfer learning from unlabeled data, International Conference on Machine Learning, 2007.
[31] M. Kandemir, Asymmetric transfer learning with deep gaussian processes, International Conference on Machine Learning (2015) 730–738.
[32] J. Pan, X. Hu, Y. Zhang, P. Li, Y. Lin, H. Li, W. He, L. Li, Quadruple transfer learning: Exploiting both shared and non-shared concepts for text classification, Knowledge-Based Systems 90 (2015) 199–210.
[33] Y. Xiao, B. Liu, P.S. Yu, Z. Hao, A robust one-class transfer learning method with uncertain data, Knowledge and Information Systems 44 (2) (2015) 407–438.
[34] J. Zheng, Z. Jiang, R. Chellappa, Cross-view action recognition via transferable dictionary learning, IEEE Transaction on Image Processing 25 (6) (2016) 2542–2556.
[35] S. Zhao, Q. Cao, J. Chen, Y. Zhang, J. Tang, Z. Duan, A multi-atl method for transfer learning across multiple domains with arbitrarily different distribution, Knowledge-Based Systems 94 (2016) 60–69.
[36] X. Hu, J. Pan, P. Li, H. Li, W. He, Y. Zhang, Multi-bridge transfer learning, Knowledge-Based Systems 97 (2016) 60–74.
[37] Y. Xiao, B. Liu, L. Cao, J. Yin, X. Wu, Smile: A similarity-based approach for multiple instance learning, IEEE International Conference on Data Mining, 2010.
[38] Q. Zhang, S.A. Goldman, Em-dd: An improved multiple-instance learning technique, Advances in Neural Information Processing Systems (2001) 1073–1080.
[39] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, Advances in Neural Information Processing Systems (2002) 561–568.
[40] J. Wu, Y. Yu, C. Huang, K. Yu, Deep multiple instance learning for image classification and auto-annotation, IEEE Conference on Computer Vision and Pattern Recognition (2015) 3460–3469.
[41] M.A. Carbonneau, E. Granger, A.J. Raymond, G. Gagnon, Robust multiple-instance learning ensembles using random subspace instance selection, Pattern Recognition 58 (2016) 83–99.
[42] D. Zhang, L. Si, Multiple instance transfer learning, IEEE International Conference on Data Mining Workshop, 2009.
[43] Q. Wang, L. Ruan, L. Si, Adaptive knowledge transfer for multiple instance learning in image classification, the Association for the Advance of Artificial Intelligence (2014) 1334–1340.
[44] S. Mei, H. Zhu, Multi-label multi-instance transfer learning for simultaneous reconstruction and cross-talk modeling of multiple human signaling pathways, BMC Bioinformatics 417 (2015).
[45] J. Yuan, X. Huang, H. Liu, Submil: Discriminative subspaces for multi-instance learning, NEUROCOMPUTING 173 (4) (2016) 1768–1774.
[46] Q. Wu, M.K. Ng, Y. Ye, Markov-miml: A markov chain based multi-instance multi-label learning algorithm, Knowledge and Information Systems 37 (1) (2014) 83–104.
[47] Q. Wu, M.K. Ng, Y. Ye, Cotransfer learning using coupled markov chains with restart, IEEE Intelligent Systems 29 (4) (2014) 26–33.
[48] Y. Xu, H. Min, Q. Wu, H. Song, B. Ye, Multi-instance metric transfer learning for genome-wide protein function prediction, Scientific Reports (2017) 41831.
[49] G.P.C. Fung, J.X. Yu, H. Lu, P.S. Yu, Text classification without negative examples revisit, IEEE Transactions on Knowledge and Data Engineering 18 (2006) 6–20.
[50] F. Seabastiani, Machine learning in automated text categorization, ACM Computing Surveys 34 (1) (2002) 1–47.
[51] Z. Zhou, Y. Sun, Y. Li, Multi-instance learning by treating instances as non-i.i.d. samples, International Conference on Machine Learning, 2009.
[52] E. Zhong, W. Fan, J. Peng, K. Zhang, Cross domain distribution adaptation via kernel mapping, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2009.
[53] M. Zhang, Z. Zhou, M^3miml: A maximum margin method for multi-instance multi-label learning, IEEE International Conference on Data Mining, 2008.
[54] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys 34 (1) (2002) 1–47.
[55] C. Chang, C. Lin, Libsvm: A library for support vector machines, Available: http://www.csie.ntu.edu.tw/cjlin/libsvm (2001).
[56] J. Demsar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.

[57] J. Derrac, S. Garcia, D. Molina, F. Herrera, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms, Swarm and Evolutionary Computation 1 (1) (2011) 3–18.

[58] F. Shai, S. Katya, Efficient svm training using low-rank kernel representation, Journal of machine learning research 2 (2) (2002) 243–264.

[59] Y.J. Lee, O.L. Mangasarian, Ssvm: A smooth support vector machine for classification, Computational Optimization and Applications 20 (1) (1999) 5–22.

Bo Liu is a professor at the Department of Automation, Guangdong University of Technology. His research interests include machine learning and data mining such as SVM-based multi-class classification, multiple instance learning, data streams and outlier detection. He has published papers on IEEE Transactions on Pattern Analysis and Machine, IEEE Transactions on Neural Networks, IEEE Transactions on Knowledge and Data Engineering, International Joint Conference on Artificial Intelligence (IJCAI), IEEE Conference on Data Mining (ICDM), SIAM Conference on Data Mining (SDM) and International Conference on Information and Knowledge Management (CIKM).

Yanshan Xiao is a professor at the Department of Computer, Guangdong University of Technology. Her research interests include multi-instance learning and machine learning. She has published papers on IEEE Transactions on Pattern Analysis and Machine, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, International Joint Conference on Artificial Intelligence (IJCAI), IEEE Conference on Data Mining (ICDM) and International Conference on Information and Knowledge Management (CIKM).

Zhifeng Hao is a professor at School of Mathematics and Big Data, Foshan University, Foshan, China. His current research interests include design and analysis of algorithm, mathematical modeling, and combinatorial optimization.