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