Which Option is a Better Way to Improve Transfer Learning Performance?

Honghui Xu, Zhipeng Cai, and Wei Li

Department of Computer Science, Georgia State University.
Atlanta, GA 30302-3965, USA
hxu16@student.gsu.edu
zcai@gsu.edu
wli28@gsu.edu

Abstract. Transfer learning has been widely applied in Artificial Intelligence of Things (AIoT) to support intelligent services. Typically, collection and collaboration are two mainstreaming methods to improve transfer learning performance, whose efficiency has been evaluated by real-data experimental results but lacks validation of theoretical analysis. In order to provide guidance of implementing transfer learning in real applications, a theoretical analysis is in desired need to help us fully understand how to efficiently improve transfer learning performance. To this end, in this paper, we conduct comprehensive analysis on the methods of enhancing transfer learning performance. More specifically, we prove the answers to three critical questions for transfer learning: i) by comparing collecting instances and collecting attributes, which collection approach is more efficient? ii) is collaborative transfer learning efficient? and iii) by comparing collection with collaboration, which one is more efficient? Our answers and findings can work as fundamental guidance for developing transfer learning.

Keywords: AIoT · Transfer Learning · Collaborative Transfer Learning

1 Introduction

A compound annual growth rate of 47% in network traffic indicates the proliferation of connected Internet of Things (IoT) devices in recent years [9]. Meanwhile, known as the next generation IoT, Artificial Intelligence of Things (AIoT) [19] relies on machine learning (ML) techniques to extract information from data collected by a number of IoT devices [24] and has been leveraged by real-world services, such as Facebook, Google, Amazon, and Microsoft, to provide individuals with more efficient services. Transfer learning, which takes an advantage of learned knowledge from known categories to novel/unknown categories with limited training instances collected by IoT devices for prediction, is one of the widely-used machine learning models in AIoT [3]. Therefore, improving transfer learning performance plays a critical role to promote the quality of AIoT-based services, which has attracted lots of research attention.

Currently, researchers have realized the performance enhancement of transfer learning through collection and collaboration. On the one hand, the transfer learning performance can be improved by collecting more training instances [12, 18, 34] and more attributes since transfer learning's prediction is based on shared attributes [17, 11, 32]. Even though these two collection methods have been validated by conducting real-data experiments, no study has been carried out to theoretically investigate which collection option is more efficient. On the other hand, inspired by the idea of collaborative learning [10, 14], collaborative transfer learning models have been proposed to improve transfer learning performance by sharing data among institutions [35, 13, 4, 5, 20, 30, 21]. Although these works have demonstrated the efficiency of collaborative transfer learning through the real-data experimental results [20, 4], there lacks theoretically analysis whether collaboration can indeed bring the performance enhancement for transfer learning. Therefore, in literature, the challenging question, "how to efficiently improve transfer learning performance", has not been answered yet. However, with the wide applications of transfer learning models, a thorough analysis is urgent to be done before implementing these models so as to provide individuals, institutions, and organizations with theoretical guidance illustrating how to choose collection or collaboration for enhancing transfer learning performance efficiently.

To fill this gap, in this paper, a series of theoretical analysis is well implemented: i) first of all, we study how to select a more efficient collection option from two collection ways (*i.e.* collecting training instances and collecting more attributes) for performance enhancement in transfer learning; ii) next, we investigate whether collaborative transfer learning in a two-party collaboration scenario is worth doing to enhance performance; and iii) finally, we compare collection and collaboration to analyze which one is more efficient for promoting transfer learning. In our analysis, the conclusions are obtained through rigorous proof and can be used as fundamental guidance for applying transfer learning in real applications. Our multi-fold contributions are addressed as below.

- To the best of our knowledge, this is the first work to perform theoretical analysis on the methods of improving transfer learning performance.
- We provide a theoretical basis for entities to select a more efficient collection option from two collection ways to improve transfer learning performance.
- We prove whether the collaboration in a two-party scenario is efficient to promote transfer learning performance.
- We offer guidance to choose collection or collaboration for performance improvement in transfer learning.

The remainder of this paper is organized as follows. Related works are briefly summarized in Section 2. After formulating problems in Section 3, we detail our analysis process in Section 4, Section 5, and Section 6. Finally, Section 7 concludes this paper and discusses our future work.

2 Related Works

Transfer learning [3, 31], which refers to domain adaptation or sharing of knowledge and representations, is a kind of learning model to achieve the goal of knowledge transfer by transferring information of learned models to changing or unknown data distributions. Transfer learning approaches can be broadly classified into four categories: instance-based, feature-based, parameter-based, and relational-based approaches. i) Instance-based transfer learning approaches [12, 18, 34] are mainly based on the instance weighting strategy. ii) Feature-based approaches [17, 11, 32] map the original features into a new feature space, which can be further divided into two subcategories, including asymmetric and symmetric feature-based transfer learning. The asymmetric approaches [2, 17] transform the features in the source domain to match the features in the target domain, while symmetric approaches [28, 32] attempt to find a common feature space and then map the source and target features into the common feature space. iii) Parameter-based transfer learning approaches [16, 29, 7] keep an eye on transferring the shared knowledge at the parameter level. iv) Relational-based approaches [22, 6] focus on transferring the logical relationship learned in the source domain to the target domain. These transfer learning approaches reduce the effort to collect new labeled training samples and have been widely applied in different real applications, such as disease prediction [23], sign language recognition [8], and target online display advertising [26]. However, no existing work theoretically investigates how to enhance transfer learning performance efficiently.

Collaborative transfer learning was proposed to promote transfer learning performance by sharing data among institutions [4, 5, 20, 30, 21], which is motivated by the idea of collaborative learning [10, 14]. For instance, in the process of healthcare and financial marketing analysis, sharing data among institutions can help enhance the accuracy of medical diagnosis [4] and financial marketing forecasting [20]. However, these works only validate the efficiency of collaboration through the real-data experimental results but lack theoretical analysis.

In a nutshell, the question of how to enhance transfer learning performance efficiently has not been answered with theoretical analysis. To fill this blank, in this paper, we will conduct comprehensively and deeply analysis on the methods of enhancing transfer learning performance, aiming to offer theoretical guidance for the improvement of transfer learning.

3 Problem Formulation

In this section, we mathematically formulate the problem of transfer learning based on PAC learning framework [15].

Transfer learning essentially trains a classifier by using the training instances associated with known labels to predict the testing instances associated with novel/unknown labels [33]. Inspired by the idea of [25], transfer learning can be defined to learn a classifier as follows.

Definition 1. *A classifier consists of two mapping functions, where the first one is the map from a training instance space $\mathcal{X}$ to an attribute space $\mathcal{T}$ (*i.e.* $\mathcal{X} \rightarrow \mathcal{T}$), and the second one is the map from $\mathcal{T}$ to a label space $\mathcal{Y}$ (*i.e.* $\mathcal{T} \rightarrow \mathcal{Y}$).*

In the learning process of $\mathcal{X} \rightarrow \mathcal{T}$, we aim to train T binary classifiers using $\mathcal{X}$, where $T \in \mathbb{N}^+$ is the number of attributes in $\mathcal{T}$; and in the learning process of $\mathcal{T} \rightarrow \mathcal{Y}$, we use these T binary classifiers to predict class labels in $\mathcal{Y}$. Especially, every instance's label is represented by a 0-1 binary attribute vector, where a vector element is set as 1 if and only if the element's corresponding attribute exists in this instance. Let $L_{\mathcal{X}\rightarrow\mathcal{Y}}$ denote the training loss (*e.g.*, prediction error) of the entire process $\mathcal{X} \rightarrow \mathcal{Y}$. Suppose the upper bound of $L_{\mathcal{X}\rightarrow\mathcal{Y}}$ is $\tau \in (0, 1)$. The probability that the training loss does not exceed τ can be represented as:

$$\mathbb{P}(L_{\mathcal{X}\rightarrow\mathcal{Y}} \leq \tau) = 1 - \gamma, \tag{1}$$

where $\gamma \in (0, 1)$ is the error probability. Different from traditional learning processes, transfer learning technically focuses on how to use the training instances to learn T binary classifiers instead of a classifier for prediction. Thus, considering the requirement of $L_{\mathcal{X}\rightarrow\mathcal{Y}} \leq \tau$, the training loss of each binary classifier in the process $\mathcal{X} \rightarrow \mathcal{T}$, denoted by $L_{\mathcal{X}\rightarrow\mathcal{T}}$, should not be larger than $\frac{\tau}{T}$. The probability of holding $L_{\mathcal{X}\rightarrow\mathcal{T}} \leq \frac{\tau}{T}$ for any a classifier in $\mathcal{T}$ is represented by:

$$\mathbb{P}(L_{\mathcal{X}\rightarrow\mathcal{T}} \leq \frac{\tau}{T}) = 1 - \delta, \tag{2}$$

where $\delta \in (0, 1)$ is the error probability. Accordingly, the probability to ensure $L_{\mathcal{X}\rightarrow\mathcal{T}} \leq \frac{\tau}{T}$ for T classifiers is $(1-\delta)^T$, and the probability to ensure $L_{\mathcal{X}\rightarrow\mathcal{Y}} \leq \tau$ with T classifiers is $\mathrm{BinoCDF}(\tau; T, \frac{\tau}{T})$ that is the binomial cumulative distribution probability. On the other hand, according to PAC learning framework [15], there should be at least N training instances in $\mathcal{X}$ such that Eq. (1) and Eq. (2) can be satisfied, in which N can be computed as follows:

$$N = \frac{T^2}{\tau^2} \lceil \ln(\frac{2}{\delta}) \rceil. \tag{3}$$

From Eq. (3), we have $\delta = \frac{2}{e^{\frac{N\tau^2}{T^2}}}$.

To sum up, with N training instances and T classifiers, the performance of the transfer learning process $\mathcal{X} \rightarrow \mathcal{Y}$ can be defined as the total probability $P(N, T)$ that the training loss does not exceed the upper bound τ, *i.e.*,

$$P(N, T) = (1 - \frac{2}{e^{\frac{N\tau^2}{T^2}}})^T \cdot \mathrm{BinoCDF}(\tau; T, \frac{\tau}{T}) \cdot (1 - \gamma). \tag{4}$$

Generally speaking, there are two major methods to improve the performance of transfer learning. i) **Collection:** collecting more training instances for learning better binary classifiers and/or collecting more attributes for predicting novel labels more accurately. ii) **Collaboration:** collaborative transfer learning aims to enhance the learning performance by sharing collected training instances and attributes among paticipants.

To thoroughly understand how to improve transfer learning performance efficiently, the intent of this paper is to deeply investigate the following three problems:

- By comparing collecting instances and collecting attributes, which collection approach is more efficient?
- Is collaborative transfer learning efficient?
- By comparing collection with collaboration, which one is more efficient?

In the following, we elaborate on our theoretical analysis on these three problems in Section 4, Section 5, and Section 6 in order.

4 Collecting Instances vs. Collecting Attributes

First of all, for presentation simplicity, we let

$$f(N,T) = (1 - \frac{2}{e^{\frac{N\tau^2}{T^2}}})^T, \tag{5}$$

and

$$g(T) = \mathrm{BinoCDF}(\tau; T, \frac{\tau}{T}). \tag{6}$$

Then, we can rewrite $P(N,T)$ as below,

$$P(N,T) = f(N,T)g(T)(1-\gamma). \tag{7}$$

The benefit of collecting one more instance r_n and the benefit of collecting one more attribute r_t can be computed by Eq. (8) and Eq. (9), respectively.

$$r_n = \Delta P_N(N,T) = P(N,T) - P(N-1,T). \tag{8}$$

$$r_t = \Delta P_T(N,T) = P(N,T) - P(N,T-1). \tag{9}$$

The benefit difference $k(N,T)$ is calculated as:

$$\begin{aligned} k(N,T) = r_t - r_n =& P(N,T) - P(N,T-1) - [P(N,T) - P(N-1,T)] \\ =& P(N-1,T) - P(N,T-1) \\ =& \left[f(N-1,T)g(T) - f(N,T-1)g(T-1)\right](1-\gamma). \end{aligned} \tag{10}$$

We further use Taylor Theorem [27] and Newton's forward interpolation formula [1] to get a bivar linear function $\tilde{k}(N,T)$ to approximate $k(N,T)$. Meanwhile, to guarantee the existence of gradients in the following calculation process, we approximate $k(N,T)$ at the point $(3,3)$, *i.e.*,

$$\tilde{k}(N,T) = k(3,3) + (N-3)\Delta k_N(3,3) + (T-3)\Delta k_T(3,3). \tag{11}$$

From Eq. (10), we have

$$k(3,3) = \left[f(2,3)g(3) - f(3,2)g(2)\right](1-\gamma). \tag{12}$$

According to Eq. (10) and Newton's forward interpolation formula [1], we can obtain the calculation of gradients in the following.

$$\begin{aligned}\Delta k_N(3,3) &= \frac{k(3,3)-k(2,3)}{3-2} = k(3,3)-k(2,3) \\ &= [f(2,3)g(3)-f(3,2)g(2)-f(1,3)g(3)+f(2,2)g(2)](1-\gamma).\end{aligned} \tag{13}$$

$$\begin{aligned}\Delta k_T(3,3) &= \frac{k(3,3)-k(3,2)}{3-2} = k(3,3)-k(3,2) \\ &= [f(2,3)g(3)-f(3,2)g(2)-f(2,2)g(2)+f(3,1)g(1)](1-\gamma).\end{aligned} \tag{14}$$

By substituting Eq. (12), Eq. (13), and Eq. (14), we can rewrite $\tilde{k}(N,T)$ as:

$$\begin{aligned}\tilde{k}(N,T) = &[f(2,3)g(3)-f(3,2)g(2)]\,(1-\gamma) \\ &+(N-3)[f(2,3)g(3)-f(3,2)g(2)-f(1,3)g(3)+f(2,2)g(2)](1-\gamma) \\ &+(T-3)[f(2,3)g(3)-f(3,2)g(2)-f(2,2)g(2)+f(3,1)g(1)](1-\gamma).\end{aligned} \tag{15}$$

If $\tilde{k}(N,T) \leq 0$, collecting one more training instance is more efficient for performance enhancement; otherwise, collecting one more attribute is more efficient. Thus, we can compare $\tilde{k}(N,T)$ with 0 to decide which collection option is more efficient for one party to improve transfer learning performance.

Theorem 1. *Suppose that one party has N training instances and T attributes. Collecting one more training instance is more efficient than collecting one more attribute to improve transfer learning performance, when any one of the following two conditions holds:*
(i) Condition 1: N, T, and τ satisfy Eq. (16) *and Eq.* (17)*;*
(ii) Condition 2: N, T, and τ satisfy Eq. (18) *and Eq.* (19)*.*

$$\begin{aligned}N \leq &-(T-3)\frac{f(2,3)g(3)-f(3,2)g(2)-f(2,2)g(2)+f(3,1)g(1)}{f(2,3)g(3)-f(3,2)g(2)-f(1,3)g(3)+f(2,2)g(2)} \\ &-\frac{f(2,3)g(3)-f(3,2)g(2)}{f(2,3)g(3)-f(3,2)g(2)-f(1,3)g(3)+f(2,2)g(2)}+3.\end{aligned} \tag{16}$$

$$f(2,3)g(3)-f(3,2)g(2)-f(1,3)g(3)+f(2,2)g(2) > 0. \tag{17}$$

$$\begin{aligned}N \geq &-(T-3)\frac{f(2,3)g(3)-f(3,2)g(2)-f(2,2)g(2)+f(3,1)g(1)}{f(2,3)g(3)-f(3,2)g(2)-f(1,3)g(3)+f(2,2)g(2)} \\ &-\frac{f(2,3)g(3)-f(3,2)g(2)}{f(2,3)g(3)-f(3,2)g(2)-f(1,3)g(3)+f(2,2)g(2)}+3.\end{aligned} \tag{18}$$

$$f(2,3)g(3)-f(3,2)g(2)-f(1,3)g(3)+f(2,2)g(2) < 0. \tag{19}$$

Note that $f(2,3) = (1-\frac{2}{e^{\frac{2\tau^2}{9}}})^3$, $f(3,2) = (1-\frac{2}{e^{\frac{3\tau^2}{4}}})^2$, $f(2,2) = (1-\frac{2}{e^{\frac{\tau^2}{2}}})^2$, $f(3,1) = (1-\frac{2}{e^{3\tau^2}})$, $f(1,3) = (1-\frac{2}{e^{\frac{\tau^2}{9}}})^3$ *according to Eq.* (5)*, and* $g(1) = \mathrm{BinoCDF}(\tau;1,\frac{\tau}{T})$, $g(2) = \mathrm{BinoCDF}(\tau;2,\frac{\tau}{T})$, $g(3) = \mathrm{BinoCDF}(\tau;3,\frac{\tau}{T})$ *according to Eq.* (6)*.*

Proof. The requirement of collecting one more instance to be more efficient is $\tilde{k}(N,T) \leq 0$. According to Eq. (15), we have

$$\begin{aligned}
&[f(2,3)g(3) - f(3,2)g(2)]\,(1-\gamma) \\
&+ (N-3)[f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2)](1-\gamma) \\
&+ (T-3)[f(2,3)g(3) - f(3,2)g(2) - f(2,2)g(2) + f(3,1)g(1)](1-\gamma) \leq 0.
\end{aligned} \tag{20}$$

When Eq. (20) holds, there are three cases for consideration:
(i) if $f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2) > 0$ (*i.e.*, Eq. (17) is satisfied), then we can get Eq. (16);
(ii) if $f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2) < 0$ (*i.e.*, Eq. (19) is satisfied), then Eq. (18) is obtained;
and (iii) if $f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2) = 0$, this is meaningless for our investigated problem.

Thus, Theorem 1 is proved.

Theorem 2. *Suppose that one party has N training instances and T attributes. Collecting one more attribute is more efficient than collecting one more training instance to enhance transfer learning performance, when any one of the following two conditions holds:*
(i) Condition 1: N, T, and τ satisfy Eq. (21) and Eq. (22);
(ii) Condition 2: N, T, and τ satisfy Eq. (23) and Eq. (24).

$$\begin{aligned}
N > &-(T-3)\frac{f(2,3)g(3) - f(3,2)g(2) - f(2,2)g(2) + f(3,1)g(1)}{f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2)} \\
&- \frac{f(2,3)g(3) - f(3,2)g(2)}{f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2)} + 3.
\end{aligned} \tag{21}$$

$$f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2) > 0. \tag{22}$$

$$\begin{aligned}
N < &-(T-3)\frac{f(2,3)g(3) - f(3,2)g(2) - f(2,2)g(2) + f(3,1)g(1)}{f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2)} \\
&- \frac{f(2,3)g(3) - f(3,2)g(2)}{f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2)} + 3.
\end{aligned} \tag{23}$$

$$f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2) < 0. \tag{24}$$

Note that $f(2,3) = (1 - \frac{2}{e^{\frac{2\tau^2}{9}}})^3$, $f(3,2) = (1 - \frac{2}{e^{\frac{3\tau^2}{4}}})^2$, $f(2,2) = (1 - \frac{2}{e^{\frac{\tau^2}{2}}})^2$, $f(3,1) = (1 - \frac{2}{e^{3\tau^2}})$, $f(1,3) = (1 - \frac{2}{e^{\frac{\tau^2}{9}}})^3$ *according to Eq. (5), and* $g(1) = \mathrm{BinoCDF}(\tau; 1, \frac{\tau}{T})$, $g(2) = \mathrm{BinoCDF}(\tau; 2, \frac{\tau}{T})$, $g(3) = \mathrm{BinoCDF}(\tau; 3, \frac{\tau}{T})$ *according to Eq. (6).*

Proof. Collecting one more attribute is a more efficient option, which is equivalent to $\tilde{k}(N,T) > 0$. From Eq. (15), we have

$$\begin{aligned}
&[f(2,3)g(3) - f(3,2)g(2)]\,(1-\gamma) \\
&+ (N-3)[f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2)](1-\gamma) \\
&+ (T-3)[f(2,3)g(3) - f(3,2)g(2) - f(2,2)g(2) + f(3,1)g(1)](1-\gamma) > 0.
\end{aligned} \tag{25}$$

There are three cases for solving Eq. (25):
(i) $f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2) > 0$ (*i.e.*, Eq. (22) is met), then Eq. (21) is obtained;
(ii) $f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2) < 0$ (*i.e.*, Eq. (24) is satisfied), then we can gain Eq. (23);
and (iii) $f(2,3)g(3) - f(3,2)g(2) - f(1,3)g(3) + f(2,2)g(2) = 0$, this is a meaningless case for our investigated problem.

Therefore, Theorem 2 is proved.

5 Whether to Collaboration

Similar to the analysis in Section 4, we use Taylor Theorem [27] and Newton's forward interpolation formula [1] to obtain a bivar linear function $\tilde{P}(N,T)$ to approximate $P(N,T)$. Additionally, to ensure the existence of gradients for $\tilde{P}(N,T)$, we approximate $P(N,T)$ at $(3,3)$ as follows,

$$\tilde{P}(N,T) = P(3,3) + (N-3)\Delta P_N(3,3) + (T-3)\Delta P_T(3,3). \tag{26}$$

Based on Eq. (7) and Newton's forward interpolation formula [1], there exist

$$\begin{aligned} \Delta P_N(3,3) &= \frac{P(3,3) - P(2,3)}{3-2} = P(3,3) - P(2,3) \\ &= f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma), \end{aligned} \tag{27}$$

and

$$\begin{aligned} \Delta P_T(3,3) &= \frac{P(3,3) - P(3,2)}{3-2} = P(3,3) - P(3,2) \\ &= f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma). \end{aligned} \tag{28}$$

With the substitution of Eq. (27) and Eq. (28), $\tilde{P}(N,T)$ can be rewritten as:

$$\begin{aligned} \tilde{P}(N,T) =& P(3,3) + (N-3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\ &+ (T-3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)]. \end{aligned} \tag{29}$$

In this section, our goal is to understand whether collaborative transfer learning is efficient for one party in a two-party (including party **A** and party **B**) collaboration scenario. Without loss of generality, party **A** has N_1 training instances and T_1 attributes, and the party **B** has N_2 training instances and T_2 attributes. If the party **A** does not collaborate with party **B**, according to Eq. (29), party **A** can obtain transfer learning performance $\tilde{P}(N_1,T_1)$ as below,

$$\begin{aligned} \tilde{P}(N_1,T_1) =& P(3,3) + (N_1-3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\ &+ (T_1-3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)]. \end{aligned} \tag{30}$$

If party **A** collaborates with party **B**, according to Eq. (29), party **A**'s transfer learning performance $\tilde{P}(N_1 + N_2, T_1 \cup T_2)$ becomes

$$\begin{aligned}\tilde{P}(N_1 + N_2, T_1 \cup T_2) =& P(3,3) \\ &+ (N_1 + N_2 - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\ &+ (T_1 \cup T_2 - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)].\end{aligned} \tag{31}$$

If $\tilde{P}(N_1 + N_2, T_1 \cup T_2) \geq \tilde{P}(N_1, T_1)$, collaboration is efficient for party **A** to improve transfer learning performance; otherwise, collaboration is not efficient for party **A**. In order words, we should compare $\tilde{P}(N_1, T_1)$ with $\tilde{P}(N_1 + N_2, T_1 \cup T_2)$ in order to judge whether collaboration is efficient for party **A** to enhance transfer learning performance.

Theorem 3. *In a two-party (party* **A** *and party* **B***) collaboration scenario, party* **A** *has N_1 training instances and T_1 attributes, and party* **B** *has N_2 training instances and T_2 attributes. Collaboration is more efficient for party* **A** *to improve transfer learning performance, when T_1, N_2, T_2, and τ satisfy Eq.* (32)*.*

$$N_2 \geq -(T_1 \cup T_2 - T_1)\frac{f(3,3)g(3) - f(3,2)g(2)}{f(3,3)g(3) - f(2,3)g(3)}, \tag{32}$$

where $f(3,3) = (1 - \frac{2}{e^{\frac{\tau^2}{3}}})^3$*,* $f(3,2) = (1 - \frac{2}{e^{\frac{3\tau^2}{4}}})^2$*,* $f(2,3) = (1 - \frac{2}{e^{\frac{2\tau^2}{9}}})^3$ *according to Eq.* (5)*, and* $g(2) = \mathrm{BinoCDF}(\tau; 2, \frac{\tau}{T_1})$*,* $g(3) = \mathrm{BinoCDF}(\tau; 3, \frac{\tau}{T_1})$ *according to Eq.* (6)*.*

Proof. Collaboration is more efficient for party **A**, which means $\tilde{P}(N_1 + N_2, T_1 \cup T_2) \geq \tilde{P}(N_1, T_1)$. By substituting Eq. (30) and Eq. (31), we gain Eq. (33).

$$\begin{aligned}&P(3,3) + (N_1 + N_2 - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\ &+ (T_1 \cup T_2 - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)] \geq \\ &P(3,3) + (N_1 - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\ &+ (T_1 - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)].\end{aligned} \tag{33}$$

Since $f(3,3) > f(2,3)$ and $g(3) > 0$, we have $[f(3,3)g(3) - f(2,3)g(3)] > 0$. Then, by solving Eq. (33), Eq. (32) is obtained. Thus, Theorem 3 is proved.

Theorem 4. *In a two-party (party* **A** *and party* **B***) collaboration scenario, party* **A** *has N_1 training instances and T_1 attributes, and party* **B** *has N_2 training instances and T_2 attributes. Collaboration cannot enhance transfer learning performance for party* **A** *if T_1, N_2, T_2, and τ satisfy Eq.* (34)*.*

$$N_2 < -(T_1 \cup T_2 - T_1)\frac{f(3,3)g(3) - f(3,2)g(2)}{f(3,3)g(3) - f(2,3)g(3)}, \tag{34}$$

where $f(3,3) = (1 - \frac{2}{e^{\frac{\tau^2}{3}}})^3$*,* $f(3,2) = (1 - \frac{2}{e^{\frac{3\tau^2}{4}}})^2$*,* $f(2,3) = (1 - \frac{2}{e^{\frac{2\tau^2}{9}}})^3$ *according to Eq.* (5)*, and* $g(2) = \mathrm{BinoCDF}(\tau; 2, \frac{\tau}{T_1})$*,* $g(3) = \mathrm{BinoCDF}(\tau; 3, \frac{\tau}{T_1})$ *according to Eq.* (6)*.*

Proof. The failure of collaboration to improve performance indicates $\tilde{P}(N_1 + N_2, T_1 \cup T_2) < \tilde{P}(N_1, T_1)$. Accordingly, Eq. (35) can be gained using Eq. (30) and Eq. (31).

$$\begin{aligned} &P(3,3) + (N_1 + N_2 - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\ &+ (T_1 \cup T_2 - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)] < \\ &P(3,3) + (N_1 - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\ &+ (T_1 - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)]. \end{aligned} \tag{35}$$

In Eq. (35), $[f(3,3)g(3) - f(2,3)g(3)] > 0$ because $f(3,3) > f(2,3)$ and $g(3) > 0$. Then, the solution to Eq. (35) is Eq. (34). Thus, Theorem 4 is proved.

From Theorem 3 and Theorem 4, any party in a two-party collaboration scenario can make a judgement whether collaboration is an efficient choice.

6 Collection vs. Collaboration

In this section, we further compare the efficiency of collection and collaboration. Considering a two-party (party **A** and party **B**) collaboration scenario, party **A** has N_1 training instances and T_1 attributes, and party **B** has N_2 training instances and T_2 attributes. Via collaboration, party **A** can increase transfer learning performance to $\tilde{P}(N_1 + N_2, T_1 \cup T_2)$. On the other hand, we assume that party **A** can increase N_1 to N_{NT} and increase T_1 to T_{NT} through collection. Correspondingly, party **A** can enhance transfer learning performance to $\tilde{P}(N_{NT}, T_{NT})$ as shown in Eq. (36).

$$\begin{aligned} \tilde{P}(N_{NT}, T_{NT}) = &P(3,3) + (N_{NT} - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\ &+ (T_{NT} - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)]. \end{aligned} \tag{36}$$

If $\tilde{P}(N_1 + N_2, T_1 \cup T_2) \geq \tilde{P}(N_{NT}, T_{NT})$, collaboration is more efficient for party **A** to improve transfer learning performance; otherwise, collection is more efficient for party **A**. Thus, we need to compare $\tilde{P}(N_{NT}, T_{NT})$ and $\tilde{P}(N_1 + N_2, T_1 \cup T_2)$ so as to help party **A** compare the efficiency of collection and collaboration for performance improvement.

Theorem 5. *In a two-party (party* **A** *and party* **B***) collaboration scenario, party* **A** *has N_1 training instances and T_1 attributes, and party* **B** *has N_2 training instances and T_2 attributes. Assume that party* **A** *can increase N_1 to N_{NT} and increase T_1 to T_{NT} via collection. Collaboration is more efficient than collection for party* **A** *to enhance transfer learning performance, when N_1, N_2, N_{NT}, T_1, T_2, T_{NT}, and τ satisfy Eq. (37).*

$$N_1 + N_2 - N_{NT} \geq -(T_1 \cup T_2 - T_{NT}) \frac{f(3,3)g(3) - f(3,2)g(2)}{f(3,3)g(3) - f(2,3)g(3)}, \tag{37}$$

where $f(3,3) = (1 - \frac{2}{e^{\frac{\tau^2}{3}}})^3$, $f(3,2) = (1 - \frac{2}{e^{\frac{3\tau^2}{4}}})^2$, *and* $f(2,3) = (1 - \frac{2}{e^{\frac{2\tau^2}{9}}})^3$ *according to Eq. (5), and* $g(2) = \mathrm{BinoCDF}(\tau; 2, \frac{\tau}{T_1})$, $g(3) = \mathrm{BinoCDF}(\tau; 3, \frac{\tau}{T_1})$ *according to Eq. (6).*

Proof. If collaboration is more efficient, we have

$$\tilde{P}(N_1 + N_2, T_1 \cup T_2) \geq \tilde{P}(N_{NT}, T_{NT}),$$

which is can be equivalently expressed by Eq. (38) via substituting Eq. (31) and Eq. (36).

$$\begin{aligned}
&P(3,3) + (N_1 + N_2 - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\
&+ (T_1 \cup T_2 - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)] \geq \\
&P(3,3) + (N_{NT} - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\
&+ (T_{NT} - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)].
\end{aligned} \tag{38}$$

Since $f(3,3) > f(2,3)$ and $g(3) > 0$, $[f(3,3)g(3) - f(2,3)g(3)] > 0$. Then, by solving Eq. (38), Eq. (37) is obtained. Therefore, Theorem 5 is proved.

Theorem 6. *In a two-party (party* **A** *and party* **B***) collaboration scenario, party* **A** *has N_1 training instances and T_1 attributes, and party* **B** *has N_2 training instances and T_2 attributes. Assume that party* **A** *can increase N_1 to N_{NT} and increase T_1 to T_{NT} via collection. Collection is more efficient than collaboration for party* **A** *to enhance transfer learning performance, when N_1, N_2, N_{NT}, T_1, T_2, T_{NT}, and τ satisfy Eq. (39).*

$$N_1 + N_2 - N_{NT} < -(T_1 \cup T_2 - T_{NT}) \frac{f(3,3)g(3) - f(3,2)g(2)}{f(3,3)g(3) - f(2,3)g(3)}, \tag{39}$$

where $f(3,3) = (1 - \frac{2}{e^{\frac{\tau^2}{3}}})^3$, $f(3,2) = (1 - \frac{2}{e^{\frac{3\tau^2}{4}}})^2$, $f(2,3) = (1 - \frac{2}{e^{\frac{2\tau^2}{9}}})^3$ according to Eq. (5), and $g(2) = \mathrm{BinoCDF}(\tau; 2, \frac{\tau}{T_1})$, $g(3) = \mathrm{BinoCDF}(\tau; 3, \frac{\tau}{T_1})$ according to Eq. (6).

Proof. Similar to the analysis in Theorem 5, $\tilde{P}(N_1 + N_2, T_1 \cup T_2) < \tilde{P}(N_{NT}, T_{NT})$ is the condition of collection to be more efficient. By substituting Eq. (31) and Eq. (36), we obtain Eq. (40) as follows.

$$\begin{aligned}
&P(3,3) + (N_1 + N_2 - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\
&+ (T_1 \cup T_2 - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)] < \\
&P(3,3) + (N_{NT} - 3)[f(3,3)g(3)(1-\gamma) - f(2,3)g(3)(1-\gamma)] \\
&+ (T_{NT} - 3)[f(3,3)g(3)(1-\gamma) - f(3,2)g(2)(1-\gamma)].
\end{aligned} \tag{40}$$

As $f(3,3) > f(2,3)$ and $g(3) > 0$, $[f(3,3)g(3) - f(2,3)g(3)] > 0$. Thus, for Eq. (40), the solution is Eq. (39); that is, Theorem 6 is proved.

Theorem 5 and Theorem 6 demonstrate the conditions of selecting collaboration or collection as an efficient method for a party to improve transfer learning performance in a two-party scenario.

7 Conclusion and Future Work

In this paper, we comprehensively investigate how to improve transfer learning performance more efficiently. To the best of our knowledge, this is the first work to theoretically analyze the methods of improving transfer learning performance. Specifically, Theorem 1 and Theorem 2 are proposed in Section 4 to help select a more efficient collection option from two collection ways for promoting transfer learning performance. Besides, Theorem 3 and Theorem 4 are presented in Section 5 to help judge whether collaboration is more efficient to enhancing transfer learning performance without considering collection. Moreover, Theorem 5 and Theorem 6 are put forward in Section 6 to help judge whether collaboration is still efficient while considering both collection and collaboration. To sum up, our proposed theorems and conclusions provide the thoroughly theoretical decision-making guidance for improving transfer learning performance.

In our future work, we will advance our theoretical analysis for transfer learning performance by taking into account the cost of collection and collaboration.

Acknowledgment

This work was partly supported by the National Science Foundation of U.S. (2118083, 1912753, 1704287, 1741277, 1829674).

References

1. AL-Sammarraie, O.A., Bashir, M.A.: Generalization of newtons forward interpolation formula. International Journal of Scientific and Research Publications (2015)
2. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine learning **73**, 243–272 (2008)
3. Blanke, U., Schiele, B.: Remember and transfer what you have learned-recognizing composite activities based on activity spotting. In: International Symposium on Wearable Computers. pp. 1–8. IEEE (2010)
4. Chen, Y., Qin, X., Wang, J., Yu, C., Gao, W.: Fedhealth: A federated transfer learning framework for wearable healthcare. IEEE Intelligent Systems pp. 84–94 (2020)
5. Daga, H., Nicholson, P.K., Gavrilovska, A., Lugones, D.: Cartel: A system for collaborative transfer learning at the edge. In: Proceedings of the ACM Symposium on Cloud Computing. pp. 25–37. ACM (2019)
6. Davis, J., Domingos, P.: Deep transfer via second-order markov logic. In: Proceedings of the 26th annual international conference on machine learning. pp. 217–224. ACM (2009)
7. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 109–117. ACM (2004)
8. Farhadi, A., Forsyth, D., White, R.: Transfer learning in sign language. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007)

9. Forecast, G.: Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022. Update **2017**, 2022 (2019)
10. Hsieh, K., Harlap, A., Vijaykumar, N., Konomis, D., Ganger, G.R., Gibbons, P.B., Mutlu, O.: Gaia: Geo-distributed machine learning approaching {LAN} speeds. In: 14th {USENIX} Symposium on Networked Systems Design and Implementation. pp. 629–647. USENIX (2017)
11. Jebara, T.: Multi-task feature and kernel selection for svms. In: Proceedings of the 21st international conference on Machine learning. p. 55. ACM (2004)
12. Jiang, J., Zhai, C.: Instance weighting for domain adaptation in nlp. In: Proceedings of the 45th annual meeting of the association of computational linguistics. pp. 264–271. ACL (2007)
13. Ju, C., Gao, D., Mane, R., Tan, B., Liu, Y., Guan, C.: Federated transfer learning for EEG signal classification. In: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2020, Montreal, QC, Canada, July 20-24, 2020. pp. 3040–3045. IEEE (2020). https://doi.org/10.1109/EMBC44109.2020.9175344, `https://doi.org/10.1109/EMBC44109.2020.9175344`
14. Kang, Y., Hauswald, J., Gao, C., Rovinski, A., Mudge, T., Mars, J., Tang, L.: Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. ACM SIGARCH Computer Architecture News **45**, 615–629 (2017)
15. Kearns, M.J., Vazirani, U.V., Vazirani, U.: An Introduction to Computational Learning Theory. MIT Press (1994)
16. Lawrence, N.D., Platt, J.C.: Learning to learn with the informative vector machine. In: Proceedings of the 21st international conference on Machine learning. p. 65. ACM (2004)
17. Lee, S.I., Chatalbashev, V., Vickrey, D., Koller, D.: Learning a meta-level prior for feature relevance from multiple related tasks. In: Proceedings of the 24th international conference on Machine learning. pp. 489–496. ACM (2007)
18. Liao, X., Xue, Y., Carin, L.: Logistic regression with an auxiliary data source. In: Proceedings of the 22nd international conference on Machine learning. pp. 505–512. ACM (2005)
19. Luo, C., Zhang, F., Huang, C., Xiong, X., Chen, J., Wang, L., Gao, W., Ye, H., Wu, T., Zhou, R., et al.: Aiot bench: towards comprehensive benchmarking mobile and embedded device intelligence. In: International Symposium on Benchmarking, Measuring and Optimization. pp. 31–35. Springer (2018)
20. Ma, Z., Ma, J., Miao, Y., Choo, K.K.R., Liu, X., Wang, X., Yang, T.: Pmkt: Privacy-preserving multi-party knowledge transfer for financial market forecasting. Future Generation Computer Systems **106**, 545–558 (2020)
21. Ma, Z., Ma, J., Miao, Y., Choo, K.R., Liu, X., Wang, X., Yang, T.: PMKT: privacy-preserving multi-party knowledge transfer for financial market forecasting. Future Gener. Comput. Syst. **106**, 545–558 (2020). https://doi.org/10.1016/j.future.2020.01.007, `https://doi.org/10.1016/j.future.2020.01.007`
22. Mihalkova, L., Huynh, T., Mooney, R.J.: Mapping and revising markov logic networks for transfer learning. In: AAAI. vol. 7, pp. 608–614. AAAI (2007)
23. Ogoe, H.A., Visweswaran, S., Lu, X., Gopalakrishnan, V.: Knowledge transfer via classification rules using functional mapping for integrative modeling of gene expression data. BMC bioinformatics **16**, 1–15 (2015)
24. Olmedilla, D.: Applying machine learning to ads integrity at facebook. In: Proceedings of the 8th ACM Conference on Web Science. pp. 4–4. ACM (2016)

25. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Advances in Neural Information Processing Systems. pp. 1410–1418. MIT Press (2009)
26. Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O., Provost, F.: Machine learning for targeted display advertising: Transfer learning in action. Machine learning **95**, 103–127 (2014)
27. Rababah, A.: Taylor theorem for planar curves. Proceedings of the American Mathematical Society **119**, 803–810 (1993)
28. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th international conference on Machine learning. pp. 759–766. ACM (2007)
29. Schwaighofer, A., Tresp, V., Yu, K.: Learning gaussian process kernels via hierarchical bayes. In: Advances in neural information processing systems. pp. 1209–1216. MIT Press (2005)
30. Sharma, S., Xing, C., Liu, Y., Kang, Y.: Secure and efficient federated transfer learning. In: 2019 IEEE International Conference on Big Data. pp. 2569–2576. IEEE (2019)
31. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**, 854–869 (2007)
32. Wang, C., Mahadevan, S.: Manifold alignment using procrustes analysis. In: Proceedings of the 25th international conference on Machine learning. pp. 1120–1127. ACM (2008)
33. Wang, W., Zheng, V.W., Yu, H., Miao, C.: A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology **10**, 1–37 (2019)
34. Wu, P., Dietterich, T.G.: Improving svm accuracy by training on auxiliary data sources. In: Proceedings of the 21st international conference on Machine learning. p. 110. ACM (2004)
35. Yang, H., He, H., Zhang, W., Cao, X.: Fedsteg: A federated transfer learning framework for secure image steganalysis. IEEE Transactions on Network Science and Engineering **14**, 78–88 (2018)