Online Multitask Relative Similarity Learning

Shuji Hao[1], **Peilin Zhao**[2*], **Yong Liu**[3], **Steven C. H. Hoi**[4], **Chunyan Miao**[5]
[1]Institute of High Performance Computing, A*STAR, Singapore
[2]Artificial Intelligence Department, Ant Financial, China
[3]Institute for Infocomm Research, A*STAR, Singapore
[4]School of Information Systems, SMU, Singapore
[5]School of Computer Science and Engineering, NTU, Singapore
haosj@ihpc.a-star.edu.sg, peilin.zpl@antfin.com, liuyo@i2r.a-star.edu.sg,
chhoi@smu.edu.sg, ascymiao@ntu.edu.sg

Abstract

Relative similarity learning (RSL) aims to learn similarity functions from data with relative constraints. Most previous algorithms developed for RSL are batch-based learning approaches which suffer from poor scalability when dealing with real-world data arriving sequentially. These methods are often designed to learn a single similarity function for a specific task. Therefore, they may be sub-optimal to solve multiple task learning problems. To overcome these limitations, we propose a scalable RSL framework named OMTRSL (Online Multi-Task Relative Similarity Learning). Specifically, we first develop a simple yet effective online learning algorithm for multi-task relative similarity learning. Then, we also propose an active learning algorithm to save the labeling cost. The proposed algorithms not only enjoy theoretical guarantee, but also show high efficacy and efficiency in extensive experiments on real-world datasets.

1 Introduction

The objective of relative similarity learning (RSL) is to learn similarity functions from training data with relative constraints, instead of the explicit labels commonly used in conventional classification tasks. RSL has been extensively studied and widely used for many real-world applications, such as web search, image retrieval, data mining [Schultz and Joachims, 2004; Yang and Jin, 2006]. However, existing RSL approaches usually have two main drawbacks. On one hand, most of them are batch-based learning approaches. They may suffer from very expensive re-training cost in the application scenarios (e.g, online social media platforms) where new examples are continuously arriving. On the other hand, previous RSL methods are mainly designed for the single task learning scenario, i.e., learning a single similarity function for only one given task. When applying these methods for multi-task relative similarity learning — a scenario which is common for many real-world applications (e.g., learning multiple similarity functions for different retrieval tasks), one may have to either learn a local similarity metric for each task *independently* or learn a *global* similarity metric for all tasks by combining all training data. Nevertheless, these two solutions are often sub-optimal for solving the multi-task relative similarity learning problems.

In the literature, multi-task learning has been actively studied for classification problems [Cohen and Crammer, 2014; Gonçalves *et al.*, 2016]. However, very few research work has been explored for multi-task relative similarity learning. The most relevant one is [Parameswaran and Weinberger, 2010], in which a multi-task metric learning algorithm, namely mtLMNN, has been proposed by extending the popular LMNN algorithm [Weinberger and Saul, 2009] to multi-task learning scenarios. mtLMNN jointly learns a global metric for multiple tasks and several local metrics, each of which is designed for an individual task. There exist several major limitations of mtLMNN. First of all, mtLMNN is based on batch learning, thus it is usually time consuming and difficult to scale up for large-scale applications. Moreover, mtLMNN enforces to learn a Positive Semi-definite (PSD) distance matrix, which is computationally intensive for large-scale applications. Furthermore, in [Parameswaran and Weinberger, 2010], the mtLMNN model is built based on explicit labeling information which is not as flexible as the relative constraints adopted in this work.

To tackle these challenges, we propose a novel framework, namely Online Multi-Task Relative Similarity Learning (OMTRSL). Specifically, the proposed method simultaneously learns multiple similarity metrics from the relative constraints data via an online learning algorithm. In addition to the high efficiency of the adopted online learning scheme, the learned similarity matrices are also not required to be PSD. These characteristics of OMTRSL make it much efficient and scalable compared to existing RSL approaches. Moreover, we also propose an extension of OMTRSL in an online active learning setting, named as OMTRSL-Active. This extension can reduce the labeling cost and thus further improve the efficiency of the proposed model. In this paper, we theoretically analyze the mistake bounds of both proposed algorithms. Then, we also perform extensive experiments on real-world datasets to demonstrate the empirical performances of these algorithms.

The rest of this paper is organized as follows. Section 2 re-

*Corresponding author

views the most relevant research work. Section 3 introduces the details of the proposed algorithms, as well as their theoretical analysis. Then, in Section 4, we empirically validate the proposed algorithms in terms of both efficacy and efficiency. Finally, Section 5 concludes this study.

2 Related Work

This work is related to two main categories of research studies: online multi-task learning and similarity learning. Next, we briefly review the most relevant work in each category.

2.1 Online Multi-task Learning

Multi-task learning aims to boost the performances of all tasks via information sharing mechanism across all of the tasks. The implicit assumption is that these tasks are related and the labeled data for each task is limited [Hoi *et al.*, 2014; Bakker and Heskes, 2003; Calandriello *et al.*, 2014; Cohen and Crammer, 2014; Zhang *et al.*, 2016]. This work is related to the multi-task learning methods under online settings. In the literature, Cavallanti et al. proposed the first online multi-task learning algorithms [Cavallanti *et al.*, 2010] based on the Perceptron algorithm [Block, 1962]. The learned relationship matrix among tasks was fixed. In [Saha *et al.*, 2011], Saha et al. proposed to adaptively update the relationship matrix via an incremental method. However, the performances of these algorithms were usually limited due to the adopted underlying learning scheme (i.e., Perceptron) [Ammar *et al.*, 2014; Lugosi *et al.*, 2009]. In this work, we adopt a superior first-order based online learning scheme [Crammer *et al.*, 2006]. By assuming all tasks share a sparse basis, Ruvolo and Eaton [Eaton and Ruvolo, 2013; Ruvolo and Eaton, 2014] proposed a series of Efficient Lifelong Learning (ELLA) algorithms. In [Ammar *et al.*, 2014; Calandriello *et al.*, 2014], these ELLA algorithms were used for reinforcement learning. Under this scenario, these algorithms were interpreted as online learning algorithms to learn the tasks one-by-one. However, it is too restrictive to collect the entire data for one task before to learn the other tasks. Moreover, all these algorithms are designed for conventional classification problems. In this work, we study the online multi-task learning for similarity learning problems.

2.2 Similarity Learning

Similarity learning has been extensively studied in machine learning and data mining communities [Chechik *et al.*, 2010; Crammer and Chechik, 2012; Hao *et al.*, 2015; Schultz and Joachims, 2004; Shalev-Shwartz *et al.*, 2004; Yang and Jin, 2006]. Here, we restrict our discussions on the most related multi-task metric learning progresses in the literature. To the best of our knowledge, the first multi-task metric learning approach was proposed by Parameswaran and Weinberger [Parameswaran and Weinberger, 2010], in which a global metric and individual metric for each task were learned together based on the large-margin neighborhood scheme [Weinberger and Saul, 2009]. Based on similar idea, Yang et al. [Yang *et al.*, 2013] proposed to learn multiple metrics by using von Neumann divergence among metrics. Zhang and Yeung [Zhang and Yeung, 2010] proposed to learn a target metric from several related source tasks via multi-task learning approach. Moreover, there were other work applying multi-task metric learning to network data [Fang and Rockmore, 2015] and person re-identification application [Ma *et al.*, 2014]. However, these algorithms were mostly designed for classification problems. In addition, all these algorithms were offline methods, and the learned similarity matrices were required to be PSD. These factors made these algorithms very inefficient and unsuitable for large scale problems. Moreover, these algorithms also required explicit class labels, which were not as flexible as the relative constraints.

3 Online Multi-Task Relative Similarity Learning (OMTRSL)

3.1 Problem Setting

In this work, we investigate the problem of online learning K similarity functions for K related tasks simultaneously. More formally, we denote the similarity function for the k-th task by $S^k(\mathbf{x}, \mathbf{x}'), k \in [K]$, where $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ are two instances (aka samples) from the k-th task, and $[K] = \{1, \ldots, K\}$ denotes the indices of all K tasks. We assume a bi-linear form for the similarity function as follows

$$S^k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{M}^k \mathbf{x}', \tag{1}$$

where $\mathbf{M}^k \in \mathbb{R}^{d \times d}$ is the similarity matrix learned for the k-th task. Note that it is possible to learn a similarity matrix between two heterogeneous spaces, where $\mathbf{M}^k \in \mathbb{R}^{d \times d'}$ and $d \neq d'$. For simplicity, in this paper, we simply restrict the rest discussions by assuming $d = d'$.

For online multi-task relative similarity learning, we receive a sequence of triplet data from multiple tasks. Each triplet contains training data information of three instances, the relative label information, and the ID of the task in the triplet. More formally, at the t-th round, the received triplet is denoted as

$$\left\{(\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2; y_t; t_k) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \{-1, +1\} \times [K]\right\},$$

where $\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2$ are three received instances at the t-th iteration, $t_k \in [K]$ denotes the task's ID for which the triplet belongs to, y_t indicates the relative similarity relationship between the instance pairs $(\mathbf{x}_t, \mathbf{x}_t^1)$ and $(\mathbf{x}_t, \mathbf{x}_t^2)$. Specifically, $y_t = +1$ implies that the instance $\mathbf{x}_t$ is more similar to $\mathbf{x}_t^1$ than $\mathbf{x}_t^2$, which can be formally expressed as $S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^1) \geq S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^2)$. On the contrary, $y_t = -1$ implies that $\mathbf{x}_t$ is more similar to $\mathbf{x}_t^2$ than $\mathbf{x}_t^1$. This can be formally expressed as $S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^1) < S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^2)$. The objective of online multi-task relative similarity learning is to simultaneously learn the set of K similarity matrices $\mathbf{M}^k$ that assign higher similarity scores to similar pairs and lower similarity scores to dissimilar pairs, via an online learning approach.

3.2 Proposed Method

Formally, for any triplet denoted by $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2; y_t; t_k)$, an online learner is expected to optimize the set of similarity functions in order to ensure the following constraint:

$$y_t[S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^1) - S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^2)] \geq 0, \quad \forall t \in [T].$$

Algorithm 1 OMTRSL: The proposed algorithm for Online Multi-Task Relative Similarity Learning.

Input: Parameters $C > 0$ and $b > 0$
Initialize: $\mathbf{M}_0^k = \mathbf{I}_d, \quad \forall k \in [K]$
for $t = 0, 1, 2, \ldots, T$ **do**
  Receive $(\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2, t_k)$
  Compute $p_t = \mathbf{w}_t^\top \boldsymbol{\phi}_t$ and $\hat{y}_t = \text{sign}(p_t)$
  Query y_t and compute $\ell(\mathbf{w}_t, \boldsymbol{\phi}_t) = [0, 1 - y_t p_t]_+$
  if $\ell(\mathbf{w}_t, \boldsymbol{\phi}_t) > 0$ **then**
    $\mathbf{M}_{t+1}^k = \mathbf{M}_t^k + y_t \tau_t \mathbf{A}_{k,t_k}^{-1} \mathbf{X}_t, \quad \forall k \in [K]$
    where $\tau_t = \min\left\{C, \ell(\mathbf{w}_t, \boldsymbol{\phi}_t) / \|\boldsymbol{\phi}_t\|^2_{\mathbf{A}_{\mathbf{I}_{dd}}^{-1}}\right\}$
  else
    $\mathbf{M}_{t+1}^k = \mathbf{M}_t^k, \quad \forall k \in [K]$
  end if
end for
Output: $\mathbf{M}_{T+1}^k, \quad \forall k \in [K]$

In practice, it is impossible to satisfy the above constraint for every triplet.

For those violated cases, we introduce some loss functions to measure the loss of the k-th similarity function on the t-th triplet. Specifically, we adopt the hinge loss and define the loss functions as follows:

$$\ell\left(\mathbf{M}^{t_k}; \mathbf{z}_t\right) = \left[1 - y_t[S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^1) - S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^2)]\right]_+,$$

where $[\cdot]_+ = \max(0, \cdot)$. Note that other loss functions, e.g., logistic loss and squared loss, can also be adopted.

For simplicity, we vectorize the matrix representations by defining $\mathbf{w}_t^k = \text{vec}(\mathbf{M}_t^k) = [\mathbf{M}_{11}; \ldots; \mathbf{M}_{d1}; \ldots \mathbf{M}_{dd}]$ and $\boldsymbol{\phi}_t^{t_k} = \text{vec}(\mathbf{X}_t)$, where $\mathbf{X}_t = \mathbf{x}_t(\mathbf{x}_t^1 - \mathbf{x}_t^2)^\top$. Moreover, the whole triplet $\mathbf{z}_t$ is represented as a $(K \times d \times d)$-dimensional compound vector $\boldsymbol{\phi}_t = (0; \ldots; 0; \boldsymbol{\phi}_t^{t_k}; 0; \ldots; 0) \in \mathbb{R}^{Kdd \times 1}$. The target matrices of all the K tasks are represented as a $(K \times d \times d)$-dimensional compound vector $\mathbf{w}_t = (\mathbf{w}_t^1; \ldots; \mathbf{w}_t^k) \in \mathbb{R}^{Kdd \times 1}$. Based on the compound vector representations, the loss function can be re-written as $\ell(\mathbf{w}_t, \boldsymbol{\phi}_t) = [1 - y_t \mathbf{w}_t^\top \boldsymbol{\phi}_t]_+$.

Similar to [Cavallanti *et al.*, 2010], we adopt a task relationship matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$, which defines the learning rate to be used in the updating rules for each matrix $\mathbf{M}^k$,

$$\mathbf{A}^{-1} = \frac{1}{(1+b)K} \begin{bmatrix} b+K & b & \ldots & b \\ b & b+K & \ldots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \ldots & b+K \end{bmatrix},$$

where b is a parameter used to control to what extent the tasks would share instances with each other. In the following sections, we denote $\mathbf{A}_{i,j}$ as the entry in the i-th row and j-th column of matrix $\mathbf{A}$.

For an incoming triplet $\mathbf{z}_t$, we define the multi-task Passive-Aggressive [Crammer *et al.*, 2006] objective function for online learning:

$$\begin{aligned} &\frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2_{\mathbf{A}_{\mathbf{I}_{dd}}} + C\xi \\ &\text{s.t.} \quad 1 - y_t \mathbf{w}^\top \boldsymbol{\phi}_t \le \xi, \text{ and } \xi \ge 0, \end{aligned} \tag{2}$$

Algorithm 2 OMTRSL-Active: The proposed algorithm for Active Online Multi-Task Relative Similarity Learning.

Input: Parameters $C > 0$, $b > 0$ and $\delta > 0$
Initialize: $\mathbf{M}_0^k = \mathbf{I}_d, \quad \forall k \in [K]$
for $t = 0, 1, 2, \ldots, T$ **do**
  Receive $(\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2, t_k)$
  Compute $p_t = \mathbf{w}_t^\top \boldsymbol{\phi}_t$ and $\hat{y}_t = \text{sign}(p_t)$
  Sample $Z_t \in \{0, 1\}$ with $\Pr(Z_t = 1) = \frac{\delta}{\delta + |p_t|}$;
  if $Z_t = 1$ **then**
    Query y_t and compute $\ell(\mathbf{w}_t, \boldsymbol{\phi}_t) = [0, 1 - y_t p_t]_+$
    if $\ell(\mathbf{w}_t, \boldsymbol{\phi}_t) > 0$ **then**
      $\mathbf{M}_{t+1}^k = \mathbf{M}_t^k + y_t \tau_t \mathbf{A}_{k,t_k}^{-1} \mathbf{X}_t, \quad \forall k \in [K]$
      where $\tau_t = \min\left\{C, \ell(\mathbf{w}_t, \boldsymbol{\phi}_t) / \|\boldsymbol{\phi}_t\|^2_{\mathbf{A}_{\mathbf{I}_{dd}}^{-1}}\right\}$
    end if
  else
    $\mathbf{M}_{t+1}^k = \mathbf{M}_t^k, \quad \forall k \in [K]$
  end if
end for
Output: $\mathbf{M}_{T+1}^k, \quad \forall k \in [K]$

where $C > 0$ is a parameter used to trade-off between minimizing the adjustment of the model and minimizing the loss of the new model on current triplet. In Eq. 2, $\mathbf{A}_{\mathbf{I}_{dd}} \in \mathbb{R}^{Kdd \times Kdd}$ denotes $\mathbf{A} \otimes \mathbf{I}_{dd}$, where $\mathbf{I}_{dd}$ is a d^2 identity matrix. The objective function enjoys a closed-form updating rule:

$$\mathbf{w}_{t+1} = \mathbf{w} + y_t \tau_t \mathbf{A}_{\mathbf{I}_{dd}}^{-1} \boldsymbol{\phi}_t, \tag{3}$$

where $\tau_t = \min\left\{C, \ell(\mathbf{w}_t, \boldsymbol{\phi}_t) / \|\boldsymbol{\phi}_t\|^2_{\mathbf{A}_{\mathbf{I}_{dd}}^{-1}}\right\}$. For the k-th task, Eq. (3) is essentially equivalent to

$$\mathbf{M}_{t+1}^k = \mathbf{M}_t^k + y_t \tau_t \mathbf{A}_{k,t_k}^{-1} \mathbf{X}_t. \tag{4}$$

Algorithm 1 summarizes the details of the proposed algorithm OMTRSL.

3.3 An Active Learning Extension

In Section 3.2, the proposed OMTRSL algorithm only requires relative label information y_t that indicates which pair of instances are more similar. This greatly reduces the labeling effort compared to computing the exact similarity score between any two instances. However, the effort is still required to collect the relative label information. For many real-world applications, this may result in a huge amount of human labeling cost. Because the number of potential triplets is often much larger than the number of instances in a relative similarity learning task. To reduce the labeling cost, we propose an active learning algorithm, namely OMTRSL-Active, based on the fully-supervised OMTRSL algorithm. Specifically, we propose a simple yet effective margin-based query strategy [Cesa-Bianchi *et al.*, 2006], which has been successfully used for active learning [Cesa-Bianchi *et al.*, 2006; Hao *et al.*, 2015; 2016].

A stochastic active sampling scheme is adopted to decide whether it is necessary to query the true label y_t of an incoming triplet. This strategy attempts to draw a Bernoulli trial on a random variable $Z_t \in \{0, 1\}$, where $Z_t = 1$ indicates the

true label y_t should be queried at the t-th step and $Z_t = 0$ otherwise. In this paper, we define the sampling probability at the t-th step as follows:

$$\Pr(Z_t = 1) = \frac{\delta}{\delta + |p_t|}, \tag{5}$$

where $p_t = \mathbf{w}_t^\top \boldsymbol{\phi}_t$ indicates the difference between the similarity scores $S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^1)$ and $S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^2)$. A small value of $|p_t|$ usually means that the t-th triplet is more difficult to be predicated thus more informative to train the similarity function than a large value of $|p_t|$. Therefore, this triplet has a high probability to be queried for the true label in order to update the similarity matrix S^{t_k}. In Eq. 5, $\delta > 0$ is a smoothing parameter used to decide the amount of queries. The details of the algorithm OMTRSL-Active are shown in Algorithm 2.

3.4 Theoretical Analysis

Theorem 1 *Let $\{(\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2, y_t, k_t)|t \in [T]\}$ be a sequence of examples where $\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2 \in \mathbb{R}^d$, $y_t \in \{-1, +1\}$, $k_t \in [K]$ and $\|\mathbf{x}_t\|, \|\mathbf{x}_t^1\|, \|\mathbf{x}_t^2\| \leq R$ for all t. Then for any matrix $M \in \mathbb{R}^{d\times d}$, the number of mistakes for Online Multi-Task Relative Similarity Learning (OMTRSL) is bounded by:*

$$\sum_{t=1}^{T} m_t \leq \max\Big(\frac{4(b+K)R^4}{(1+b)K}, 1/C\Big) \times \Big[\sum_{i=1}^{K}\|M^i\|_F^2 + b\sum_{i=1}^{K}\|M^i - \frac{1}{K}\sum_{i=1}^{K}M^i\|_F^2 + C\sum_{t=1}^{T}\ell(M^{k_t}; \mathbf{z}_t)\Big]$$

where $m_t = \mathbf{I}(y_t[S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^1) - S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^2)] < 0)$.

Remark: It can be observed that, when $b = 0$, this bound corresponds to the case where all the K tasks are learnt separately; while when b is large enough, this bound corresponds to the case where all the tasks are treated as one. In practice, we can tune the parameter b to achieve a good trade-off.

Theorem 2 *Let $\{(\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2, y_t, k_t)|t \in [T]\}$ be a sequence of examples where $\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2 \in \mathbb{R}^d$, $y_t \in \{-1, +1\}$, $k_t \in [K]$ and $\|\mathbf{x}_t\|, \|\mathbf{x}_t^1\|, \|\mathbf{x}_t^2\| \leq R$ for all t. Then for any matrix $M \in \mathbb{R}^{d\times d}$, the expected number of mistakes for the OMTRSL-Active algorithm is bounded by:*

$$\mathbb{E}\Big[\sum_{t=1}^{T} m_t\Big] \leq \frac{1}{\delta}\max\left(\frac{4(b+K)R^4}{(1+b)K}, 1/C\right) \times \left[(\frac{\delta+1}{2})^2\sum_{i=1}^{K}\|M^i\|_F^2 + (\frac{\delta+1}{2})^2 b\sum_{i=1}^{K}\|M^i - \frac{1}{K}\sum_{i=1}^{K}M^i\|_F^2 + (\delta+1)C\sum_{t=1}^{T}\ell(M^{k_t}; \mathbf{z}_t)\right]$$

where $m_t = \mathbf{I}(y_t[S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^1) - S^{t_k}(\mathbf{x}_t, \mathbf{x}_t^2)] < 0)$.

Here, we omit the detailed proofs due to space limitation.

4 Experiments

In this section, we present our empirical studies on two real datatsets for evaluating both the efficacy and efficiency of the proposed algorithms.

4.1 Experimental Settings

Datasets

The performances of the proposed methods are evaluated on two real-world datasets: the *Isolet* spoken alphabet recognition dataset [Fanty and Cole, 1990] and the *news20* dataset [1]. These two datasets have been widely used for multi-task learning research. The *Isolet* dataset consists of $7,797$ examples, which are collected from 150 speakers who have uttered all characters in the English alphabet twice. On average, each speaker has contributed 52 training examples. The learning task on this dataset is to classify which letter has been uttered based on several acoustic features, including spectral coefficients, contour-, sonorant-, and post-sonorant features. On this dataset, the speakers are categorized into smaller groups, each contains 30 similar speakers. This gives rise to 5 disjoint groups of training examples called "Isolet1-5". Each group has its own classification task with 26 labels. Thus, there are 5 learning tasks on *Isolet* dataset. More details of the *Isolet* dataset can be found in [Fanty and Cole, 1990; Parameswaran and Weinberger, 2010]. The *news20* dataset is for news document classification. This dataset consists of 20 classes. In total, there are 15,935 examples for training and 3,993 examples for testing. We adopt 4 major categories (i.e., *comp*, *rec*, *sci*, and *talk*) in this dataset to form 4 learning tasks in our multi-task learning experiments.

Setup and Metrics

On both datasets, we used standard 5-fold cross validation for evaluation, in which 80% of the data are used for training, and the remaining 20% are used for testing. We report the averaged results on the 5 test sets. In each fold, the triplet data streams used in the experiments were generated based on the training set. Specifically, to generate the t-th triplet $(\mathbf{x}_t, \mathbf{x}_t^1, \mathbf{x}_t^2; y_t)$, we first randomly choose two examples $\mathbf{x}_a$ and $\mathbf{x}_b$ which belong to the same class, and another example $\mathbf{x}_c$ from another different class. We then flip a coin to decide the value of y_t. If $y_t = +1$, we assign the third example $\mathbf{x}_c$ to $\mathbf{x}_t^2$ in the triplet, and assign the rest two examples from the same class to $\mathbf{x}_t$ and $\mathbf{x}_t^1$, respectively. If $y_t = -1$, the third example $\mathbf{x}_c$ is assigned to $\mathbf{x}_t^1$, and the rest two are assigned to $\mathbf{x}_t$ and $\mathbf{x}_t^2$, respectively. Note that under the online active learning setting, y_t is *only* disclosed to the online learner upon receiving the query request by the online active learner OMTRSL-Active.

The performances of all algorithms are evaluated using precision at top k, a standard measure for ranking algorithms. For each query instance in the test set, all other test instances are ranked according to their similarities to the query instance calculated using Eq. (1). Among the top k instances, the percentage of instances from the same class with the query instance can be computed, and then averaged over all test instances. So that, we can obtain the average precision at-top-k, denoted by AP@k. Moreover, we also compute the mean Average Precision at top k, denoted by mAP@k. This measure that has been widely used in the information retrieval community. It considers the number of instances whose classes are the same with the query instance, as well as the ranking

[1] Available on the LIBSVM Machine Learning Repository.

order of the retrieved instances. For both of the AP@k and mAP@k measures, $k \in [10]$ are considered.

4.2 Comparison Schemes

We compare the proposed algorithms with the state-of-the-art multi-task learning algorithms.

- mtLMNN [Parameswaran and Weinberger, 2010]: This is the state-of-the-art multi-task similarity learning algorithm. Although it was originally proposed for classification, it implicitly outputs distance metrics for similarity learning.
- OGRSL [Chechik *et al.*, 2010]: This algorithm combines all the tasks into a single task to yield a global similarity matrix.
- OSTRSL [Chechik *et al.*, 2010]: This is a online single task relative similarity learning method, which treats each task independently.
- OMTRSL-Random: This is the random query version of OMTRSL for active learning. It randomly decides when to query the true label of an incoming triplet.
- OMTRSL: This is the proposed online multi-task relative similarity learning method (Algorithm 1).
- OMTRSL-Active: This is the proposed active variant of the OMTRSL algorithm (Algorithm 2).

For each evaluated algorithm, the optimal parameter are chosen by using cross-validation.

4.3 Evaluation on Efficacy

In this section, we evaluate the proposed OMTRSL algorithm on the test set, in terms of AP@k and mAP@k, where k varies from 1 to 10. Figure 1 and Figure 2 show the average performances over all tasks achieved by different methods on the *isolet* dataset and *news20* dataset, respectively. Based on these results, we make the following observations:

- The online global relative similarity learning algorithm OGRSL consistently performs much worse than the other algorithms. This indicates that it is necessary to learn a model for each task, rather than learn one global model for all tasks.
- The proposed online multi-task learning algorithm OMTRSL consistently outperforms the single task learning algorithm OSTRSL which learns each similarity metric independently. This is consistent with the findings in [Caruana, 1997; Parameswaran and Weinberger, 2010]. Moreover, it also confirms that the proposed OMTRSL algorithm could improve the overall performances of all tasks by sharing information between related tasks.
- The performances of all algorithms decrease when k gradually increases. This is consistent with our intuition that the task becomes harder when the list of retrieved instances becomes longer. In addition, we also notice that mtLMNN can achieve higher performance when k is small. One possible reason is that, in the learning process, for each query instance, mtLMNN tries to make sure its top k closest instances are from the same class.This learning strategy makes mtLMNN outperform other methods when $k \in [2]$. However, for larger k, the proposed OMTRSL algorithm consistently achieve better results than the mtLMNN algorithm, on both datasets.

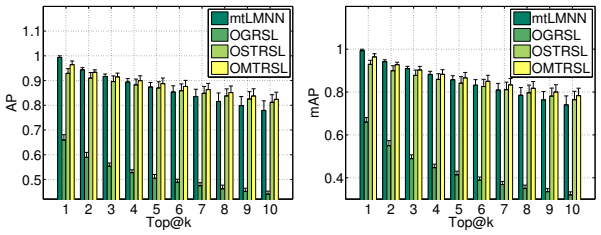


Figure 1: Performance with varied k on *isolet* datasets (left: AP, right: mAP).

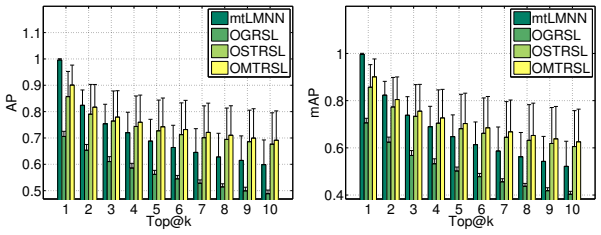


Figure 2: Performance with varied k on *news20* datasets (left: AP, right: mAP).

4.4 Evaluation on Efficiency

We also study the efficiency of the evaluation algorithms. Table 1 shows the time costs of these algorithms on both *isolet* dataset and *news20* dataset.

Table 1: Time (s) cost on training the algorithms.

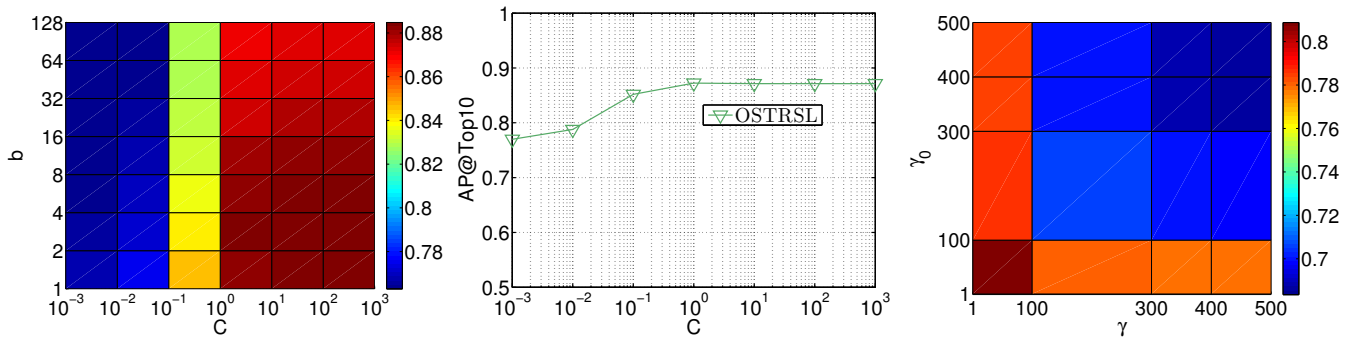
Data	mtLMNN	OGTRSL	OSTRSL	OMTRSL
isolet	241.3	12.5	12.6	14.7
news20	5371.4	111.2	114.9	126.4

From Table 1, we can observe that the offline algorithm mtLMNN takes more than 16 times learning time compared to other online algorithms on *isolet* dataset, and more than 40 times learning time on *news20* dataset due to the high dimension. In addition, it also should be noted that we implement our algorithms with pure Matlab language, and the mtLMNN algorithm provided by the author [2] is implemented by a combination of Matlab and C languages. These observations indicate that online algorithms are more advantageous, in terms of time cost. Moreover, the proposed online algorithm OMTRSL takes a little extra time compared to other two online algorithms (OGRSL, OSTRSL), due to updating several tasks simultaneously on each triplet. However, considering the efficiency of the online learning scheme, this extra time could be ignored. These observations confirm the efficiency of the proposed OMTRSL algorithm, which makes it very suitable to large-scale similarity learning problems.

4.5 Parameter Sensitivity Analysis

In this section, we study the sensitivity of the model parameters. Figures 3 (a), (b) and (c) show the sensitivity of param-

[2] http://www.cs.cornell.edu/~kilian/code/code.html



(a) OMTRSL (b) OSTRSL (c) mtLMNN

Figure 3: Performance with different parameters on validation set of *isolet* dataset.

eters in the OMTRSL, OSTRSL and mtLMNN algorithms, respectively. From Figure 3 (a), we can observe that the performance of the OMTRSL algorithm is not sensitive to the settings of parameter b. However, we can also observe that the performance decreases as b increases. This is consistent with our theoretical analysis in Section (3.3). Moreover, the OMTRSL algorithm is a little sensitive to the choice of parameter C. An empirical setting of C is 1. In Figure 3 (b), we can make similar observations as in Figure 3 (a). A too small value of C can make the performance of the OSTRSL algorithm decrease. In addition, Figure 3 (c) indicates that a large value for any of γ and γ_0 would make the performance of the mtLMNN algorithm decrease. This is consistent with the findings in [Parameswaran and Weinberger, 2010].

4.6 Experiments of Active Learning

In the experiments, we also evaluate the performance of the proposed active learning algorithm OMTRSL-Active. Figure 4 shows the results on *isolet* dataset. From Figure 4, we can notice that the proposed OMTRSL-Active method consistently outperforms the random version OMTRSL method, i.e., *OMTRSL-Random*. The active learning approach achieves more than 5% better performance than the random approach with less than 20% query ratio. More impressively, using around 40% of training data, the active learning approach (i.e., OMTRSL-Active) can achieve highly comparable performance with the OSTRSL and mtLMNN algorithms, which use all of the training data. These results demonstrate the effectiveness of the proposed active learning approach in reducing labeling cost. Similar observations can be made on news20 dataset, and we omit the results due to space limitation.

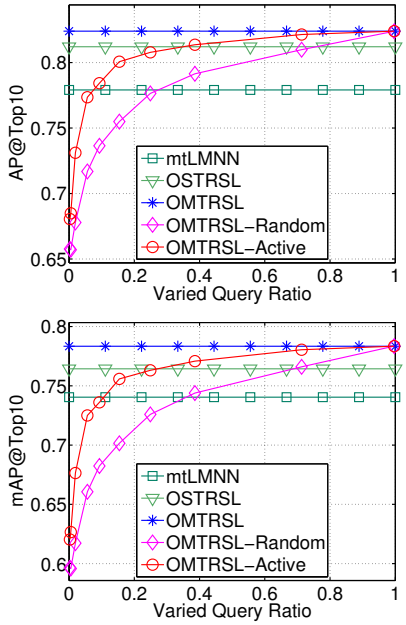


Figure 4: Performance of varied query ratios on *isolet* dataset.

5 Conclusion

This paper investigates online multi-task learning techniques for relative similarity learning tasks. In particular, we propose a novel Online Multi-Task Relative Similarity Learning algorithm (OMTRSL), which overcomes the drawbacks of existing approaches that are often of low efficacy and inefficiency. To further reduce the human labeling effort, we develop an active variant of OMTRSL, namely OMTRSL-Active, to avoid labeling of each incoming triplet. We theoretically analyze the mistake bounds of both OMTRSL and OMTRSL-Active algorithms, and empirically conduct extensive experiments on real datasets. We have found very encouraging results by comparing with the state-of-the-art algorithms. For future work, we would like to sparse multi-task learning for similarity problems [Yao *et al.*, 2015] and design adaptive relationship matrix method for online multi-task RSL.

Acknowledgments

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative and the International Research Centres in Singapore Funding Initiative. This research is also partially supported by the NTU-PKU Joint Research Institute, a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation.

References

[Ammar *et al.*, 2014] Haitham Bou Ammar, Eric Eaton, Paul Ruvolo, and Matthew Taylor. Online multi-task learning for policy gradient methods. In *2014 ICML*, pages 1206–1214, 2014.

[Bakker and Heskes, 2003] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *JMLR*, 4:83–99, 2003.

[Block, 1962] HD Block. The perceptron: A model for brain functioning. i. *Reviews of Modern Physics*, 34(1), 1962.

[Calandriello *et al.*, 2014] Daniele Calandriello, Alessandro Lazaric, and Marcello Restelli. Sparse multi-task reinforcement learning. In *NIPS*, pages 819–827, 2014.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[Cavallanti *et al.*, 2010] Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *JMLR*, 11:2901–2934, 2010.

[Cesa-Bianchi *et al.*, 2006] Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. Worst-case analysis of selective sampling for linear classification. *JMLR*, 7:1205–1230, December 2006.

[Chechik *et al.*, 2010] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11:1109–1135, 2010.

[Cohen and Crammer, 2014] Haim Cohen and Koby Crammer. Learning multiple tasks in parallel with a shared annotator. In *NIPS*, pages 1170–1178, 2014.

[Crammer and Chechik, 2012] Koby Crammer and Gal Chechik. Adaptive regularization for similarity measures. In *ICML*, 2012.

[Crammer *et al.*, 2006] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585, 2006.

[Eaton and Ruvolo, 2013] Eric Eaton and Paul L Ruvolo. Ella: An efficient lifelong learning algorithm. In *ICML*, pages 507–515, 2013.

[Fang and Rockmore, 2015] Chen Fang and Daniel N Rockmore. Multi-task metric learning on network data. In *Advances in Knowledge Discovery and Data Mining*, pages 317–329. Springer, 2015.

[Fanty and Cole, 1990] Mark Fanty and Ronald Cole. Spoken letter recognition. In *NIPS*, 1990.

[Gonçalves *et al.*, 2016] André R Gonçalves, Fernando J Von Zuben, and Arindam Banerjee. Multi-task sparse structure learning with gaussian copula models. *Journal of Machine Learning Research*, 17(33):1–30, 2016.

[Hao *et al.*, 2015] Shuji Hao, Peilin Zhao, Steven C. H. Hoi, and Chunyan Miao. Learning relative similarity from data streams: Active online learning approaches. In *CIKM*, pages 1181–1190. ACM, 2015.

[Hao *et al.*, 2016] Shuji Hao, Peilin Zhao, Jing Lu, Steven C. H. Hoi, Chunyan Miao, and Chi Zhang. Soal: Second-order online active learning. In *2016 ICDM*, pages 931–936, 2016.

[Hoi *et al.*, 2014] Steven CH Hoi, Jialei Wang, and Peilin Zhao. Libol: A library for online learning algorithms. *The Journal of Machine Learning Research*, 15(1):495–499, 2014.

[Lugosi *et al.*, 2009] Gábor Lugosi, Omiros Papaspiliopoulos, and Gilles Stoltz. Online multi-task learning with hard constraints. *arXiv preprint arXiv:0902.3526*, 2009.

[Ma *et al.*, 2014] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *Image Processing, IEEE Transactions on*, 23(8):3656–3670, 2014.

[Parameswaran and Weinberger, 2010] Shibin Parameswaran and Kilian Q Weinberger. Large margin multi-task metric learning. In *NIPS*, pages 1867–1875, 2010.

[Ruvolo and Eaton, 2014] Paul Ruvolo and Eric Eaton. Online multi-task learning via sparse dictionary optimization. In Carla E. Brodley and Peter Stone, editors, *AAAI*, pages 2062–2068. AAAI Press, 2014.

[Saha *et al.*, 2011] Avishek Saha, Piyush Rai, Hal Daumé III, and Suresh Venkatasubramanian. Online learning of multiple tasks and their relationships. In *AISTATS*, volume 15 of *JMLR Proceedings*, pages 643–651. JMLR.org, 2011.

[Schultz and Joachims, 2004] Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. *NIPS*, page 41, 2004.

[Shalev-Shwartz *et al.*, 2004] Shai Shalev-Shwartz, Yoram Singer, and Andrew Y Ng. Online and batch learning of pseudo-metrics. In *ICML*, page 94. ACM, 2004.

[Weinberger and Saul, 2009] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009.

[Yang and Jin, 2006] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2, 2006.

[Yang *et al.*, 2013] Peipei Yang, Kaizhu Huang, and Cheng-Lin Liu. Geometry preserving multi-task metric learning. *Machine learning*, 92(1):133–175, 2013.

[Yao *et al.*, 2015] Dezhong Yao, Peilin Zhao, Chen Yu, Hai Jin, and Bin Li. Sparse online relative similarity learning. In *2015 ICDM*, pages 529–538, 2015.

[Zhang and Yeung, 2010] Yu Zhang and Dit-Yan Yeung. Transfer metric learning by learning task relationships. In *SIGKDD*, pages 1199–1208. ACM, 2010.

[Zhang *et al.*, 2016] Chi Zhang, Peilin Zhao, Shuji Hao, Yeng Chai Soh, and Bu-Sung Lee. Rom: A robust online multi-task learning approach. In *2016 ICDM*, pages 1341–1346, 2016.