MULTI TASK LEARNING WITH POSITIVE AND UNLABELED DATA AND ITS APPLICATION TO MENTAL STATE PREDICTION

Hirotaka Kaji[1]*, Hayato Yamaguchi*[1]*, and Masashi Sugiyama*[2,3]

[1]Frontier Research Center, Toyota Motor Corp., Shizuoka, Japan
[2]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan
[3]Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan

ABSTRACT

In real-world machine learning applications, we are often faced with a situation where only a small number of training samples is available due to high sampling costs. For instance, prediction of mental states such as drowsiness from physiological information is a typical example. To cope with this problem, classifier training methods only from positive and unlabeled data and multi-task learning methods for improving the classification performance by solving multiple related tasks simultaneously have been actively investigated recently. In this paper, we combine these methods and propose a multi-task learning method that can handle positive-unlabeled tasks and positive-negative tasks in a unified manner. Through experiments on drivers' drowsiness prediction, we demonstrate the effectiveness of the proposed method.

Index Terms— Multi-task learning, positive and unlabeled learning, mental state prediction

1. INTRODUCTION

In real-world machine learning applications, we are often faced with a situation where only a small number of labeled training samples is available due to high sampling costs.

An example of such a small sample situation that we will consider throughout the paper is prediction of a person's mental states such as drowsiness and stress from physiological information such as heart beats. Labeling of drowsiness is usually carried out by subjective measures such as the *Karolinska sleepiness scale*[1], experts' facial expression scoring, or measuring the response time of executing some task, which involve time-consuming manual annotation processes.

To cope with such small sample problems, various approaches have been explored so far. *Positive-unlabeled learning* (PU learning) allows us to train a binary classifier only from positive and unlabeled data [2, 3, 4]. PU learning is effective when negative samples are expensive to collect, while positive and unlabeled samples are easy to collect. Drowsiness prediction matches well with this situation since negative samples (being drowsy) are expensive to collect, while positive samples (non-drowsy) and unlabeled samples can be obtained abundantly, thanks to recent advances in wearable sensors. Another popular approach to compensating for the small number of training samples is *multi-task learning* [5, 6, 7], which solves multiple learning tasks simultaneously by sharing information among related tasks. Drowsiness prediction is suited also to multi-task learning because drowsiness classifiers may have variations depending on subjects and thus drowsiness classification problems for multiple subjects can be naturally formulated as multi-task learning [8, 9].

In this paper, we combine these two approaches and propose a novel multi-task learning method that can handle positive-unlabeled tasks (PU tasks) and positive-negative tasks (PN tasks) in a unified manner.

[1]The *Karolinska sleepiness scale* is a 9-point Likert scale rated from "very alert (1)" to "fighting sleepiness (9)" [1].

2. PROBLEM FORMULATION

In this section, we formulate a binary classification problem and review the frameworks of positive and unlabeled learning and multi-task learning.

Positive and Negative Learning: Let $\boldsymbol{x} \in \mathbb{R}^d$ be a pattern and $y \in \{+1, -1\}$ be its class label, which are regarded as random variables equipped with unknown joint probability density $p(\boldsymbol{x}, y)$. In a standard binary classification problem, we are given independent and identically distributed training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from the joint probability density $p(\boldsymbol{x}, y)$. The goal of binary classification is to, from the training samples, obtain classifier $g : \mathbb{R}^d \to \mathbb{R}$ that classifies a test pattern $\boldsymbol{x}$ to its true class y by $\mathrm{sign}(g(\boldsymbol{x}))$.

More precisely, we want to obtain classifier g that minimizes the *risk* defined as $R(g) = \mathbb{E}[\ell(yg(\boldsymbol{x}))]$, where $\mathbb{E}$ denotes the expectation over $p(\boldsymbol{x}, y)$ and $\ell(yg(\boldsymbol{x}))$ is a *loss* when y is predicted as $g(\boldsymbol{x})$. Typically, the squared loss $\ell(z) = (1 - z)^2$ is used. Approximating the expectation over unknown $p(\boldsymbol{x}, y)$ with the average over training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, a classifier is trained as $\min_g \widehat{R}(g)$, where $\widehat{R}(g) = \frac{1}{n}\sum_{i=1}^n \ell(y_i g(\boldsymbol{x}_i))$. Since positive and negative samples are used for classifier training, we refer to this problem setting as *positive and negative (PN) learning*.

Positive and Unlabeled Learning: When negative samples are not obtained easily while positive and unlabeled samples can be obtained abundantly, *positive and unlabeled (PU) learning* is a useful approach to binary classification [2, 3, 4].

Suppose we are given positive and unlabeled datasets:

$$\mathcal{X}^{\mathrm{p}} := \{\boldsymbol{x}_i^{\mathrm{p}}\}_{i=1}^{n_{\mathrm{p}}} \overset{\text{i.i.d.}}{\sim} p^{\mathrm{p}}(\boldsymbol{x}|y=+1),$$
$$\mathcal{X}^{\mathrm{u}} := \{\boldsymbol{x}_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u}}} \overset{\text{i.i.d.}}{\sim} p^{\mathrm{u}}(\boldsymbol{x}),$$

where $p(\boldsymbol{x}|y)$ is class-conditional density, $p(\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=+1) + (1-\pi)p(\boldsymbol{x}|y=-1)$ is the marginal density, and $\pi = p(y=+1)$ is the class-prior probability for the positive class.

The risk R can be decomposed as

$$R(g) = \pi \mathbb{E}_{\mathrm{p}}[\ell(g(\boldsymbol{x}))] + (1-\pi)\mathbb{E}_{\mathrm{n}}[\ell(-g(\boldsymbol{x}))], \quad (1)$$

where $\mathbb{E}_{\mathrm{p}}[\cdot]$ and $\mathbb{E}_{\mathrm{n}}[\cdot]$ denote the expectations over positive and negative class-conditional distributions, respectively. Since negative samples are not available in the PU learning setup, let us use the following relation induced from the definition of marginal density:

$$\mathbb{E}_{\mathrm{u}}[\ell(-g(\boldsymbol{x}))] = \pi \mathbb{E}_{\mathrm{p}}[\ell(-g(\boldsymbol{x}))] + (1-\pi)\mathbb{E}_{\mathrm{n}}[\ell(-g(\boldsymbol{x}))],$$

where $\mathbb{E}_{\mathrm{u}}$ denotes the expectation over unlabeled samples. We then obtain the following risk expression for PU classification by eliminating $(1-\pi)\mathbb{E}_{\mathrm{n}}[\ell(-g(\boldsymbol{x}))]$ in Eq.(1) [4]:

$$R(g) = \pi \mathbb{E}_{\mathrm{p}}\left[\tilde{\ell}(g(\boldsymbol{x}))\right] + \mathbb{E}_{\mathrm{u}}[\ell(-g(\boldsymbol{x}))],$$

where $\tilde{\ell}(z) = \ell(z) - \ell(-z)$ is called the *composite loss*. In practice, the expectations over positive and unlabeled samples are replaced with corresponding sample averages.

du Plessis et al. [4] showed that, for convex $\ell(z)$, $\tilde{\ell}(z)$ is linear if and only if it is convex. For example, the squared loss, the logistic loss, and the double hinge loss yield a linear composite loss. Without loss of generality, let $\tilde{\ell}(z) = -z$. Then the risk for PU classification is given as

$$R(g) = \pi \mathbb{E}_{\mathrm{p}}[-g(\boldsymbol{x})] + \mathbb{E}_{\mathrm{u}}[\ell(-g(\boldsymbol{x}))].$$

Multi-Task Learning: Suppose that we have T binary classification tasks and training samples $\{(\boldsymbol{x}_i, y_i, t_i)\}_{i=1}^N$, where $t_i \in \{1, \ldots, T\}$ denotes the task index. The idea of *multi-task learning* (MTL) is to share information across related tasks by solving all tasks simultaneously [5]. Among various MTL formulations, *regularized MTL* is one of the most practical approaches [6, 7], which imposes solutions of two different tasks to be close when these tasks are similar:

$$J(g) = \sum_{i=1}^N \ell(g_t(\boldsymbol{x}_i, \boldsymbol{\alpha}_{t_i}), y_i) + \sum_{t=1}^T r_t(\boldsymbol{\alpha}_t) + \sum_{t,t'=1}^T s_{t,t'}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t'}),$$

where $\boldsymbol{\alpha}_t$ is a model parameter vector for the t-th task, r is a regularization function and s is an information sharing function. Typically, the ℓ_2-norm is used for r and s.

3. MULTI TASK LEARNING WITH POSITIVE AND UNLABELED DATA

In this section, we introduce our novel method called *multi-task learning with positive and unlabeled data* (PU-MTL), where tasks with positive and negative samples (PN tasks) and tasks with positive and unlabeled samples (PU tasks) are treated in a unified manner in the multi-task learning framework.

Formulation: Consider a multi-task learning problem which includes k PN tasks and $T-k$ PU tasks. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a pattern, $y \in \{+1, -1\}$ be its class label, and $t \in \{1, \ldots, T\}$ be the task index. We assume that, for the t'-th PN task, we are given a positive dataset $\mathcal{X}_{t'}^{\tilde{\mathrm{p}}}$ and a negative dataset $\mathcal{X}_{t'}^{\mathrm{n}}$:

$$\mathcal{X}_{t'}^{\tilde{\mathrm{p}}} := \left\{\boldsymbol{x}_i^{\tilde{\mathrm{p}}}\right\}_{i=1}^{m_{\mathrm{p},t'}} \sim p_{t'}^{\tilde{\mathrm{p}}}(\boldsymbol{x}|y=+1),$$
$$\mathcal{X}_{t'}^{\mathrm{n}} := \{\boldsymbol{x}_i^{\mathrm{n}}\}_{i=1}^{m_{\mathrm{n},t'}} \sim p_{t'}^{\mathrm{n}}(\boldsymbol{x}|y=-1),$$

where $m_{\mathrm{p},t'}$ and $m_{\mathrm{n},t'}$ denote the number of positive and negative samples of the t'-th PN task. Similarly, we assume that, for the t-th PU task, we are given a positive dataset $\mathcal{X}_t^{\mathrm{p}}$ and an unlabeled dataset $\mathcal{X}_t^{\mathrm{u}}$:

$$\mathcal{X}_t^{\mathrm{p}} := \{\boldsymbol{x}_i^{\mathrm{p}}\}_{i=1}^{n_{\mathrm{p},t}} \sim p_t^{\mathrm{p}}(\boldsymbol{x}|y=+1),$$
$$\mathcal{X}_t^{\mathrm{u}} := \{\boldsymbol{x}_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u},t}} \sim p_t^{\mathrm{u}}(\boldsymbol{x}),$$

where $n_{\mathrm{p},t}$ and $n_{\mathrm{u},t}$ denote the number of positive and unlabeled samples of the t-th PU task. Let $L = M + N$, $M = M_{\mathrm{p}} + M_{\mathrm{n}}$, $N = N_{\mathrm{p}} + N_{\mathrm{u}}$, $M_{\mathrm{p}} = m_{\mathrm{p},1} + \cdots + m_{\mathrm{p},k}$ be the total number of positive samples in PN tasks, $M_{\mathrm{n}} = m_{\mathrm{n},1} + \cdots + m_{\mathrm{n},k}$ be the total number of negative samples in PN tasks, $N_{\mathrm{p}} = n_{\mathrm{p},T-k} + \cdots + n_{\mathrm{p},T}$ be the total number of positive samples in PU tasks, and $N_{\mathrm{u}} = n_{\mathrm{u},T-k} + \cdots + n_{\mathrm{u},T}$ be the total number of unlabeled samples in PU tasks. Then the PU-MTL criterion we propose is given by

$$\widehat{J}(\boldsymbol{\alpha}) = \frac{N}{L}\widehat{J^{\mathrm{pu}}}(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T) + \frac{M}{L}\widehat{J^{\mathrm{pn}}}(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T) + \frac{1}{2}\sum_{t=1}^T \lambda_t \boldsymbol{\alpha}_t^\top \boldsymbol{\alpha}_t + \frac{w}{4}\sum_{t,t'}^T \gamma_{t,t'}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t'})^\top(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t'}),$$

where $\widehat{J^{\mathrm{pu}}}$ denotes the PU learning criterion, $\widehat{J^{\mathrm{pn}}}$ denotes the PN learning criterion, $\lambda_t \geq 0$ is the regularized parameter of the t-th task, $w \geq 0$ is a parameter to control the magnitude of information sharing, and $\gamma_{t,t'} \in [0, 1]$ is the similarity between the t-th and t'-th tasks.

Efficient Implementation: For the t-th classification task, let us employ a linear-in-parameter model given by $g_t(\boldsymbol{x}) = \boldsymbol{\alpha}_t^\top \boldsymbol{\varphi}_t(\boldsymbol{x})$, where $\boldsymbol{\alpha}_t = (\alpha_{t,1}, \ldots, \alpha_{t,b})^\top$ and $\boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \ldots, \varphi_b(\boldsymbol{x}))^\top$ are a parameter vector and a basis function vector, respectively. As the basis function, we employ the Gaussian kernel $\varphi_\ell(\boldsymbol{x}) = \exp\left(\frac{-||\boldsymbol{x}-\boldsymbol{c}_\ell||^2}{2h^2}\right)$, where $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_L\} =$

Algorithm 1 Similarity estimation

Initialize similarity $\gamma_{t,t'} = \gamma \geq 0,\ \forall t, t'$
repeat
  Estimate $\boldsymbol{\alpha}$ using current $\gamma_{t,t'}$
  Update $\gamma_{t,t'}$ as $\gamma_{t,t'} = \exp\left(-\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t'}\|^2\right)$.
until $\boldsymbol{\alpha}$ converges

$\{\boldsymbol{x}^{\tilde{\mathrm{p}}}_1, \ldots, \boldsymbol{x}^{\tilde{\mathrm{p}}}_{M_\mathrm{p}}, \boldsymbol{x}^{\mathrm{n}}_1, \ldots, \boldsymbol{x}^{\mathrm{n}}_{M_\mathrm{n}}, \boldsymbol{x}^{\mathrm{p}}_1, \ldots, \boldsymbol{x}^{\mathrm{p}}_{N_\mathrm{p}}, \boldsymbol{x}^{\mathrm{u}}_1, \ldots, \boldsymbol{x}^{\mathrm{u}}_{N_\mathrm{u}}\}$ are the Gaussian centers and $h > 0$ is the Gaussian bandwidth.

Below, we consider the squared loss function $\ell(z) = \frac{1}{4}(z-1)^2$. Then $\widehat{J^{\mathrm{pu}}}$ and $\widehat{J^{\mathrm{pn}}}$ included in PU-MTL criterion can be expressed with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_T^\top)^\top \in \mathbb{R}^{bT}$ as

$$\widehat{J^{\mathrm{pu}}}(\boldsymbol{\alpha}) = \frac{1}{4N_\mathrm{u}}\boldsymbol{\alpha}^\top \boldsymbol{\Psi}_\mathrm{u}^\top \boldsymbol{\Psi}_\mathrm{u}\boldsymbol{\alpha} + \frac{1}{2N_\mathrm{u}}\mathbf{1}^\top \boldsymbol{\Psi}_\mathrm{u}\boldsymbol{\alpha} - \frac{\pi}{N_p}\mathbf{1}^\top \boldsymbol{\Psi}_\mathrm{p}\boldsymbol{\alpha},$$

$$\widehat{J^{\mathrm{pn}}}(\boldsymbol{\alpha}) = \frac{1}{4M}\boldsymbol{\alpha}^\top \boldsymbol{\Psi}_{\tilde{\mathrm{p}}}^\top \boldsymbol{\Psi}_{\tilde{\mathrm{p}}}\boldsymbol{\alpha} + \frac{1}{4M}\boldsymbol{\alpha}^\top \boldsymbol{\Psi}_\mathrm{n}^\top \boldsymbol{\Psi}_\mathrm{n}\boldsymbol{\alpha} + \frac{1}{2M}\mathbf{1}^\top \boldsymbol{\Psi}_\mathrm{n}\boldsymbol{\alpha} - \frac{1}{2M}\mathbf{1}^\top \boldsymbol{\Psi}_{\tilde{\mathrm{p}}}\boldsymbol{\alpha},$$

where

$$\boldsymbol{\Psi}_\mathrm{p} = \left(\boldsymbol{\psi}_{t_1}(\boldsymbol{x}_1), \ldots, \boldsymbol{\psi}_{t_{N_\mathrm{p}}}(\boldsymbol{x}^{\mathrm{p}}_{N_\mathrm{p}})\right)^\top \in \mathbb{R}^{N_\mathrm{p} \times bT},$$
$$\boldsymbol{\Psi}_\mathrm{u} = \left(\boldsymbol{\psi}_{t_1}(\boldsymbol{x}_1), \ldots, \boldsymbol{\psi}_{t_{N_\mathrm{u}}}(\boldsymbol{x}^{\mathrm{u}}_{N_\mathrm{u}})\right)^\top \in \mathbb{R}^{N_\mathrm{u} \times bT},$$
$$\boldsymbol{\Psi}_{\tilde{\mathrm{p}}} = \left(\boldsymbol{\psi}_{t_1}(\boldsymbol{x}'_1), \ldots, \boldsymbol{\psi}_{t_{M_\mathrm{p}}}(\boldsymbol{x}'^{\mathrm{p}}_{M_\mathrm{p}})\right)^\top \in \mathbb{R}^{M_\mathrm{p} \times bT},$$
$$\boldsymbol{\Psi}_\mathrm{n} = \left(\boldsymbol{\psi}_{t_1}(\boldsymbol{x}'_1), \ldots, \boldsymbol{\psi}_{t_{M_\mathrm{n}}}(\boldsymbol{x}'^{\mathrm{n}}_{M_\mathrm{n}})\right)^\top \in \mathbb{R}^{M_\mathrm{n} \times bT},$$
$$\boldsymbol{\psi}_t(\boldsymbol{x}) = \left(\mathbf{0}^\top_{b(t-1)}, \boldsymbol{\varphi}_t(\boldsymbol{x})^\top, \mathbf{0}^\top_{b(T-t)}\right)^\top \in \mathbb{R}^{bT},$$

and $\mathbf{0}_b$ is the b-dimensional vector with all zeros. Then the minimizer $\widehat{\boldsymbol{\alpha}}$ of the learning criterion $\widehat{J}(\boldsymbol{\alpha})$ can be analytically computed as

$$\widehat{\boldsymbol{\alpha}} = \left(\frac{N}{L}\frac{1}{2N_\mathrm{u}}\boldsymbol{\Psi}_\mathrm{u}^\top \boldsymbol{\Psi}_\mathrm{u} + \frac{1}{2L}\boldsymbol{\Psi}_{\tilde{\mathrm{p}}}^\top \boldsymbol{\Psi}_{\tilde{\mathrm{p}}} + \frac{1}{2L}\boldsymbol{\Psi}_\mathrm{n}^\top \boldsymbol{\Psi}_\mathrm{n} + \boldsymbol{C} \otimes \boldsymbol{I}_b\right)^{-1} \left(\frac{N}{L}\frac{\pi}{N_p}\boldsymbol{\Psi}_\mathrm{p}^\top \mathbf{1} - \frac{N}{L}\frac{1}{2N_\mathrm{u}}\boldsymbol{\Psi}_\mathrm{u}^\top \mathbf{1} + \frac{1}{2L}\boldsymbol{\Psi}_{\tilde{\mathrm{p}}}^\top \mathbf{1} - \frac{1}{2L}\boldsymbol{\Psi}_\mathrm{n}^\top \mathbf{1}\right),$$

where $\boldsymbol{C}$ is the $T \times T$ matrix with the (t, t') element given by

$$C_{t,t'} = \begin{cases} \lambda_t + w\sum_{t''=1}^T \gamma_{t,t''} - w\gamma_{t,t} & (t = t'), \\ -w\gamma_{t,t'} & (t \neq t'), \end{cases}$$

and $\otimes$ denotes the Kronecker product. Thanks to the analytic-form solution, the minimizer of the PU-MTL criterion for the squared loss function can be computed efficiently.

In our PU-MTL implementation, we estimated the model parameters and the task-task similarity $\gamma_{t,t'}$ alternately, as described in Algorithm 1.

4. REAL-WORLD DROWSINESS PREDICTION

In this section, we demonstrate the effectiveness of PU-MTL through experiments on drivers' drowsiness prediction.

Background: Drivers' drowsiness prediction is an important research topic for traffic accident prevention. We focus on drowsiness prediction from heart beat information, because recent advances in wearable technologies allow us to measure heart beat information on a daily basis. The cardiovascular system reflects the activity of the autonomic nervous system [10], which is deeply related to mental states such as stress [11] and drowsiness [12]. Since the transition of drowsiness during driving is in the course of nature, manual annotation is necessary to collect drowsy samples. On the other hand, awake samples can be obtained easily if drivers are assumed to be in an arousal state in the early stage of driving. In the following, we consider a situation where positive and negative samples can be collected from several subjects in laboratory experiments, and positive and unlabeled samples are collected from a specific driver.

Data Collection: Three healthy males in their 20's to 40's participated in this experiment[2]. The subjects drove our driving simulator along an expressway at around 100 km/h with overtaking other cars which run at 80 km/h, until an expert observed the subjects' strong drowsiness or the subjects completed the whole driving task (about 150 km distance). Each subject performed the experiment 10 times. We adopted AP108 (TEAC Corp.) as a wireless physiological signal amplifier for data acquisition to measure an electrocardiogram (ECG) during driving. The disposable electrodes were attached on the subject's chest. The sampling frequency was 500 Hz. At the same time, the subject's face was recorded by a video camera mounted on the cabinet of the driving simulator. This movie was used for detailed drowsiness scoring by experts after the experiment.

Data Processing: To avoid the influence on driving, we employed a sleepiness level based on facial expressions [13], which is commonly used for driver evaluations [14], as the ground truth of drowsiness. Experts checked the movie of subjects' face during driving and rated the sleepiness level from 1 ("Not sleepy") to 5 ("Very sleepy") every 60 seconds. To convert the target problem into a binary classification problem, we defined the drowsiness scores from 1 to 2 as the positive class ("awake") and from 3 to 5 as the negative class ("drowsy"). In this experiment, only five trials were annotated and remaining five trials were completely unlabeled.

We processed ECG to extract the following seven features as input vector $\boldsymbol{x}$: **LF**: The spectral power of the low frequency (LF) component (0.04-0.15 Hz), **HF**: The spectral power of the high frequency (HF) component (0.15-0.4 Hz),

[2]The design of experiments was approved and conducted according to the "Ethical guidelines for research involving with human subject" of Toyota Motor Corporation. We sufficiently explained the details before the experiment and obtained informed consents from all the subjects.

Pnn50: The number of the adjacent RRIs' differences which exceed 50 ms, **RMSSD**: The root mean square of the adjacent RRIs' differences, **SD/RMSSD**: The standard deviation of RRI divided by RMSSD, **RRV**: The variance of RRI, **The number of Peak**: The number of peaks of RRI time series. Here, the RR-interval (RRI) which is the peak-to-peak interval of R-wave of ECG was calculated at first. We then computed these features at 60 seconds intervals with 120 second sliding windows. Finally, these features were normalized to have zero mean and unit standard deviation for each subject.

Evaluation Method and Result: We consider a scenario that positive and unlabeled samples can be obtained from a new driver, and positive and negative samples were collected from two subjects beforehand. In this scenario, the goal is to provide a good drowsiness predictor to the new driver. Therefore, we evaluate the prediction performance of the PU task for the new driver. The following methods are compared: **PU-MTL**: The proposed method, **PUL**: Ordinary PU learning only from a new driver's PU samples, **PNL**: Ordinary PN learning with labeled samples obtained from the two subjects. The squared loss was used for all the methods as the loss function. We set the number of PN training samples to $(m_{\mathrm{p}}, m_{\mathrm{n}}) = (5,5), (20,20)$ for PU-MTL and PNU, and the number of PU training samples to $(n_{\mathrm{p}}, n_{\mathrm{u}}) = (10,10), (10,30), (10,100), (10,200)$ for PU-MTL and PUL. All combinations of these settings were investigated.

Here, we randomly divided the samples into the training, validation and test sets. We chose positive and negative samples from the annotated five trials and unlabeled samples from the non-annotated five trials. We assumed that the class prior of unlabeled samples was equal to that of labeled samples and was known at training time[3]. Hyperparameters λ, h, w were selected by using a validation set in terms of the misclassification rate (i.e., the zero-one loss) over all PU and PN tasks. The validation samples are composed of 10 positive samples and 30 unlabeled samples. Moreover, we chose 60 positive samples and 60 negative samples from the annotated five trials of the PU task as test samples. This procedure was repeated 100 times with different random seeds.

Tables 1 and 2 show the mean and standard deviation of the misclassification rate of subject t (which is the PU task) for different sample size. The bold face indicates the best and comparable methods according to paired t-test at the significance level 0.05. When the PN training sample size was $(5,5)$, PU-MTL and PUL tended to outperform PNL, implying that the PN sample size was not sufficient for PNL and PU-MTL used PU samples more effectively. As the number of unlabeled samples was increased, the misclassification rate of PU-MTL and PUL was gradually improved. On the other hand, when the PN training sample size was increased to $(20,20)$, the performance of PUL was worse than that of PU-MTL and PNL. With this condition, information sharing worked well since PN tasks had enough training samples. Overall, we experimentally found that PU-MTL is effective in drowsiness prediction.

[3]Practically, class prior estimation methods from PU samples such as [15] may be used.

Table 1. Comparison of the mean and standard deviation of the misclassification rate over 100 trials. The PN sample size is $(m_{\mathrm{p}}, m_{\mathrm{n}}) = (5,5)$.

$(n_{\mathrm{p}}, n_{\mathrm{u}})$	PU task	PU-MTL	PUL	PNL
(10,10)	Subject 1	**0.440**(0.077)	0.460(0.076)	0.470(0.075)
	Subject 2	**0.270**(0.051)	**0.274**(0.046)	0.493(0.093)
	Subject 3	**0.408**(0.064)	**0.415**(0.066)	**0.411**(0.084)
(10,30)	Subject 1	**0.436**(0.073)	**0.442**(0.069)	0.470(0.075)
	Subject 2	**0.260**(0.044)	**0.262**(0.045)	0.493(0.093)
	Subject 3	**0.392**(0.053)	**0.396**(0.055)	**0.411**(0.084)
(10,100)	Subject 1	**0.441**(0.074)	**0.442**(0.075)	0.470(0.075)
	Subject 2	**0.254**(0.044)	**0.256**(0.044)	0.493(0.093)
	Subject 3	**0.385**(0.056)	**0.391**(0.050)	0.411(0.084)
(10,200)	Subject 1	**0.432**(0.080)	**0.433**(0.082)	0.470(0.075)
	Subject 2	**0.257**(0.046)	**0.254**(0.045)	0.493(0.093)
	Subject 3	**0.387**(0.053)	**0.384**(0.054)	0.411(0.084)

Table 2. Comparison of the mean and standard deviation of the misclassification rate over 100 trials. The PN sample size is $(m_{\mathrm{p}}, m_{\mathrm{n}}) = (20,20)$.

$(n_{\mathrm{p}}, n_{\mathrm{u}})$	PU task	PU-MTL	PUL	PNL
(10,10)	Subject 1	**0.432**(0.070)	0.460(0.076)	**0.447**(0.066)
	Subject 2	**0.267**(0.050)	**0.274**(0.046)	0.489(0.075)
	Subject 3	0.409(0.064)	0.415(0.066)	**0.365**(0.063)
(10,30)	Subject 1	**0.424**(0.069)	0.442(0.069)	0.447(0.066)
	Subject 2	**0.257**(0.040)	**0.262**(0.045)	0.489(0.075)
	Subject 3	0.399(0.063)	0.396(0.055)	**0.365**(0.063)
(10,100)	Subject 1	**0.434**(0.074)	**0.442**(0.075)	**0.447**(0.066)
	Subject 2	**0.256**(0.043)	**0.256**(0.044)	0.489(0.075)
	Subject 3	0.385(0.052)	0.391(0.050)	**0.365**(0.063)
(10,200)	Subject 1	**0.439**(0.082)	**0.433**(0.082)	**0.447**(0.066)
	Subject 2	**0.259**(0.046)	**0.254**(0.045)	0.489(0.075)
	Subject 3	0.384(0.054)	0.384(0.054)	**0.365**(0.063)

5. CONCLUSION

In this paper, we proposed a novel method called multi-task learning with positive and unlabeled data (PU-MTL) to handle small sample problems emerging in real-world machine learning applications. Essentially, PU-MTL treats ordinary positive-negative tasks and positive-unlabeled tasks in a unified manner in a multi-task learning framework. We also provided an efficient implementation of PU-MTL for a linear-in-parameter model with the squared loss, allowing efficient computation of the globally optimal solution in a closed-form. We applied the proposed PU-MTL to drivers' drowsiness prediction, and demonstrated that PN tasks can assist the performance of the PU task via information sharing and unlabeled samples can improve the prediction performance. On the other hand, the classification accuracy was insufficient for subjects 1 and 3. In future work, we will conduct theoretical analysis of PU-MTL, improve the drowsiness predictor and apply PU-MTL to other real-world problems.

6. REFERENCES

[1] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *International Journal of Neuroscience*, vol. 52, no. 1–2, pp. 29–37, 1990.

[2] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proceedings of the 14th SIGKDD Conference on Knowledge Discovery and Data Mining*, 2008, pp. 213–220.

[3] M. C. du Plessis, G. Niu, and M. Sugiyama, "Analysis of learning from positive and unlabeled data," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, vol. 27, pp. 703–711.

[4] M. C. du Plessis, G. Niu, and M. Sugiyama, "Convex formulation for learning from positive and unlabeled data," in *Proceedings of 32nd International Conference on Machine Learning (ICML2015), JMLR Workshop and Conference Proceedings*, F. Bach and D. Blei, Eds., 2015, vol. 37, pp. 1386–1394.

[5] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[6] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proceedings of the 10th SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004, pp. 109–117.

[7] T. Kato, H. Kashima, M. Sugiyama, and K. Asai, "Conic programming for multitask learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 7, pp. 957–968, 2010.

[8] N. Jaques, O. Rudovic, S. Taylor, A. Sano, and R. Picard, "Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation," in *Proceedings of IJCAI2017 Workshop on Artificial Intelligence in Affective Computing*, 2017, pp. 17–33.

[9] X. Sun, H. Kashima, R. Tomioka, N. Ueda, and P. Li, "A new multi-task learning method for personalized activity recognition," in *Proceedings of IEEE 11th International Conference on Data Mining*, 2011, pp. 1218–1223.

[10] D. Robertson, I. Biaggioni, G. Burnstock, P. A. Low, and J. F. R. Paton, Eds., *Primer on the Autonomic Nervous System, Third Edition*, Elsevier, 2012.

[11] B. Cinaz, B. Arnrich, R. La Marca, and G. Tröster, "Monitoring of mental workload levels during an everyday life office-work scenario," *Pervasive and Ubiquitous Computing*, vol. 17, no. 2, pp. 229–239, 2013.

[12] J. Vicente, P. Laguna, A. Bartra, and R. Bailon, "Drowsiness detection using heart rate variability," *Medical & Biological Engineering & Computing*, vol. 54, pp. 927–937, 2016.

[13] H. Kitajima, N. Numata, K. Yamamoto, and Y. Goi, "Prediction of automobile driver sleepiness (1st report, rating of sleepiness based on facial expression and examination of effective predictor indexes of sleepiness (in japanese)," *Transactions of the Japanese Society of Mechanical Engineers, Series C*, vol. 63, pp. 3059–3068, 1997.

[14] S. Hachisuka, K. Ishida, T. Enya, and M. Kamijo, "Facial expression measurement for detecting driver drowsiness," in *Engineering Psychology and Cognitive Ergonomics*, D. Harris, Ed., 2011, pp. 135–144.

[15] M. C. du Plessis and M. Sugiyama, "Class prior estimation from positive and unlabeled data," *IEICE Transactions on Information and Systems*, vol. E97-D, no. 5, pp. 1358–1362, 2014.