2019 IEEE Symposium Series on Computational Intelligence (SSCI)
December 6-9 2019, Xiamen, China

EEG-Based Driver Drowsiness Estimation Using Self-Paced Learning with Label Diversity

Yifan Xu
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
Wuhan, China
Email: yfxu@hust.edu.cn

Dongrui Wu
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
Wuhan, China
Email: drwu@hust.edu.cn

***Abstract*—Drowsy driving is one of the major contributors to traffic accidents. Continuously detecting the driver's drowsiness and taking actions accordingly may be one solution to improving driving safety. Electroencephalogram (EEG) signals contain information of the brain state, and hence can be utilized to estimate the driver's drowsiness level. A challenge in EEG-based drowsiness estimation is that when directly applied to a new subject without any calibration, the system's performance usually degrades significantly. Many efforts have been devoted to reducing the calibration data requirement, but there are still very few approaches that can completely eliminate the calibration process. This paper proposes a self-paced learning approach, which also takes the label diversity into consideration. The model learns from the easiest samples when the training first starts, and then more difficult ones are gradually added to the training process. This training strategy improves the generalization performance of the model on a new subject. Experiments on a simulated driving dataset with 15 subjects demonstrated that the proposed approach can better reduce the estimation error than several other approaches.**

***Keywords*—Drowsy driving; self-paced learning; EEG; brain-computer interface**

I. Introduction

Drowsy driving is one of the major contributors to road crashes. The Traffic Injury Research Foundation conducted a public opinion poll in 2007 [1] on 750 drivers. The results suggested that a majority of drivers (58.6%) had drowsy driving experience. 14.5% of drivers admitted that they had fallen asleep or "nodded off" while driving during the past year. Nearly 2% of the surveyed drivers had experienced a crash caused by drowsy driving in the past year. There are other surveys or polls [2], [3] reporting similar findings that a great proportion of drivers had the experience of driving while being fatigued or drowsy.

Monitoring the driver's drowsiness state and taking preventive actions accordingly may be one feasible solution to improving the driving safety. Commonly used approaches can be roughly divided into two categories: contactless detections and wearable sensor based detections. Contactless approaches [4], [5] are often computer vision based, and detect the drowsiness from the driver's eye, face and/or nodding activities. Wearable sensor based approaches collect and decode the driver's physiological signals, e.g., electroencephalogram (EEG) [6], electrocardiography (ECG) [7], electromyography (EMG) [8], etc., to estimate the drowsiness. We focus on the latter, especially, EEG-based drowsiness estimation, in this paper. Two commonly used features for EEG-based drowsiness estimation are the alpha (8-12Hz) and theta (4-7Hz) band powers [9]–[12]. Generally, as the drowsiness level increases, the alpha band power decreases, whereas the theta band power increases.

EEG signals contain information of the brain state, but are usually noisy and non-stationary [13]. Moreover, EEG signals also demonstrate strong individual differences, which may cause a model well-trained on data from existing subjects to perform poorly on a new subject. Subject-specific calibration is usually required when applying a brain-computer interface (BCI) to an unseen new subject. However, the calibration process is time-consuming and not user-friendly. Many efforts have been made to reduce or eliminate this calibration process for new subjects. Transfer learning [14], [15], which utilizes data from auxiliary subjects or sessions to facilitate the learning for a new subject, has been widely used in BCIs. Wei *et al.* [16] proposed to selectively exploit only the auxiliary data with relatively high transferability and achieved better performance than using all the data. Zanini *et al.* [17] proposed a covariance matrices alignment approach that centers the data of every session/subject with respect to the reference covariance matrix at the resting state in the Riemannian space. This approach can be used as a data preprocessing step. Wu *et al.* [11] proposed an online weighted adaptation regularization approach for regression problems, which can train the regression model with only a few amounts of calibration data. Although these approaches can significantly reduce the calibration effort, they cannot eliminate the calibration completely.

A plug-and-play BCI has no access to the subject-specific calibration data at all. We consider such a scenario in this paper. To train a subject-independent model, we should pay special attention to its generalization performance. Instead of transferring information from the auxiliary subjects, this paper adopts another strategy called curriculum learning [18], or self-paced learning (SPL) [19], which helps avoid local minima and hence improves the generalization performance.

Curriculum learning was proposed by Bengio *et al.* [18]

978-1-7281-2485-8/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:20:02 UTC from IEEE Xplore. Restrictions apply.

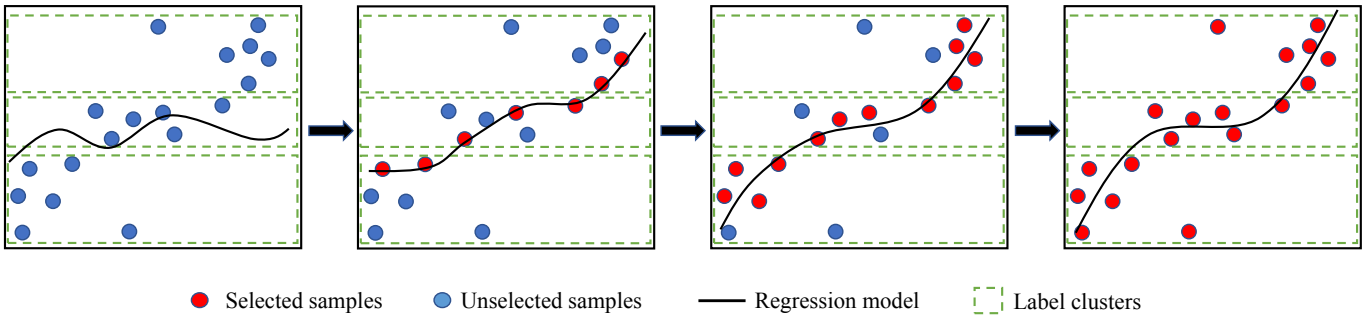

Fig. 1. Illustration of the learning process in SPLLD. Assume each sample has only one feature and one label, and we aim to build a regression model. In each subfigure, the red (blue) points are the samples involved (not involved) in current training epoch, and the black solid curve is the regression model updated on the red points. The dashed green rectangles represent different clusters according to the label. In the first subfigure, the model is randomly initialized. The other subfigures demonstrate the change of the model in training. An increasing number of samples from different label clusters are involved in training until the entire training set is used. The final model is shown in the last subfigure.

in 2009. It was inspired by the human learning process: we start from very basic concepts, and then advance to more difficult ones gradually. A child will get frustrated and unable to learn well if the difficulty level of a task is too high at the early stage of learning. Similarly, a machine learning model may be confused if the training samples are very difficult at the beginning. The key in curriculum learning is to design a curriculum, i.e., the order of samples to be fed into the model, according to their difficulty. The curriculums in different tasks are generally different.

Kumar *et al.* [19] proposed SPL, which automatically constructs the curriculum by explicitly defining the difficulty levels of the samples. The model can selectively learn from a subset of the training data according to its current performance. The loss function of SPL includes a regularization term on the weights of the samples, which can be applied to various tasks with different loss functions.

SPL has demonstrated promising performance in multiple applications [20]–[22]. However, SPL only considers the difficulty levels of the samples, but ignores their diversity. Thus, the easy samples selected by SPL may be redundant, which is not good for model training. To fix this deficiency, Jiang *et al.* [23] proposed self-paced learning with diversity (SPLD), which takes both the difficulty level and the diversity of the samples into consideration. SPLD has been empirically demonstrated to improve the generalization performance of the model.

This paper extends SPLD from classification to regression. We propose self-paced learning with label diversity (SPLLD), which considers the label diversity instead of the feature diversity, as illustrated in Fig. 1. We applied SPLLD to EEG-based driver drowsiness estimation and validated its improved performance.

The remainder of this paper is organized as follows: Section II introduces the proposed SPLLD algorithm. Section III compares SPLLD with a few other approaches in EEG-based driver drowsiness estimation. Section IV draws conclusions.

II. Self-Paced Learning with Label Diversity (SPLLD)

This section introduces SPL, SPLD, and our proposed SPLLD.

A. Self-Paced Learning (SPL)

Let the training set be $\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$ denotes the feature vector of the i^{th} sample, and y_i the corresponding label. Let $f(\mathbf{x};\boldsymbol{\theta})$ be a decision model on $\mathbf{x}$, parameterized by $\boldsymbol{\theta}$, and $L(y_i, f(\mathbf{x}_i;\boldsymbol{\theta}))$ be the loss function in evaluating the performance of $f(\mathbf{x};\boldsymbol{\theta})$. The loss can assume different forms, and the mean squared error loss is used in this paper.

The loss function of SPL is a weighted sum of $L(y_i, f(\mathbf{x}_i;\boldsymbol{\theta}))$ and a regularization term on the sample weights $\mathbf{v} \in \{0,1\}^n$:

$$\min_{\boldsymbol{\theta},\mathbf{v}} E(\boldsymbol{\theta}, \mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} v_i L(y_i, f(\mathbf{x}_i;\boldsymbol{\theta})) - \lambda\|\mathbf{v}\|_1, \quad (1)$$

where λ is a parameter indicating the learning pace, which can be viewed as the age of the model. The sample weight, v_i, is either 1 or 0, meaning the sample is either selected or not selected in the current iteration. Note that the l_1-norm regularizer can also take other forms [20], [24].

Parameters $\boldsymbol{\theta}$ and $\mathbf{v}$ are optimized alternatively. When $\mathbf{v}$ is fixed, gradient descent can be used to optimize $\boldsymbol{\theta}$. When $\boldsymbol{\theta}$ is fixed, the optimal $\mathbf{v}$ can be calculated by:

$$v_i = \begin{cases} 1, & L(y_i, f(\mathbf{x}_i;\boldsymbol{\theta})) < \lambda \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

i.e., only samples with loss smaller than the threshold λ are involved in the next round update of $\boldsymbol{\theta}$.

At the end of each training iteration, λ is multiplied by a constant step size μ (μ>1) to ensure that more difficult examples will be added to the training set in the next iteration. When λ is large enough, all samples are used in training, and the regularization term in (1) can be ignored. This ensures that the performance of SPL would not be worse than a traditional

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:20:02 UTC from IEEE Xplore. Restrictions apply.

training approach, i.e., using all training samples from the beginning.

It is very important to note that when optimizing $\boldsymbol{\theta}$ in a new iteration, it is initialized as the optimal $\boldsymbol{\theta}$ in the previous iteration. If $\boldsymbol{\theta}$ in a new iteration is initialized randomly, then there is no benefit to train the model iterative, and it is equivalent to training the model using all samples directly. Theoretically, both SPL and traditional learning have the same global optimum, but it is not easy to reach in practice. SPL, which starts with easier samples, may be less easily to be trapped in a local minimum.

Finally, if the loss function is convex and has a closed-form solution, then there is no benefit to use SPL, because the global optimum can always be reached.

B. Self-Paced Learning with Diversity (SPLD)

SPL has demonstrated promising performance. However, it does not take the diversity of the samples into account, and hence the selected samples may be very similar, and hence be redundant. Jiang *et al.* [23] proposed SPLD to improve SPL. It first divides the training set into several groups. Samples in the same group are more similar than those from a different group.

Let the training samples, $\mathbf{x} \in \mathbb{R}^{d\times n}$, be divided into b groups and denoted as $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(b)}$, where $\mathbf{x}^{(j)} \in \mathbb{R}^{d\times n_j}$ represents the j^{th} group with n_j samples and $\sum_{j=1}^{b} n_j = n$. To be consistent, the weight vector is denoted as $\mathbf{v} = [\mathbf{v}^{(1)}, ..., \mathbf{v}^{(b)}]$. The diversity of the samples is reflected in the scatter of non-zero elements in different $\mathbf{v}^{(j)}$. The loss function of SPLD is:

$$\min_{\boldsymbol{\theta},\mathbf{v}} E(\boldsymbol{\theta}, \mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} v_i L(y_i, f(\mathbf{x}_i; \boldsymbol{\theta})) - \lambda\|\mathbf{v}\|_1 - \gamma\|\mathbf{v}\|_{2,1} \tag{3}$$

where λ and γ are the parameters that weight the importance of the difficulty level and the diversity respectively. The $l_{2,1}$-norm is introduced to obtain group-sparse $\mathbf{v}$.

The optimization of SPLD is the same as SPL. $\boldsymbol{\theta}$ and $\mathbf{v}$ are updated alternatively. When $\mathbf{v}$ is fixed, $\boldsymbol{\theta}$, initialized from the previous iteration, can be optimized by gradient descent. When $\boldsymbol{\theta}$ is fixed, the optimal $\mathbf{v}^{(j)}$ in $\mathbf{v}$ is computed by:

$$v_i^{(j)} = \begin{cases} 1, & L(y_i^{(j)}, f(\mathbf{x}_i^{(j)}; \boldsymbol{\theta})) < \lambda + \frac{\gamma}{\sqrt{i}+\sqrt{i-1}} \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

where i is the sample's rank after all the samples in the j^{th} group are sorted in the ascending order according to their loss L. As in SPL, $\mathbf{v}$ in SPLD is still determined by a threshold. The difference is that the threshold has an additional term $\frac{\gamma}{\sqrt{i}+\sqrt{i-1}}$, which penalizes samples selected from the same group.

At the end of each training iteration, both λ and γ are multiplied by a constant step size so that eventually all samples are used in training. Note that (3) degrades to (1) when only one group is used, or the group number equals the number of training samples.

Algorithm 1: SPLLD for regression problems.

Input: Training set $\mathcal{D}$;
Step size μ;
Number of groups b;
Output: Regression model parameter $\boldsymbol{\theta}$.
Cluster the training samples into b groups $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(b)}\}$, according to their labels;
Randomly initialize the regression model parameters $\boldsymbol{\theta}$;
Compute $L(y_i^{(j)}, f(\mathbf{x}_i^{(j)}; \boldsymbol{\theta}))$ for each $\mathbf{x}_i$ in each group;
Set λ and γ to half the median of all samples' loss values;
while *training* **do**
  for $j = 1 : b$ **do**
    Sort the samples in $\mathbf{x}^{(j)}$ as $(\mathbf{x}_1^{(j)}, ..., \mathbf{x}_{n_j}^{(j)})$, in ascending order of their loss value L;
    for $i = 1 : n_j$ **do**
      Compute the weight $v_i^{(j)}$ of the sample $\mathbf{x}_i^{(j)}$ using (4);
    end
  end
  Update $\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}, \mathbf{v})$ in (3) using gradient descent, where $\boldsymbol{\theta}$ is initialized as the optimal $\boldsymbol{\theta}$ from the previous iteration;
  $\lambda = \mu \cdot \lambda$;
  $\gamma = \mu \cdot \gamma$;
end

C. Self-Paced Learning with Label Diversity (SPLLD)

The groups in SPLD are obtained by clustering the samples in the feature space so that the feature diversity is taken into consideration. We hypothesize that the label diversity is important as well. Thus, in SPLLD we cluster the samples according to their labels. The remaining procedure is the same as SPLD.

The pseudocode of SPLLD is shown in Algorithm 1. For simplicity, we set λ and γ to be equal, i.e., the importance of the difficulty level and the diversity are the same. More specifically, we set λ and γ to be half the median of all samples' initial losses to ensure that around half of the samples are included in the first training iteration.

III. Experiments and Discussion

This section compares the performance of SPLLD with a few other approaches.

A. Dataset and Feature Extraction

The dataset used in this study was identical to that used in [10], [11]. Sixteen healthy subjects with normal or corrected to normal vision participated in a sustained-attention driving experiment [25], [26], using a real vehicle mounted on a motion platform with 6 degrees of freedom immersed in a 360-degree virtual-reality scene. The experiment simulated driving on an empty highway at 100km/h, with lane-departure events

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:20:02 UTC from IEEE Xplore. Restrictions apply.

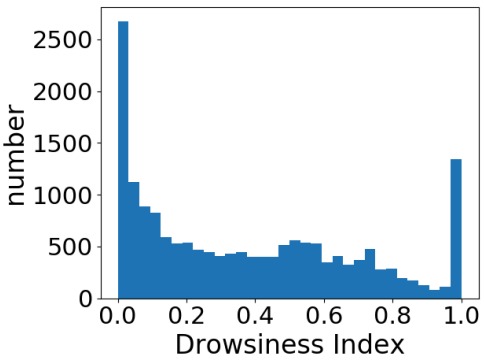

Fig. 2. Histogram of the drowsiness index from all 15 subjects.

randomly activated every 5-10 seconds. The subjects needed to steer the car back to the center of the lane as quickly as possible.

The time between the lane-departure event onset and the driver reaction onset was recorded and later converted to a drowsiness index (DI) [16],

$$DI = \max\left(0, \frac{1 - e^{-(\tau - \tau_0)}}{1 + e^{-(\tau - \tau_0)}}\right), \tag{5}$$

where τ_0 was set to 1 in our work, as in [10], [11]. The DIs were then smoothed by a 90s moving-average window to reduce variations.

Each experiment lasted for about 60-90 minutes and was conducted in the afternoon when people tend to feel sleepy. Participants' scalp EEG signals were recorded during the experiment using a 32-channel Neuroscan system (30-channel EEGs plus 2-channel earlobes). Data from one subject was not recorded correctly, so we only used data from the other 15 subjects. To ensure the fairness of comparison, we used the first 3,600 seconds data from each subject. The histogram of all drowsiness index values from the 15 subjects is shown in Fig. 2.

We used EEGLAB [27] for signal preprocessing. A band-pass filter (1-50 Hz) was first applied to reduce the artifacts, noise and DC drift. Then, the EEG data were downsampled from 500 Hz to 250 Hz and re-referenced to averaged earlobes. We used 30-second EEG signal before each sample point to predict the DI for each subject every 3 seconds. The average power spectral density (PSD) in theta and alpha bands were then computed using Welch's method [28] for each channel. The PSDs were then converted into dBs and used as the features in our experiment. There were $30 \times 2 = 60$ features for each sample.

B. Experimental Setting and Performance Measures

We compared the performances of the following approaches using leave-one-subject-out cross-validation:

1) *Baseline*: All data from 14 subjects were combined to train a neural network regression model for the remaining subject.

TABLE I
AVERAGE RMSES AND CCS OF THE 15 SUBJECTS.

	Baseline	SPL	SPLD	SPLLD
RMSE	0.3061	0.2938	0.2881	**0.2774**
CC	0.4680	0.5076	0.5022	**0.5234**

2) *SPL*, which has been introduced in Section II-A. λ was set to the median of the initial losses of all samples. The step size μ was set to 1.2, a little smaller than that in [19].
3) *SPLD*, which has been introduced in Section II-B. λ and γ were set to half the median of all samples' initial losses. The step size μ was set to 1.2, and the number of clusters was 32.
4) *SPLLD*, which has been introduced in Section II-C. Its parameters were the same as those in SPLD.

A neural network using one 40-node hidden layer and ReLU activation function was trained in all four approaches for regression. It was optimized using stochastic gradient descent with momentum 0.9, batch size 32, learning rate 0.001, and weight decay 0.00005. We mixed all data from the remaining 14 subjects, reserved 10% for validation in early-stopping, using a patience of 10 epochs. Except for the baseline that used early-stopping directly, the other three approaches activated early-stopping only when all samples were involved in training. The maximum number of training epochs was 500. We repeated each approach five times and report the average results.

Root mean squared error (RMSE) and the Pearson correlation coefficient (CC) were used as our performance measures.

C. Experiment Results

The RMSEs and CCs for each subject, averaged across five runs, are shown in Fig. 3. The average RMSEs and CCs of all 15 subjects are shown in Table I. The performances of the three SPL-based approaches on the individual subjects were all better than or comparable with the baseline. Especially, for Subjects 11 and 15, on which the baseline performed poorly, the three SPL-based approaches performed much better. On average, the three SPL-based approaches all outperformed the baseline.

Among the three SPL-based approaches, both SPLD and SPLLD outperformed the SPL, suggesting that the diversity did matter. Although the proposed SPLLD achieved the best average performance, it was only slightly better than SPLD, because both the features and the label contain useful information about the diversity. However, clustering the labels is much faster than clustering the features, because of its low dimensionality. So, SPLLD is more efficient than SPLD.

D. Parameter Sensitivity Study

The hyper-parameters in SPLLD include: b, the number of clusters; the initial weights of the two regularization terms in (3); λ and γ, and their step size μ. Experiments were performed to find out the sensitivity of SPLLD to them.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:20:02 UTC from IEEE Xplore. Restrictions apply.

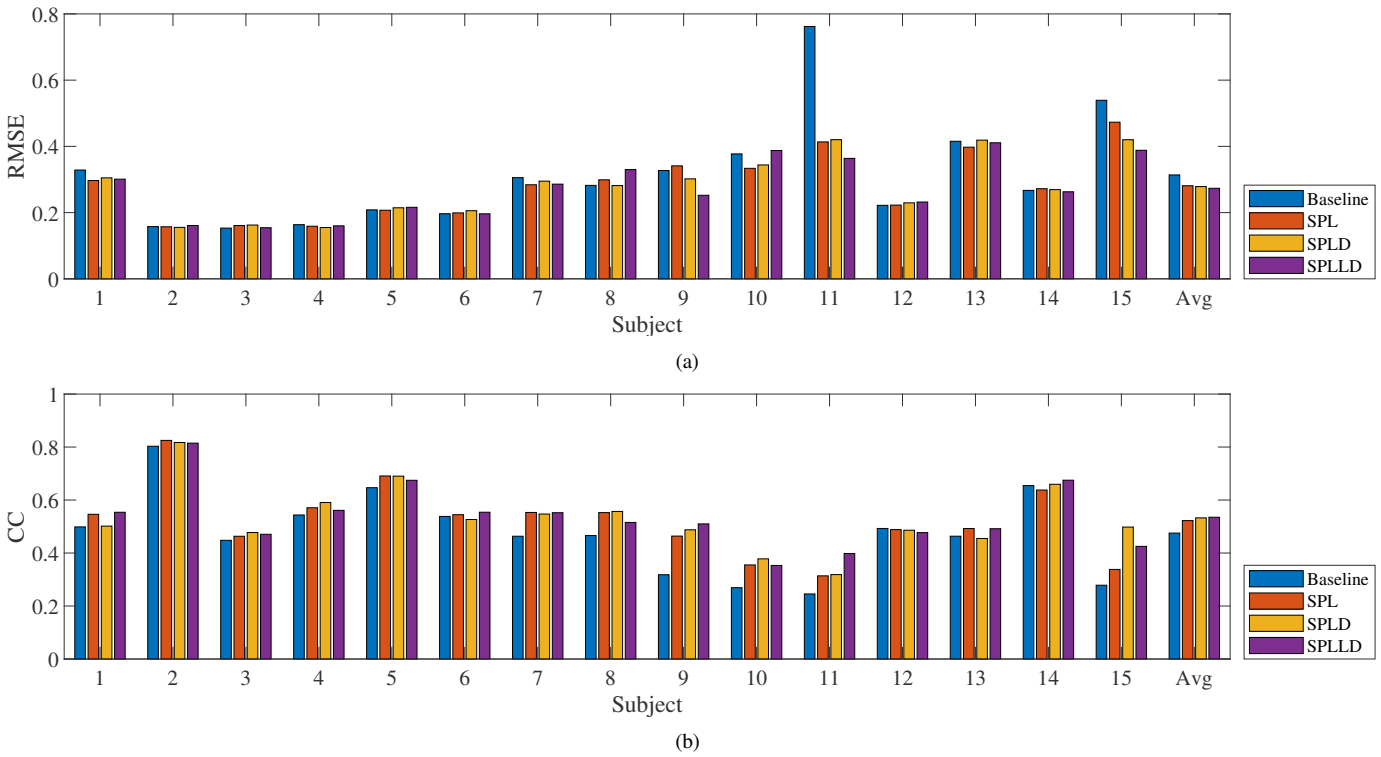

Fig. 3. (a) RMSEs and (b) CCs in leave-one-subject-out cross-validation. The experiments were repeated five times on each subject, and the averages are shown.

TABLE II
AVERAGE RMSES AND CCS FOR DIFFERENT NUMBER OF CLUSTERS.

Number of Clusters	16	32	64
RMSE	0.2789	**0.2735**	0.3090
CC	0.5281	**0.5351**	0.5008

TABLE III
AVERAGE RMSES AND CCS FOR DIFFERENT STEP SIZE.

Step Size	1.05	1.2	1.5	1.8
RMSE	0.2872	**0.2735**	0.2868	0.2914
CC	0.5097	**0.5351**	0.4972	0.4903

1) b, the Number of Clusters: In [24], the number of groups b was selected from $\{32, 64, 128, 256\}$. In our experiments, there were 14 auxiliary subjects and one testing subject in each run. Each subject had around one thousand samples. The histogram in Fig. 2 shows that the drowsiness index labels biased towards 0 and 1, but the distributions in-between did not vary much. So, we may not need too many clusters. We performed experiments on $b \in \{16, 32, 64\}$ and show the results in Table II. All other parameters were the same as those in Section III-B. Intuitively, b should not be too small or too big, because when b equals one or the number of training samples, SPLLD degrades to SPL, and hence the diversity information is not used at all. Table II confirms this.

2) μ, the Step Size of the Regularization Weights: In our experiment, we simply set the initial λ and γ to be the same, and their step size $\mu = 1.2$. $\mu = 1.3$ was used in [19]. In [23], the loss function was convex, and the final results was irrelevant to μ. Our loss function is non-convex, so it may be influenced by the self-paced step size μ. The average RMSEs and CCs for $\mu \in \{1.05, 1.2, 1.5, 1.8\}$ are shown in Table III.

It can be seen that the results obtained from a small or large step size were a little worse than that from a moderate step size. The loss of samples tends to decrease as the training proceeds and the number of the samples incorporated into training would increase automatically even if the two thresholds stay constant. Therefore, it is not necessary to set the step size too large, or the model would learn at a too fast pace. When the step size becomes extremely large, there would be no difference between training with and without SPLLD. On the other hand, when the step size is too small, the model is likely to repeatedly learn from almost the same set of samples, resulting in bad generalization.

3) The Initial Value of λ and γ: In the SPL [19], the initial λ was set such that more than half of the samples were selected for training. In [23], a number of randomly selected samples were used in the first training iteration and after that λ and γ were used to select the training samples. They tuned λ and γ on the validation set through a linear search strategy.

For simplicity, we set the initial λ and γ to half of the median of all samples' losses in previous comparisons to ensure that at least around half of the training set were involved in the first training epoch. To study how initial λ and γ affect the performance, we tried to set the sum of them to

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:20:02 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
AVERAGE RMSES AND CCS WITH DIFFERENT INITIALIZATIONS OF λ AND γ.

Percentile	35	50	65
RMSE	0.2902	**0.2735**	0.3047
CC	0.5193	**0.5351**	0.4926

different percentiles of all samples' initial losses. The results are shown in Table IV. Except for the initial λ and γ, all other parameters were the same as those in section III-B. The results demonstrate that too many samples being added to training at first may introduce noisy samples, which may confuse the model. On the other hand, a small sample set may result in overfitting. Therefore, half of the number of samples seems to be reasonable and suitable for the model to learn at the beginning.

E. Discussion

As mentioned in [23], SPL based approaches usually have the limitation that being unstable to random starting values. Since we evaluate the easiness of the samples referring to their loss values with the randomly initialized model and generate the training set for the next training epoch, the final model much depends on the initialization. If the samples used in the first training iteration can be properly selected, SPL approaches are more likely to obtain better performance.

There are many parameters requiring pre-definition in the algorithm. The initial values of λ and γ and the step size do not need much tuning and can be set referring to the previous work. The number of clusters can be chosen according to the distribution of the data.

IV. CONCLUSION

Drowsy driving is pervasive among drivers, and is one of the major contributors to vehicle accidents. Detecting the driver's drowsiness level in real-time and taking preventive actions accordingly may help improve driving safety. EEG signals can be used to estimate the driver's drowsiness level. However, individual differences make it challenging to apply a model trained on existing subjects to a new subject, without tuning its parameters on some subject-specific calibration data. This paper proposed an SPLLD approach that exploits the label diversity in self-paced learning to train a more robust nonlinear regression model that generalizes better to new subjects in EEG-based driver drowsiness estimation.

REFERENCES

[1] W. Vanlaar, H. Simpson, D. Mayhew, and R. Robertson, "Fatigued and drowsy driving: A survey of attitudes, opinions and behaviors," *Journal of Safety Research*, vol. 39, no. 3, pp. 303–309, 2008.

[2] D. Royal *et al.*, "National survey of distracted and drowsy driving attitudes and behavior: 2002, vol 1: Findings," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep. DOT HS 809 566, 2003.

[3] B. C. Tefft, "Asleep at the wheel: The prevalence and impact of drowsy driving," Washington, DC, 2010, [Online] Available: https://www.aaafoundation.org/sites/default/files/2010DrowsyDrivingReport_1.pdf.

[4] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Trans. on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 63–77, 2006.

[5] T. DOrazio, M. Leo, C. Guaragnella, and A. Distante, "A visual approach for driver inattention detection," *Pattern Recognition*, vol. 40, no. 8, pp. 2341–2355, 2007.

[6] C.-T. Lin, R.-C. Wu, S.-F. Liang, W.-H. Chao, Y.-J. Chen, and T.-P. Jung, "EEG-based drowsiness estimation for safety driving using independent component analysis," *IEEE Trans. on Circuits and Systems-I*, vol. 52, no. 12, pp. 2726–2738, 2005.

[7] G. Jahn, A. Oehme, J. F. Krems, and C. Gelau, "Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 3, pp. 255–275, 2005.

[8] M. Akin, M. B. Kurt, N. Sezgin, and M. Bayram, "Estimating vigilance level by using EEG and EMG signals," *Neural Computing and Applications*, vol. 17, no. 3, pp. 227–236, 2008.

[9] Y. Cui and D. Wu, "EEG-based driver drowsiness estimation using convolutional neural networks," in *Proc. Int'l Conf. on Neural Information Processing*. Guangzhou, China: Springer, Nov. 2017, pp. 822–832.

[10] D. Wu, C.-H. Chuang, and C.-T. Lin, "Online driver's drowsiness estimation using domain adaptation with model fusion," in *Proc. Int'l Conf. on Affective Computing and Intelligent Interaction*. Xi'an, China: IEEE, Sep. 2015, pp. 904–910.

[11] D. Wu, V. J. Lawhern, S. Gordon, B. J. Lance, and C.-T. Lin, "Driver drowsiness estimation from EEG signals using online weighted adaptation regularization for regression (OwARR)," *IEEE Trans. on Fuzzy Systems*, vol. 25, no. 6, pp. 1522–1535, 2017.

[12] D. Wu, V. J. Lawhern, S. Gordon, B. J. Lance, and C.-T. Lin, "Offline EEG-based driver drowsiness estimation using enhanced batch-mode active learning (EBMAL) for regression," in *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics*, Budapest, Hungary, October 2016, pp. 730–736.

[13] A. M. Azab, J. Toth, L. S. Mihaylova, and M. Arvaneh, *Signal processing and machine learning for brain-machine interfaces*. Institution of Engineering and Technology, 2018, ch. 5, pp. 81–101.

[14] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[15] H. He and D. Wu, "Transfer learning for brain-computer interfaces: A Euclidean space data alignment approach," *IEEE Trans. on Biomedical Engineering*, 2019, in press.

[16] C.-S. Wei, Y.-P. Lin, Y.-T. Wang, T.-P. Jung, N. Bigdely-Shamlo, and C.-T. Lin, "Selective transfer learning for EEG-based drowsiness detection," in *Proc. IEEE Int.l Conf, on Systems, Man, and Cybernetics*. Hong Kong: IEEE, oct 2015, pp. 3229–3232.

[17] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Transfer learning: A Riemannian geometry framework with applications to brain-computer interfaces," *IEEE Trans. on Biomedical Engineering*, vol. 65, no. 5, pp. 1107–1116, 2017.

[18] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int'l Conf. on Machine Learning*. Montreal, Canada: ACM, Jun. 2009, pp. 41–48.

[19] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2010, pp. 1189–1197.

[20] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proc. Int'l Conf. on Multimedia*. Orlando, FL: ACM, Nov. 2014, pp. 547–556.

[21] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *Proc. Advances in Neural Information Processing Systems*, Lake Tahoe, NV, Dec. 2012, pp. 638–646.

[22] M. P. Kumar, H. Turki, D. Preston, and D. Koller, "Learning specific-class segmentation from diverse data," in *Proc. Int'l Conf. on Computer Vision*. Barcelona, Spain: IEEE, Nov. 2011, pp. 1800–1807.

[23] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2014, pp. 2078–2086.

[24] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *Proc. AAAI Conf. on Artificial Intelligence*, Austin, TX, Jan. 2015.

[25] S.-W. Chuang, L.-W. Ko, Y.-P. Lin, R.-S. Huang, T.-P. Jung, and C.-T. Lin, "Co-modulatory spectral changes in independent brain processes

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:20:02 UTC from IEEE Xplore. Restrictions apply.

are correlated with task performance," *Neuroimage*, vol. 62, no. 3, pp. 1469–1477, 2012.
[26] C.-H. Chuang, L.-W. Ko, T.-P. Jung, and C.-T. Lin, "Kinesthesia in a sustained-attention driving task," *Neuroimage*, vol. 91, pp. 187–202, 2014.
[27] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.
[28] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:20:02 UTC from IEEE Xplore. Restrictions apply.