BEB Conference on Biomedical Engineering and Biotechnology

BEB6155

Semi-supervised sparse representation classification for sleep EEG recognition with imbalanced sample sets

Xiaolei Wuzheng, Shigang Zuo, Li Yao, Xiaojie Zhao*

School of Artificial Intelligence, Beijing Normal University, Beijing, China

**** Corresponding author, E-mail: zhaox86@163.com***

I. Introduction

The accuracy of pseudo-labels in semi-supervised learning may affect the performance of the classification model. Based on semi-supervised sparse representation classification (SemiSRC), this study proposed an improved sparse concentration index to estimate the confidence of pseudo-labels data for sleep EEG recognition, considering both interclass differences and intraclass concentration. In view of class imbalance in sleep EEG data, the synthetic minority oversampling technique was also improved to remove mixed samples at the boundary between minority and majority classes. The results showed that the proposed method achieved better classification performance, in which the classification accuracy after class balancing was obviously higher than that before class balancing

II. Methods

1. Data Preprocessing

- 20 subjects (11 male and 9 female, age 28.6 ± 2.8).
- Pz-Oz channel with a sampling frequency of 100 Hz.
- a finite impulse response (FIR) bandpass filter at 0.4-35 Hz.

2.Feature Extraction

- 5-layer wavelet packet with the db4 wavelet base.
- Energy ratio, variance, and kurtosis.
- Sample entropy with thresholds of 0.2 and 0.15.

3. Improved Confidence Estimation for Unlabeled Samples

- The sparse concentration index (SCI): an estimation criterion of confidence to exclude samples with low intraclass concentration.
- **The improved SCI (SCI_Impro)** optimizes the **SCI-based** confidence estimation by **combining** the mean of sparse coefficients (MSC).

$$SCI\_Impro(\alpha_i)_m = \left[\left(\lambda \frac{\sum_{j=1}^{m-1} SCI(\alpha_i)_j}{m-1}\right) \cap \left(\lambda \frac{\sum_{j=1}^{m-1} MSC(\alpha_i)_j}{m-1}\right) \right]$$

- **MSC**: Correlation between the real class and the sparse representation class of test samples.

4. Expansion and Optimization from Imbalanced Data

- Expansion:
  - The improved SMOTE algorithm focuses on both the reasonable **choice** of the **minority** and **removing** of the **majority** at the classification boundary in the process of class balancing.
- Optimization:
  - Optimization **A**/**B** balanced the training set after adding **all/each** pseudo-label samples/ sample.

III. Results

- **SemiSRC with SCI and SCI_Impro in labeling samples**

If NS=2, it means Awake (AWA) stage, sleep stage1(S1+S2+S3+S4+REM stage); If NS=3, it means AWA stage, REM stage, sleep stage 1(S1+S2+S3+S4 stage); If NS=4, it means AWA stage, REM stage, sleep stage 1(S1+S2), sleep stage 3(S3+S4 stage); If NS=5, it means AWA stage, REM stage, sleep stage 1(S1), sleep stage 2(S2), sleep stage 3(S3+S4 stage); If NS=6, it means AWA stage, REM stage, S1 stage, S2 stage, S3 stage and S4 stage. The performances of semiSRC with SCI and SCI_Impro were compared in the number of sleep stages (NS) from 2 to 6 (Figure 1).

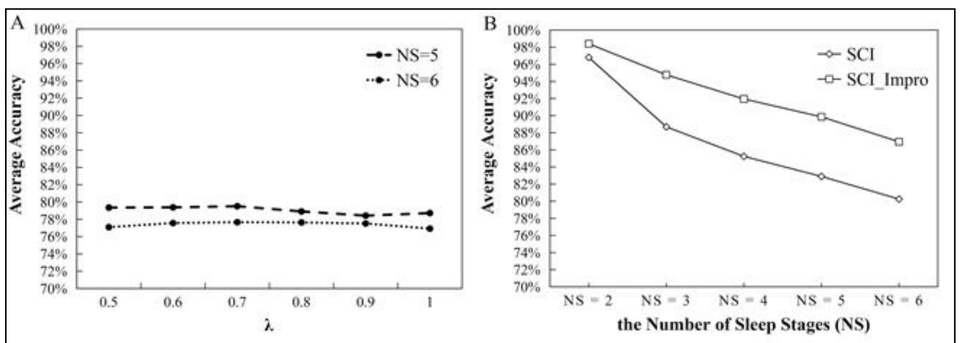

Figure 1. Comparison of the average accuracy between the SCI and SCI_Impro algorithms. A. The λ corresponding to the highest classification accuracy of sleep stage was used as the optimal parameter of 0.7. B. The labeling accuracy of the pseudo-label sample by two kinds of confidence estimation at different numbers of sleep stages.

With the reduction in NS, the classification accuracy increased, and the difference between the SCI and SCI_Impro gradually decreased.

- **SemiSRC with SCI and SCI_Impro in sleep stage classification**

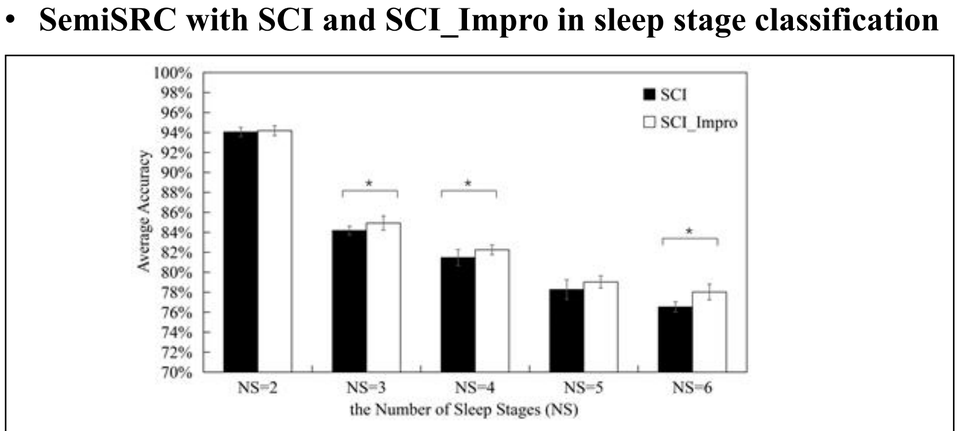

Figure 2. Average accuracy of the semiSRC model using different confidence estimations. * : $p < 0.05$.

With a significant difference, the average accuracy of SCI_Impro was higher than that of SCI except that no significant difference occurred when NS was 2.

- **Comparison with other classifiers**

To further verify the effectiveness of the proposed semiSRC, three other classification models were compared with our model at NS values from 2 to 6 (Table 1).

NS	SRC	naïve_ semiSRC	semi SVM	semi SRC
NS=2	94.20%	**94.27%**	*91.22%	94.18%
NS=3	94.20%	**94.27%**	*91.22%	94.18%
NS=4	82.18%	81.87%	*81.46%	**82.25%**
NS=5	78.20%	79.06%	*78.58%	**79.53%**
NS=6	*77.12%	*77.31%	*76.00%	**78.02%**

Table 1. The performance of four classifiers (SRC, naïve_semiSRC, semi_SVM, and semiSRC) at different numbers of sleep stages. * : $p < 0.05$. Specially, p represents the level of significance, which is compared to semiSRC.

The accuracy of our semiSRC was generally higher than that of SRC, naïve_semiSRC, and semiSVM. With increasing NS, our semiSRC showed better performance than the other models.

- **The improved SMOTE algorithm was compared with the traditional SMOTE in data distribution**

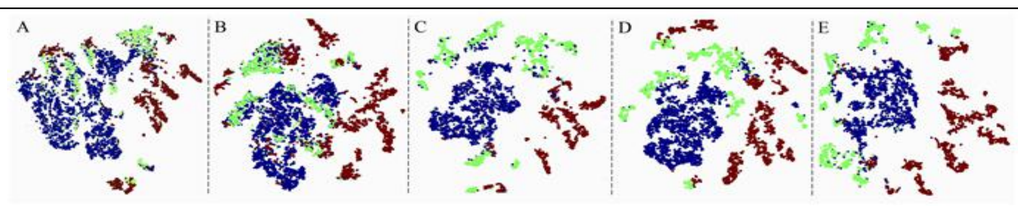

Figure 3. Data distribution after using the SMOTE or improved SMOTE algorithm.The green point represents samples from the REM stage, the red point represents samples from the AWA stage, and the blue point represents samples from the non-REM (S1+S2+S3+S4) stage. (A) Data distribution without class balancing (imbalanced data). (B) Data distribution with the SMOTE algorithm combined with optimization A. (C) Data distribution with the SMOTE algorithm combined with optimization B. (D) Data distribution with the improved SMOTE algorithm combined with optimization A. (E) Data distribution with the improved SMOTE algorithm combined with optimization B.

In Figure 3, the new synthetic samples increased the number of samples of the minority class and concentrated the samples of the minority class (Figure 3B to 3E).

- **Comparison between SMOTE or improved SMOTE with two optimizations in sleep stage classification**.

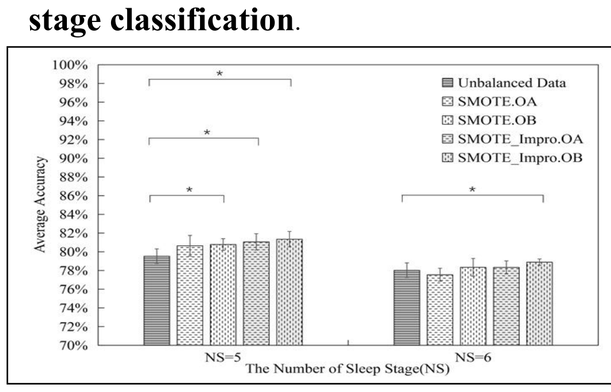

Figure 4. Average accuracy of the proposed semiSRC after class balancing using the SMOTE or improved SMOTE combined with optimization A (OA) or optimization B (OB). * : $p < 0.05$. Specially, p represents the level of significance, which is compared to Unlabeled Data.

After class balancing, improved SMOTE algorithm improved the sleep staging performance compared with SMOTE at NS = 5 and NS = 6. Moreover, optimization B showed more improved accuracy than optimization A in both SMOTE and improved SMOTE.

IV. Conclusions

- The proposed semiSRC algorithm uses a large amount of unlabeled data to reduce the burden of manual labeling. In semiSRC, SCI_Impro was used as a estimation criterion of confidence of pseudo-label samples and selected the samples not only with high concentrations within a class but also with high differences between classes, showing that SCI_Impro selected more reliable and effective samples than SCI. After the first high-confidence sample was selected by the SCI_Impro, the confidence of the subsequent m samples will increase and expand into the training set, resulting in better performance in the classification of sleep stages than other commonly used models.
- Because the proportion of each stage in all-night sleep time is quite different, a serious class imbalance affects the performance of the classifier. The proposed semiSRC could obtain a better classification boundaries of the dataset after the adjustment of the balancing algorithm. Besides, optimization B showed better overall accuracy than optimization A, regardless of the balancing algorithm (SMOTE or improved SMOTE).