Multi-Level Multi-Task Structured Sparse Learning for Diagnosis of Schizophrenia Disease

Mingliang Wang[1], Xiaoke Hao[1], Jiashuang Huang[1], Kangcheng Wang[2], Xijia Xu[3], Daoqiang Zhang[1(✉)]

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
dqzhang@nuaa.edu.cn
[2] Department of Psychology, Southwest University, Chongqing 400715, China
[3] Department of Psychiatry, Affiliated Nanjing Brain Hospital of Nanjing Medical University Nanjing University, Nanjing 210029, China

Abstract. In recent studies, it has attracted increasing attention in multi-frequency bands analysis for diagnosis of schizophrenia (SZ). However, most existing feature selection methods designed for multi-frequency bands analysis do not take into account the inherent structures (i.e., both frequency specificity and complementary information) from multi-frequency bands in the model, which are limited to identify the discriminative feature subset in a single step. To address this problem, we propose a multi-level multi-task structured sparse learning (MLMT-TS) framework to explicitly consider the common features with a hierarchical structure. Specifically, we introduce two regularization terms in the hierarchical framework to impose the common features across different bands and the specificity from individuals. Then, the selected features are used to construct multiple support vector machine (SVM) classifiers. Finally, we adopt an ensemble strategy to combine outputs of all SVM classifiers to achieve the final decision. Our method has been evaluated on 46 subjects, and the superior classification results demonstrate the effectiveness of our proposed method as compared to other methods.

1 Introduction

Schizophrenia is the most common chronic and devastating mental disorders affecting 1% of the population worldwide [1]. Until now, the pathological mechanism of schizophrenia remains unclear and there is no definitive standard in the diagnosis of schizophrenia. While it has been reported that there is a significant change in the structure, function and metabolism of the brain in schizophrenia patients [2]. Moreover, some related studies have suggested that resting-state

This work was supported in part by the National Natural Science Foundation of China (61422204; 61473149) and the NUAA Fundamental Research Funds (No. NE2013105).

functional analyses are beneficial to achieve more complete information of functional connectivity. In fact, most resting-state functional magnetic resonance imaging (RS-fMRI) studies have examined spontaneous low-frequency oscillations (LFO) at a specific frequency band of 0.01-0.1Hz [3]. Therefore, we can use the RS-fMRI data for whole-brain analysis in this study.

In addition, as the complexity of the schizophrenia itself, some studies have reported mixed results even opposite conclusions [3], which may be due to the different frequency bands used in these studies. Hence, it has been recognized that neural disorder specific changes could be restricted to the specific frequency bands recently. Also, more and more studies have indicated that taking the high-frequency bands into consideration is helpful to measure the intrinsic brain activity of schizophrenia. Therefore, in this study we decompose RS-fMRI LFO into four distinct frequency bands based on prior work (slow-5(0.01-0.027Hz), slow-4(0.027-0.073Hz), slow-3(0.073-0.198Hz), slow-2(0.198-0.25Hz)) [4].

To exploit potential information sharing among different frequency bands, instead of treating each frequency band as a single-task classification problem, multi-task learning (MTL) paradigm learns several related tasks simultaneously to improve performance [5]. However, the main limitation of existing multi-task works is that all tasks are considered in a single level, which may miss out some relevant features shared between a smaller group of tasks. Meanwhile, it may be not enough to model such complex structure information in schizophrenia studies through a single level method.

Accordingly, in this paper, we propose a multi-level multi-task structured sparse (MLMT-TS) learning method, to explicitly model the structure information of multi-frequency data for diagnosis of schizophrenia. In our hierarchical framework, $\ell_{1,1}$-norm is brought to induce the sparsity and select the specific features, meanwhile a new regularization term is introduced to capture the common features across different bands. Hence, contrary to the single level manner, the hierarchical framework gradually enforces different levels of features sharing to model the complex structure information efficiently.

2 Method

Data and Pre-Processing: In this study, we use 46 subjects in total from the Department of Psychiatry, Affiliated Nanjing Brain Hospital of Nanjing Medical University. Among them, 23 are schizophrenia patients, and the rest 23 subjects are normal controls (NC). All subjects RS-fMRI images were processed as described in [3], and after preprocessing the fractional amplitude of LFO were calculated using REST software[1]. Because the size of the RS-fMRI image is $61 \times 73 \times 61$, the voxel-based analysis is too large and noisy to directly used for disease diagnosis. Thus, we adopt a simple and effective way to extract more relevant and discriminative features for neuroimage analysis and classification. We first utilize the patch-based method with patch size $3 \times 3 \times 3$ voxels to divide

[1] http://www.restfmri.net

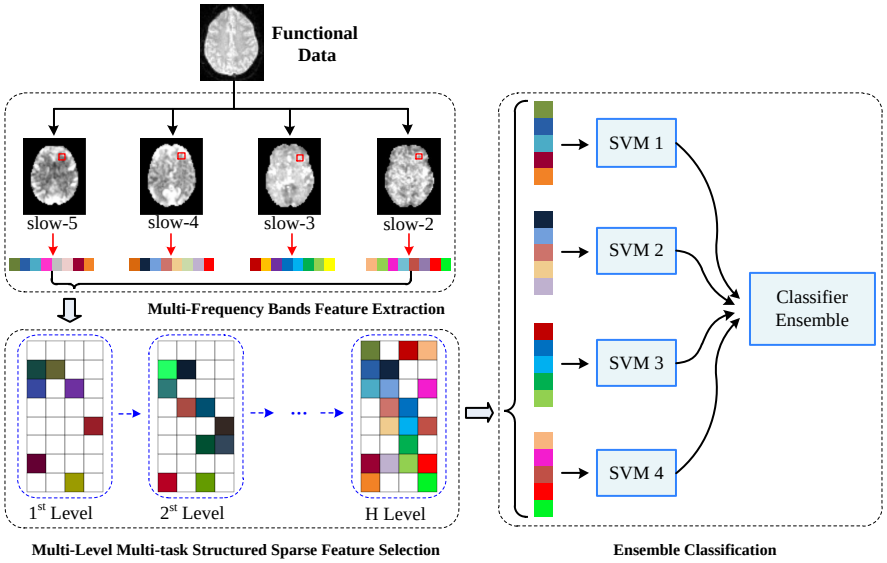


Fig. 1. The framework of the proposed classification algorithm.

the whole brain into some candidate patches. Then, we perform the t-test on the candidate patches and select the significant patch with the p-value smaller than 0.05. Finally, we calculate the mean of each patch, and treat it as the feature of the selected patch.

Multi-Level Multi-Task Structured Sparse Learning Model: The framework of the proposed method is illustrated in Fig. 1. After the patch-based multi-frequency bands features are extracted from the RS-fMRI data, our multi-level multi-task structured sparse (MLMT-TS) method is used to select the more relevant and discriminative features. The features are selected in an iterative fashion: It starts with a low level where sharing features among all tasks are induced, and then gradually enhances the incentive to share in successive levels. However the specific features of different bands are induced with a high level, then the incentive is gradually decreased. In addition, we also need to note that the learned coefficient matrix corresponding to each level is forwarded to the next hierarchy for further leaning in the same manner. In such a hierarchical manner, we gradually select the most discriminative features in order to sufficiently utilize the complementary and specific information of multi-frequency bands. Then, the selected features are used to train SVM classifiers for schizophrenia disease classification. Finally, we adopt an ensemble classification strategy, a simple and effective classifier fusion method, to combine the outputs of different SVM classifiers to make a final decision. In the following, we explain in detail how the MLMT-TS feature selection method works.

Assume that there are M supervised learning tasks corresponding to the number of frequency bands. And the training data matrix on m-th task from N

training subjects is denoted by $\mathbf{X}_m = [\mathbf{x}_{m,1}, \mathbf{x}_{m,2}, \ldots, \mathbf{x}_{m,N}]^T \in \mathbb{R}^{N\times d}$, and $\mathbf{Y} = [y_1, y_2, \ldots, y_N]^T \in \mathbb{R}^N$ as the response vector from these training subjects, where $\mathbf{x}_{m,n}$ is the feature vector of the n-th subject and the corresponding class label is y_n. Denote the coefficient matrix as $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M] \in \mathbb{R}^{d\times M}$, where $\mathbf{w}_m \in \mathbb{R}^d$ is a linear discriminant function for task m, and we assume the bias term $\mathbf{b}$ is absorbed into $\mathbf{W}$. As mentioned above, the coefficient matrix is forwarded to subsequent learning to share more structure information. Hence, we decompose the coefficient matrix into H components where each hierarchy can capture the level-specific task group structure features. Specifically, the coefficient matrix $\mathbf{W}$ can be defined as

$$\mathbf{W} = \sum_{h=1}^{H} \mathbf{W}_h \tag{1}$$

where $\mathbf{W}_h = [\mathbf{w}_{h,1}, \ldots, \mathbf{w}_{h,M}] \in \mathbb{R}^{d\times M}$ is the coefficient matrix, which is corresponding to the h-th level, and $\mathbf{w}_{h,m}$ is the m-th column of $\mathbf{W}_h$ in the h-th level. Then the objective function for MLMT-TS feature selection method can be written as:

$$min_{\mathbf{W}} \sum_{m=1}^{M} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}_m \sum_{h=1}^{H} \mathbf{w}_{h,m}\|_2^2 + R_t(\mathbf{W}) + R_s(\mathbf{W}) \tag{2}$$

where $R_t(\mathbf{W})$ and $R_s(\mathbf{W})$ are the structure regularization term and the $\ell_{1,1}$-norm regularization term, respectively, which are defined as follows

$$R_t(\mathbf{W}) = \sum_{h=1}^{H} \lambda_h \sum_{p<q}^{M} \|\mathbf{w}_{h,p} - \mathbf{w}_{h,q}\|_2 \tag{3}$$

and

$$R_s(\mathbf{W}) = \sum_{h=1}^{H} \beta_h \|\mathbf{W}_h\|_{1,1} \tag{4}$$

In Eq. (3), we impose a ℓ_2-norm on the pairwise difference among the column vectors, which encourages each pair of columns to share some smaller group features. In Eq. (4), $\|\mathbf{W}_h\|_{1,1} = \sum_{i=1}^{d} \|\mathbf{w}_h^i\|_1$ is the sum of ℓ_1-norm of the rows in matrix $\mathbf{W}_h$. The $\ell_{1,1}$-norm is used to encourage sparsity of the coefficient matrix $\mathbf{W}_h$ as well as select the specific features corresponding to each frequency band. And λ_h and β_h are positive constants used to balance the feature sharing and specific features selection. Particularly, λ_h controls the strength of sharing features and β_h controls the sparsity of the coefficient matrix. In this study, we assume a descending order for information sharing from the high-level to the first one, however the sparsity of the matrix $\mathbf{W}_h$ is gradually increasing from the low level to the high one. Hence, we set $\lambda_h = \lambda_{h-1}/\sigma$ and $\beta_h = \beta_{h-1} \times \sigma$ for $h \geq 2$ with constant $\sigma > 1$. It is worth noting that, when $\beta_1 = 0$, our method will reduce to the multi-level task grouping method (MLMT-T) [6]. Also, when $\lambda_1 = 0$, our method will reduce to a multi-level Lasso method (MLMT-S). Below, we will develop a new method for optimizing the objective function in Eq. (2).

Optimization Algorithom: To optimize the objective function in Eq. (2), we propose a top-down iterative scheme, where the problem (2) is decomposed into several sub-problems consistent with the levels, which is described as follows:

$$min_{\mathbf{W}_h} \sum_{m=1}^{M} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}_m \mathbf{w}_{h,m}\|_2^2 + \lambda_h \sum_{p<q}^{M} \|\mathbf{w}_{h,p} - \mathbf{w}_{h,q}\|_2 + \beta_h \|\mathbf{W}_h\|_{1,1} \quad (5)$$

For seeking the optimal value of $\mathbf{W}_h$ of the sub-problems which corresponds to the h-th level, we can solve the problem by the smoothing proximal gradient (SPG) method [6] fortunately. The problem solved by the SPG method takes the form

$$min_{\mathbf{W}_h} \widetilde{F}(\mathbf{W}_h) = f(\mathbf{W}_h) + g_\mu(\mathbf{W}_h) + \beta_h \|\mathbf{W}_h\|_{1,1} \quad (6)$$

According to [6], we can rewrite the smoothed approximation function of $g_\mu(\cdot)$ as

$$g_\mu(\mathbf{W}_h) = \lambda_h \sum_{p<q}^{M} \|\mathbf{w}_{h,p} - \mathbf{w}_{h,q}\|_2 = max_{\mathbf{A}\in\boldsymbol{Q}} \langle \mathbf{C}\mathbf{W}_h^T, \mathbf{A}\rangle - \mu d(\mathbf{A}) \quad (7)$$

where $\mathbf{C} \in \mathbb{R}^{\frac{m(m-1)}{2}\times m}$ is a sparse matrix with each row having only two non-zero entries, and $\mathbf{A} = (\alpha_1, \ldots, \alpha_{m(m-1)/2})^T$ is the auxiliary matrix variable with a closed and convex set domain $\boldsymbol{Q} \equiv \{\mathbf{A} | \|\alpha_j\|_2 \leq 1, \forall_j \in \mathbb{N}_{m(m-1)/2}\}$. Then the gradient of $h(\mathbf{W}_h) = f(\mathbf{W}_h) + g_\mu(\mathbf{W}_h)$ with respect to $\mathbf{W}_h$ can be computed as

$$\nabla_{\mathbf{W}_h} h(\mathbf{W}_h) = \mathbf{X}_m^T(\mathbf{X}_m \mathbf{w}_{h,m} - \mathbf{Y}) + (\mathbf{A}^*)^T \mathbf{C} \quad (8)$$

Hence, the generalized gradient update step of SPG algorithm is defined as

$$\mathbf{W}_h^{t+1} = argmin_{\mathbf{W}_h} \frac{1}{2}\|\mathbf{W}_h - (\widehat{\mathbf{W}}_h^t - \frac{1}{L}\nabla h(\widehat{\mathbf{W}}_h^t))\|_2^2 + \frac{\beta_h}{L}\|\mathbf{W}_h\|_{1,1} \quad (9)$$

where L is the Lipschitz constant which can be determined by numerical approaches, and $\mathbf{V}_h = \widehat{\mathbf{W}}_h^t - \frac{1}{L}\nabla h(\widehat{\mathbf{W}}_h^t)$. According to [6], the closed-form solution for $\mathbf{W}_h^{t+1}$ is given as $\mathbf{W}_h^{t+1,i} = sign(\mathbf{v}^i) max(0, |\mathbf{v}^i| - \frac{\beta_h}{L})$, where $\mathbf{W}_h^{t+1,i}$ and $\mathbf{v}^i$ represent the i-th row of the matrix $\mathbf{W}_h^{t+1}$ and $\mathbf{V}_h$. In addition, according to [6], instead of performing gradient descent based on $\mathbf{W}_h$, we can compute the following formulation as:

$$\widehat{\mathbf{W}}_h^{t+1} = \mathbf{W}_h^{t+1} + \eta_{t+1}(\mathbf{W}_h^{t+1} - \widehat{\mathbf{W}}_h^t) \quad (10)$$

where $\eta_{t+1} = \frac{(1-\theta_t)\theta_{t+1}}{\theta_t}$ and $\theta_{t+1} = \frac{2}{t+3}$. Let $D = max_{\mathbf{A}\in\boldsymbol{Q}} d(\mathbf{A})$ and $\mathbf{W}_h^*$ be the optimal solution to Eq. (5). If the desired accuracy is ϵ, according to [6], the SPG algorithm needs $O(\sqrt{2D}/\epsilon)$ iterations to converge.

3 Experiments

Experimental Setting: For method evaluation, we adopted leave-one-out cross-validation (LOOCV) to estimate the classification performances of different methods. All parameters were learned by conducting LOOCV. Four

classification performance including classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and the area under the receiver operating characteristic (ROC) curve (AUC) were used. We compared the proposed method with five different feature selection methods, including t-test, Lasso, MTL, and two variants of the proposed method. The t-test and Lasso were used as feature selection methods in each frequency band while the MTL and the variants of our method learned different frequency bands jointly. After that, the selected features were feed to the SVM with ensemble strategy in the final classification step. For all methods, the linear SVM implemented in LibSVM software[2] was used as the classifier. The parameters λ_1 and β_1 in Eq. (2) and two variants were chosen from $\{2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3\}$, while the constant σ was set to 1.2 as used in [6]. For the t-test method, the p-value was chosen from $\{0.05, 0.02, 0.01\}$. The parameters for Lasso and MTL were also chosen from $\{2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3\}$. The MTL method was implemented using SLEP package[3]. The hierarchy of our method is chosen by LOOCV on the training data from a set $[1, \ldots, 10]$.

Results and Disscussions: To illustrate the effectiveness of the proposed hierarchical structure, Fig. 2 shows the change of ACC and AUC with different numbers of hierarchies on the training data. It is observed that at the beginning the use of more hierarchies benefits the classification performance. When hierarchy reaches five, the method achieves the best performance. Then, the performance becomes slightly worse for larger hierarchies. Because the upper hierarchies learned from the proposed method tend to have a larger number of small groups structure due to the setting for the regularization parameter $\lambda_h = \lambda_{h-1}/\sigma$. And with the increases of hierarchy, all the hierarchies will contain more group structure, which leads to capture more noisy information in multi-frequency data. Therefore, we set the number of hierarchies to five in our experiments.

We also show in Table 1 a comparsion of the proposed method with the other competing methods and their ROC curves are given in Fig. 3. From the results of SZ vs. NC classification in Table 1 and Fig. 3, we can observe three main points. First, hierarchical methods generally achieve significantly better performance, compared with the single level methods. It is worth noting that, although MLMT-S method has not obtained the best performance in comparison with MTL, it still outperforms the t-test and Lasso methods. In contrast to MTL, MLMT-S only considers a single band and ignores the structure information. It also reveals that taking the different frequency bands into account is beneficial to improve predictions. Second, our proposed method which takes advantage of the structure information performs much better than the other methods without consider structure information in terms of classification accuracy, sensitivity and AUC. Specifically, MLMT-TS achieves a classification accuracy of 93.48%, a sensitivity of 95.65% and an AUC of 97.07%, while the best results of MLMT-S

[2] https://www.csie.ntu.edu.tw/ cjlin/libsvm/
[3] http://www.yelab.net/software/SLEP/

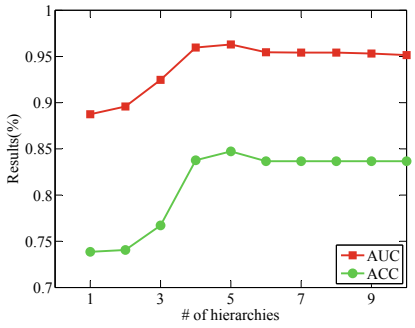


Fig. 2. Effects of using different numbers of hierarchies

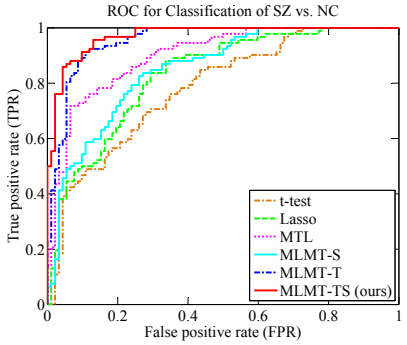


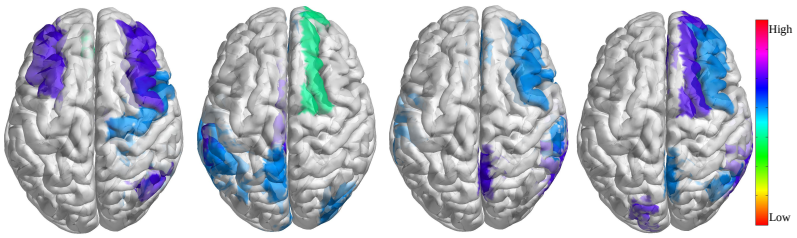
Fig. 3. ROC curves of different methods in SZ vs. NC classification

Table 1. Classification performances with LOOCV of different methods

Methods	ACC (%)	SEN (%)	SPN (%)	AUC (%)
t-test	63.04	73.91	52.17	78.07
Lasso	71.74	65.22	78.26	83.02
MTL	80.43	82.61	78.26	90.06
MLMT-S	73.91	82.61	65.22	85.01
MLMT-T	91.30	91.30	**91.30**	96.44
MLMT-TS (ours)	**93.48**	**95.65**	**91.30**	**97.07**

is 73.91%, 82.61% and 85.01%, respectively. Finally, MLMT-TS is slightly better than the MLMT-T method, which shows that specific features corresponding to each frequency band are significant to improve the classification performance.

The main finding of this study is the different contributions of LFO amplitudes in the classification of schizophrenia. Fig. 4 depicts the identified biomarkers from the four frequency bands. The color of these biomarkers indicates the contribution for identifying the schizophrenia. Specifically, key regions of cognitive control networks, including the *prefrontal cortex* and *anterior cingulate cortex* are found to contribute high weight for identifying schizophrenia in all frequency bands. For example, the *gyrus frontalis medius* region is selected in slow-5, slow-3 and slow-2. While core default mode network regions, which include *precuneus*, *angular gyrus*, and *memory associated temporal regions*, take a central position in the classification of schizophrenia and patients within specificity but not all frequency bands. For example, in slow-2 default mode network core regions, like *precuneus*, *angular gyrus* and *middle temporal* have high weights, while for slow-3 regions in the *frontotemporal network* would be the core biomarkers in classification. The results validate that taking both the specificity and complementary information into account is helpful to improve the classification performance. And the effectiveness of the selected biomarkers can be confirmed by previous reports in the literature [3, 4, 7].



(a) slow-5 (b) slow-4 (c) slow-3 (d) slow-2

Fig. 4. The biomarkers identified by the proposed MLMT-TS method in different frequency band.

4 Conclusion

In this study, we have developed a multi-level multi-task structured sparse (MLMT-TS) feature selection framework for schizophrenia diagnosis, which can make better use of the underlying specificity and structure information of multi-frequency bands data. Experimental results on the multi-frequency bands schizophrenia dataset showed that the hierarchical scheme was able to gradually refine the information sharing in multiple steps. Compared with other methods, consistently high performance demonstrates the efficacy of our proposed method.

References

1. Bhugra, D.: The global prevalence of schizophrenia. PLOS Medicine **2**(5) (2005) 372–373
2. Yan, T., Wang, L., Fang, C., Tan, L.: Identify schizophrenia using resting-state functional connectivity: an exploratory research and analysis. BioMedical Engineering OnLine **11**(50) (2012) 1–16
3. Yu, R., Chien, Y.L., Wang, H.L., Liu, C.M., Liu, C.C., Hwang, T.J., Hsieh, M.H., Hwu, H.G., Tseng, W.Y.: Frequency-specific alternations in the amplitude of low-frequency fluctuations in schizophrenia. Human Brain Mapping **35**(2) (2014) 627–637
4. Zuo, X.N., Di, M.A.: The oscillating brain: complex and reliable. Neuroimage **49**(2) (2010) 1432–1445
5. Zhang, D., Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimers disease. Neuroimage **59**(2) (2012) 895–907
6. Han, L., Zhang, Y.: Learning multi-level task groups in multi-task learning. In: AAAI 2015, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, USA, January. (2015) 2638–2644
7. Marsman, A., Mandl, R.C., Mp, V.D.H., Boer, V.O., Wijnen, J.P., Klomp, D.W., Luijten, P.R., Hilleke E, H.P.: Glutamate changes in healthy young adulthood. European Neuropsychopharmacology **23**(11) (2013) 1484–1490