Multi-task Sparse Low-Rank Learning for Multi-classification of Parkinson's Disease

Haijun Lei[1(✉)], Yujia Zhao[1], and Baiying Lei[2]

[1] College of Computer Science and Software Engineering, Guangdong Province Key Laboratory of Popular High-Performance Computers, Shenzhen University, Shenzhen, China
leiby@szu.edu.cn

[2] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China

Abstract. Identifying prodromal stages of Parkinson's disease (PD) draws increasing recognition as non-motor symptoms may appear before classical clinical diagnosis based on motor signs. To effectively develop a computer-aided diagnosis for multiple disease progression stages, neuroimaging has been widely applied for its convenience of revealing the intricate brain structure. However, the high dimensional neuroimaging features and limited sample size bring the main challenges for the diagnosis task. To handle it, a multi-task sparse low-rank learning framework is proposed to unveil the underlying relationships between input data and output targets by building a matrix-regularized feature network. Inductions of multiple tasks are simultaneously performed to capture intrinsic feature relatedness with multi-task learning. By discarding the irrelevant features and preserving the discriminative structured features, our proposed method can select the most relevant features and identify different stages of PD with different multi-classification models. Extensive experimental results on the Parkinson's progression markers initiative (PPMI) dataset demonstrate that the proposed method achieves promising classification performance and outperforms the conventional algorithms.

Keywords: Multi-task learning · Low-rank · Sparse learning
Parkinson's disease

1 Introduction

PD has gained increasing attention as the growing aging problem of the population. The chronic progression nature and imperceptible neuro-diminishment of PD make the treatment comparatively difficult [1]. There is suggestive evidence that olfaction changes, sleep behavior disorder, subtle cognitive changes and depression can be present at early PD stages, suggesting high potential of having PD [2]. Before the occurrence of motor symptoms permits the clinical diagnosis of PD, about or above 50% of the dopaminergic neurons of the substantia nigra have degenerated. The time span between the onset of neurodegeneration and manifestation of the typical motor

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 361–368, 2018.
https://doi.org/10.1007/978-3-030-00889-5_41

symptoms is referred as prodromal phase of PD (PROD) [3]. The term SWEDD (scans without evidence for dopaminergic deficit) refers to the absence of an imaging abnormality in patients clinically presumed to have PD [4]. PROD and SWEDD are different disorders of PD, whose patients require targeted treatment. Therefore, early PD diagnosis offers timely prevention treatment of the patients.

Using the rich information of neuroimaging techniques, we can monitor the minor neuro changes, which are not easy to perceive in normal clinical symptom-based diagnosis. Common neuroimaging techniques include magnetic resonance imaging (MRI), diffusion-weighted tensor imaging (DTI). Recently, many machine learning methods have been applied to utilize the neuroimages in the computer-aided diagnosis of neurodegenerative disease. A robust feature-sample selection scheme was developed for PD diagnosis [5]. Due to the challenges of high dimensionality and limited sample size, the overfitting problem could be occurred in the data analysis. Recent studies have demonstrated that feature selection is capable of overcoming this issue. A l_1-regularizer (i.e., a sparse term) is introduced in the estimation model for feature selection when the sample size is significantly smaller than the feature dimension [6]. However, sparsity regularization is insufficient in multi-classification application since there are four progressive classification targets: normal control (NC), SWEDD, PROD and PD.

In fact, the relationship between input data (i.e., MRI images) and output targets (i.e., prediction results) have more to explore. Inspired by the fact that the brain is organized with modular structures, we intend to find the most representative features to train our multi-class classifiers by extracting the low-rank structure of the matrix-regularized feature network as well as its sparseness.

On the other hand, gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) are the most significant biomarkers in the brain which are later used as features. The conventional feature extraction methods apply a simple linear combination to use the three matters without considering their own contributing factor. We model this problem as a multi-task learning framework by proposing a model that efficiently leverages the multi-modal data [7]. Our model considers the multi-classification of disease stages using each modal as one task. We assume that these tasks are related and can benefit each other for the classification purpose. Then we perform the three tasks simultaneously to capture their intrinsic relatedness to achieve better classification performance.

Moreover, clinical symptoms have been considered as a vital indicator of PD diagnosis. The judgement results of clinicians are reflected on the clinical assessment scores for each potential PD patient. The combination of constructive information with the neuroimaging information provides sufficient information for computer-aided analytical diagnosis. For this reason, we propose a multi-task sparse low-rank learning (MSLRL) framework for multi-classification of PD. The proposed MSLRL framework combines the sparsity and low-rank constraints together for each task to select the most PD related features. To the best of our knowledge, this is the first work to introduce multi-task sparse low-rank learning to PD diagnosis using neuroimages. Experimental results demonstrate the prominent performance of our proposed method on the PPMI dataset.

2 Method

The proposed method intends to find a subset of features that are most related to PD. The multi-task sparse low-rank learning framework is shown in Fig. 1. We extract our feature input data from MRI images. In order to predict the accurate labels, we add a low-rank and sparse constraint to the matrix-regularized feature network and extract the respective weighted significance by clustering for each task. Each task applies the same feature selection method in a jointly multi-task framework. The shared weight matrix leads to the selected features with reduced dimensions to train a support vector machine (SVM) based classifiers.

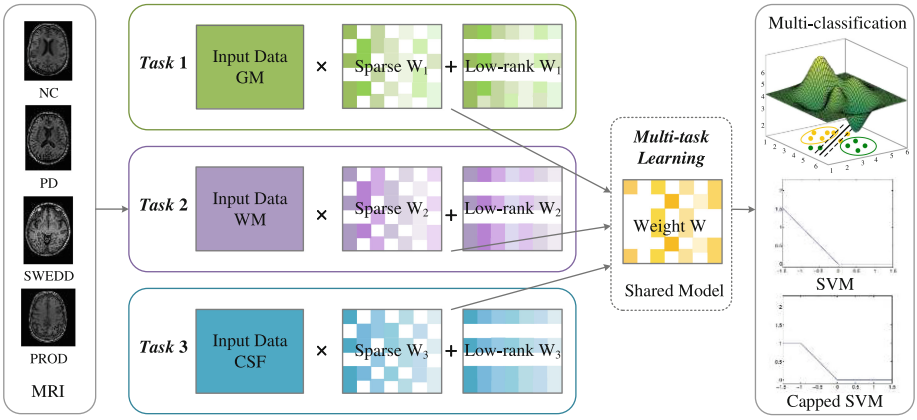


Fig. 1. Flowchart of our proposed MSLRL method. The shared model is learned from the multi-task learning by considering each tissue modal as task.

Supposing that we have m subjects and each has n features belong to k tasks. In the linear regression model $\mathbf{Y}^{(i)} = \mathbf{X}^{(i)}\mathbf{W}^{(i)}$, $\mathbf{Y}^{(i)} \in \mathbb{R}^{m\times 1}$ is the ground truth label vector of i-th task, $\mathbf{X}^{(i)} \in \mathbb{R}^{m\times n}$ is the input data matrix of i-th task, and $\mathbf{W}^{(i)} \in \mathbb{R}^{n\times 1}$ is the weight coefficient matrix for each feature of i-th task. We can get $\mathbf{W}^{(i)}$ by solving the following objective function:

$$\min_{\mathbf{W}^{(i)}} \left\|\mathbf{Y}^{(i)} - \mathbf{X}^{(i)}\mathbf{W}^{(i)}\right\|_F^2, \tag{1}$$

where $\|\mathbf{A}\|_F$ is the Frobenius norm (F-norm) of $\mathbf{A}$ which is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_i \|\mathbf{A}_i\|_2^2}$, where $\mathbf{A}_i$ is the row vector. F-norm also known as the l_2-norm or the l_2-regularizer. Equation (1) is a simple and straightforward linear regression model without constraint on any variable. However, it does not consider the properties of weight matrix, which result in inferior performance. In most machine learning applications, over-fitting is a common problem when the data matrix is unbalanced. Especially in the field of neuroimaging-aided diagnosis, the brain images are rare, and yet

they provide extensive information, leading to high dimensionality. A sparse term like l_1-regularizer is generally adopted to regulate the weight matrix by setting certain entries to zero for sparseness. Let $\|\mathbf{A}\|_1$ be the l_1-norm of $\mathbf{A}$ and is defined as $\|\mathbf{A}\|_1 = \sum_{i=1}^{N} |\mathbf{A}_i|$, we can formulate the objective function using sparse representation as:

$$\min_{\mathbf{W}^{(i)}} \left\|\mathbf{Y}^{(i)} - \mathbf{X}^{(i)}\mathbf{W}^{(i)}\right\|_F^2 + \lambda \left\|\mathbf{W}^{(i)}\right\|_1, \tag{2}$$

Equation (2) selects the most representative features under the assumption of sparsity of $\mathbf{W}^{(i)}$ and constraint of the first data-fitting term. In the model, we aim to find a weight matrix that represents the feature significance. We further explore the low-rank structure between features. It is well-known that, the brain is divided into different parts known as regions of interest (ROIs), we extract different features from these regions. Since PD is one category of neurodegenerative disease, it is influenced by a block of brain regions that are responsible for certain human actions or emotions. For this reason, we assume that a group of features are dependent on each other, leading to a low-rank structure of the coefficient weight matrix because certain rows are dependent. The sparse low-rank learning framework for each task is built on the assumption that, features are closely related with group of features while the relevance between these groups may be sparse. Multiple tasks share the same low-rank and sparse weight coefficients. Thus, the objective function for each task is reformulated as:

$$\min_{\mathbf{W}^{(i)}} \left\|\mathbf{Y}^{(i)} - \mathbf{X}^{(i)}\mathbf{W}^{(i)}\right\|_F^2 + \lambda_1 \left\|\mathbf{W}^{(i)}\right\|_1 + \lambda_2 rank\left(\mathbf{W}^{(i)}\right), \tag{3}$$

where $rank\left(\mathbf{W}^{(i)}\right)$ is the rank function of $\mathbf{W}^{(i)}$. Low-rank learning has been utilized in matrix recovery and network modeling. The weight matrix $\mathbf{W}^{(i)}$ in Eq. (3) has dimension of n rows representing the respective feature significance. The rank minimization of $\mathbf{W}^{(i)}$ explores the low-rank structure among features to obtain the intrinsic relationship. However, it is difficult to solve $\mathbf{W}^{(i)}$ since the *rank* function is non-convex and the rank minimization is a NP-hard problem. Recently, researchers have proved that trace norm function is the convex envelop of the *rank* function over the domain $\left\|\mathbf{W}^{(i)}\right\|_2 \leq 1$, which provides the lowest bounds of the rank function *rank* [11]. The trace norm $\|\mathbf{W}\|_*$ is defined as:

$$\|\mathbf{W}\|_* = \sum\nolimits_{i=1}^{\min\{n,k\}} \sigma_i = Tr\left(\left(\mathbf{W}^T\mathbf{W}\right)^{\frac{1}{2}}\right), \tag{4}$$

where σ_i is the i-th singular value of $\mathbf{W}$ and can be obtained by singular value decomposition (SVD). Thus, we can establish the final objective function with a l_1-norm $\|\mathbf{W}\|_1$ and a trace norm $\|\mathbf{W}\|_*$ as:

$$\min_{\mathbf{W}} \sum\nolimits_{i=1}^{k} \left\|\mathbf{Y}^{(i)} - \mathbf{W}^{(i)}\mathbf{X}^{(i)}\right\|_F^2 + \alpha\|\mathbf{W}\|_1 + \beta\|\mathbf{W}\|_*, \tag{5}$$

where α and β are the parameters controlling the sparse degree and the low-rank degree, respectively. When $\alpha = 0$, Eq. (5) has only the low-rank constraint. When we add a l_2-norm $\|\mathbf{W}\|_2$ to Eq. (2), we can get the standard elastic net formulation. Moreover, if we change the l_1-norm $\|\mathbf{W}\|_1$ in Eq. (2) to $l_{2,1}$-norm $\|\mathbf{W}\|_{2,1}$, we can get the classic least absolute shrinkage and selection (LASSO).

For optimization for Eq. (5), we notice that, the l_1-norm and trace norm are non-differentiable. Thus, we solve $\mathbf{W}$ using the proximal gradient descent method due to its effectiveness in solving l_1-norm involved equations. Since we have three terms in Eq. (5), we update $\mathbf{W}$ by the value of each term. First, we find the proximal operator of $\alpha\|\mathbf{W}\|_1$ according to:

$$prox_{\alpha\|\cdot\|_1}(\mathbf{W}) = \left[sign\left(w_{ij}\right) \cdot \max\left\{\left|w_{ij}\right| - \alpha, 0\right\}\right]_{n\times k}, \tag{6}$$

where $prox()$ denotes the proximal operator and $sign()$ is the sign function. Similarly, we can obtain the proximal operator of $\beta\|\mathbf{W}\|_*$ using:

$$prox_{\beta\|\cdot\|_*}(\mathbf{W}) = \mathrm{U}diag(\max\{\widehat{\sigma}_1, 0\}, \cdots, \max\{\widehat{\sigma}_l, 0\})\mathrm{V}^{\mathrm{T}}, \tag{7}$$

where U is the unitary matrix in the SVD of $\mathbf{W}$ so that $\mathbf{W} = \mathrm{U}diag(\sigma_1, \cdots, \sigma_l)\mathrm{V}^{\mathrm{T}}$ with $\widehat{\sigma}_i = \sigma_i - \beta$ and $l = \min\{n, k\}$. Then, we consider the first data-fitting term $\left\|\mathbf{Y}^{(i)} - \mathbf{W}^{(i)}\mathbf{X}^{(i)}\right\|_F^2$. Given $f_1\left(\mathbf{W}^{(i)}\right) = \left\|\mathbf{Y}^{(i)} - \mathbf{W}^{(i)}\mathbf{X}^{(i)}\right\|_F^2$, we can get the derivative of $\mathbf{W}^{(i)}$ as $\nabla f\left(\mathbf{W}^{(i)}\right) = \mathbf{X}^{(i)\mathrm{T}}\mathbf{X}^{(i)}\mathbf{W}^{(i)} - \mathbf{X}^{(i)\mathrm{T}}\mathbf{Y}^{(i)}$. Consequently, we can solve $\mathbf{W}$ by iteratively updating the values until convergence.

3 Experiments

3.1 Experimental Settings

We validate our method by classifying different stages of PD subjects. We choose SVM classifiers to construct a multi-class classification model for its efficiency in separating different class samples with the maximum margin [8]. Another classifier we apply is the capped l_p-norm SVM [9]. This upgraded classifier can deal with both light and heavy outliers, boosting classification performance. The main parameters used are α and β in Eq. (5), where α controls the sparse term $\|\mathbf{W}\|_1$ and β controls the low-rank term $\|\mathbf{W}\|_*$, respectively. The initial values are set as $\alpha \in \left\{2^{-5}, \ldots, 2^5\right\}$, $\beta \in \left\{2^{-5}, \ldots, 2^5\right\}$. The fine-tuned parameter values are specified by a 5-fold cross-validation strategy. The results are evaluated using: accuracy (ACC), sensitivity (SEN), specificity (SPEC), and area under the receiver operating characteristic curve (AUC). For fair evaluation, the classification performance of the proposed method is evaluated via a 10-fold cross-validation strategy.

3.2 Data Preprocessing

The data used in this experiment are MRI images from the PPMI dataset. All the original images are preprocessed by the anterior commissure-posterior commissure correction and skull-stripping for later operation. Then we segment the images into GM, WM, and CSF using Statistical Parametric Mapping (SPM) [10]. Following the automated anatomical labeling atlas which parcel brain into 116 regions, we compute the mean tissue density value of each region as features. In this work, we collect 643 subjects (127 NC, 380 PD, 56 SWEDD and 34 PROD). For each subject, the feature dimension is 116 for each tissue modal (116 GM, 116 WM, 116 CSF). Apart from these features, we also collect four clinical scores, namely, sleep scores, olfaction scores, depression scores, and Montreal cognitive assessment scores as features. Theses clinical scores are the clinical assessment results from the clinicians' experience and diagnosis. With the guidance of these clinical scores as features, we can build a more reliable classification model.

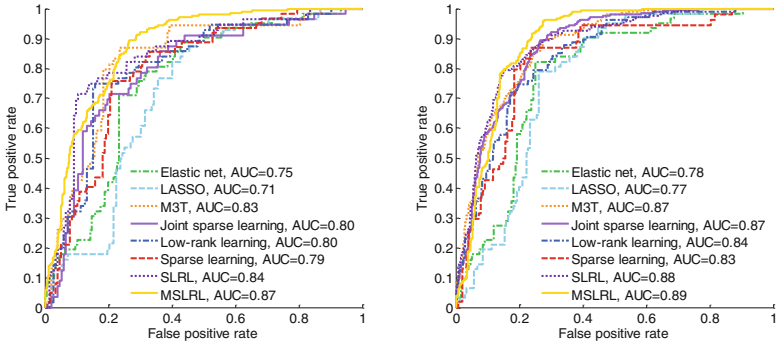
3.3 Classification Performance

To further validate the effectiveness of our MSLRL method, we compare the method with other similar methods. Apart from the elastic net and LASSO methods, we further compare MSLRL with another two sparsity-based methods. One is multi-modal multi-task (M3T) [11] and the other is joint sparse learning [12]. Furthermore, we additionally compare MSLRL with low-rank learning and sparse learning and sparse low-rank learning (SLRL). The classification performance results are summarized in Table 1. It is clear that, the MSLRL method achieves higher accuracy than classical

Table 1. Classification performance of all competing methods with different classifiers.

Method	Classifier	ACC	SEN	SPEC	AUC
Elastic net	SVM	67.84	73.17	84.11	86.23
	Capped SVM	68.66	74.31	84.66	86.93
LASSO	SVM	65.27	73.45	85.23	84.65
	Capped SVM	65.68	74.92	86.30	85.17
M3T	SVM	74.55	80.05	94.05	88.23
	Capped SVM	75.81	81.55	97.45	89.45
Joint sparse learning	SVM	72.10	75.24	85.38	87.54
	Capped SVM	73.46	78.20	87.79	89.07
Low-rank learning	SVM	72.32	73.01	88.68	88.78
	Capped SVM	73.06	78.52	90.03	89.92
Sparse learning	SVM	70.63	77.19	87.07	87.45
	Capped SVM	71.88	77.85	87.93	88.97
SLRL	SVM	75.23	84.21	93.86	90.24
	Capped SVM	77.87	85.98	95.47	92.77
MSLRL	SVM	78.76	84.62	98.32	92.21
	Capped SVM	**79.49**	**87.24**	**99.21**	**94.31**

Elastic net and LASSO as well as sparse-based M3T and joint sparse learning using both SVM classifiers. SLRL turns out to be more effective than low-rank learning and sparse learning, which validates the strategy of combining l_1-norm $\|\mathbf{W}\|_1$ and trace norm $\|\mathbf{W}\|_*$ using sparsity and low-rank structure. MSLRL outperforming SLRL in both classifiers, which proves that multi-task learning successfully explores the intrinsic relation within multi-modal features. Receiver operating characteristic curves (ROC) for algorithm comparison are shown in Fig. 2. MSLRL obtains the best performance in all competing methods in each classifier, which shows the advantage and potential for early PD diagnosis.

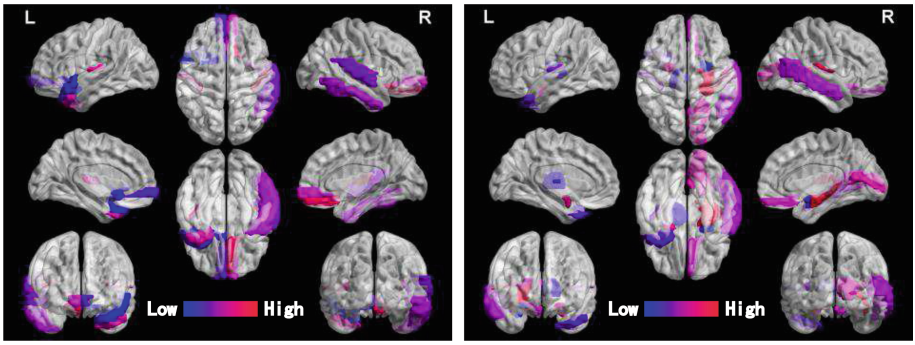


a. SVM b. Capped SVM

Fig. 2. ROC plots of the competing methods using two classifiers (SVM and Capped SVM).

3.4 Most Distinctive Brain Regions

The identification of PD-related features and the monitoring of progression are of great significance in early diagnosis. We utilize the weight coefficient matrix generated in feature selection to study the discriminative brain regions most related to PD. The regions most related with PD are visualized in Fig. 3. The selected brain regions are



a. SLRL b. MSLRL

Fig. 3. Top 10 discriminative brain regions obtained from SLRL and MSLRL. Brain regions are color-coded. High means high relevance with PD. Low means relatively low relevance with PD. (Color figure online)

slightly different in two methods. The higher relevance of MSLRL than SLRL reveals that MSLRL is more effective than SLRI for PD diagnosis. These distinctive brain regions can be further investigated for clinical practice.

4 Conclusion

In this paper, we introduce a multi-task sparse low-rank learning framework for early PD diagnosis between four progression stages. Specifically, for each task we add the sparsity and low-rank regularization to the weight coefficients with a l_1-norm and a trace norm to unveil the underlying relationships within data. By exploring the intrinsic relationships between multiple tasks, this framework can select the most representative features by jointly considering the dimension reduction of neuroimaging feature vectors and the relevant dependency properties of PD-related brain region features. Using multi-modal data from PPMI neuroimaging dataset, experiments demonstrate that our method has the best multi-class classification results among all the traditional methods.

References

1. Simons, J.A., Fietzek, U.M., Waldmann, A., Warnecke, T., Schuster, T., Ceballos-Baumann, A.O.: Development and validation of a new screening questionnaire for dysphagia in early stages of Parkinson's disease. Park. Relat. Disord. **20**(9), 992–998 (2014)
2. Postuma, R.B., et al.: Identifying prodromal Parkinson's disease: pre-motor disorders in Parkinson's disease. Mov. Disord. Off. J. Mov. Disord. Soc. **27**(5), 617–626 (2012)
3. Gaenslen, A., Swid, I., Liepelt-Scarfone, I., Godau, J., Berg, D.: The patients' perception of prodromal symptoms before the initial diagnosis of Parkinson's disease. Mov. Disord. Off. J. Mov. Disord. Soc. **26**(4), 653–658 (2011)
4. Erro, R., Schneider, S.A., Quinn, N.P., Bhatia, K.P.: What do patients with scans without evidence of dopaminergic deficit (SWEDD) have? New evidence and continuing controversies. J. Neurol. Neurosurg. Psychiatry (2015)
5. Adeli, E., et al.: Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data. NeuroImage **141**, 206–219 (2016)
6. Peng, J., An, L., Zhu, X., Jin, Y., Shen, D.: Structured sparse kernel learning for imaging genetics based alzheimer's disease diagnosis. In: MICCAI, pp. 70–78 (2016)
7. Zhou, J., Chen, J., Ye, J.: Multi-task learning: theory, algorithms, and applications. SDM Tutor. (2012)
8. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
9. Nie, F., Wang, X., Huang, H.: Multiclass capped p-Norm SVM for robust classifications. In: AAAI, pp. 2415–2417 (2017)
10. Friston, K.J.: Statistical parametric mapping (1994)
11. Zhang, D., Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage **59**(2), 895–907 (2012)
12. Lei, H., et al.: Joint detection and clinical score prediction in Parkinson's disease via multi-modal sparse learning. Expert Syst. Appl. **80**(1), 284–296 (2017)