Sparse Discriminative Feature Selection for Multi-class Alzheimer's Disease Classification

Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen*

Department of Radiology and BRIC,
University of North Carolina at Chapel Hill, USA
dgshen@med.unc.edu

Abstract. In neuroimaging studies, high dimensionality and small sample size have been always an issue, and it is common to apply a dimension reduction method to avoid the over-fitting problem. Broadly, there are two different approaches in reducing the feature dimensionality: feature selection and subspace learning. When it comes to the feature interpretability, the feature selection approach such as the sparse regularized linear regression method is preferable to the subspace learning methods, especially in Alzheimer's Disease (AD) diagnosis. However, based on recent machine learning researches, the subspace learning methods presented promising results in various applications. To this end, in this work, we propose a novel method for discriminative feature selection by combining two conceptually different methodologies of feature selection and subspace learning in a unified framework. Specifically, we integrate the ideas of Fisher's linear discriminant analysis and locality preserving projection, which consider, respectively, the global and local information inherent in observations, in a regularized least square regression model. With the help of global and local information in data, we select class-discriminative and noise-resistant features that thus help enhance classification performance. Furthermore, unlike the previous methods that mostly considered only a binary classification, in this paper, we consider a multi-class classification problem in AD diagnosis. Our experiments on the Alzheimer's Disease Neuroimaging Initiative dataset showed the efficacy of the proposed method by enhancing the performances in multi-class AD classification.

1 Introduction

Previous studies of the computer-aided Alzheimer's Disease (AD) diagnosis usually applied the sequential processes of feature extraction, feature dimensionality reduction, and classifier learning, to make a decision on the clinical status of a subject, *e.g.,* AD, Mild Cognitive Impairment (MCI), and Normal Control (NC) [4,14,16,17,20]. In this paper, we focus on the feature selection, which has the effect of lowering feature dimensionality. Furthermore, unlike the previous methods that mostly considered only binary classification of either AD vs. NC or MCI

* Corresponding author.

G. Wu et al. (Eds.): MLMI 2014, LNCS 8679, pp. 157–164, 2014.
© Springer International Publishing Switzerland 2014

vs. NC, we consider a multi-class classification problem, *e.g.,* AD vs. MCI vs. NC, for practical applications. Based on the observation that there are three or four different clinical status related to AD, *i.e.,* AD, MCI (MCI-Converter: MCI-C, MCI-NonConverter: MCI-NC), and NC, from a clinical point of view, it is more practical to build a multi-class classifier.

In neuroimaging studies, while the feature dimension is high in nature, the available sample size is very limited. It has been always an issue for high dimensionality and small sample size in computer-aided AD diagnosis [5,13,22,23]. Thus, dimensionality reduction by means of either subspace learning or feature selection has been one of the core steps in neuroimaging pattern analysis. Methodologically, feature selection methods, *e.g., t*-test and sparse regularized linear regression, select an informative feature subset from the original feature set, while the subspace learning methods, *e.g.,* Fisher's Linear Discriminant Analysis (LDA) [3] and Locality Preserving Projection (LPP) [7], transform the original feature space into a low-dimensional space. As for the interpretability of the results, the feature selection methods are preferable to subspace learning methods, in particular, in neuroimaging studies. However, according to recent studies in machine learning [6,18,19], subspace learning has shown promising performances in various fields.

In this paper, we propose a novel method that efficiently combines the methodologies of feature selection and subspace learning. Specifically, we inject the ideas of two subspace learning methods, *i.e.,* LDA and LPP, into a sparse least square regression framework. The rationale of using both LDA and LPP in our formulation is that LDA considers the global information inherent in the observations with the ratio of within-class-variance and between-class-variance, while LPP reflects the local information by means of graph Laplacian. That is, with the help of global and local information in data, we can select class-discriminative and noise-resistant features that thus help enhance classification performances.

2 Proposed Method

2.1 Multi-class Sparse Discriminative Feature Selection

Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ denote a feature matrix, where d and n are, respectively, the numbers of feature variables and samples, and $\mathbf{Y} \in \mathbb{R}^{c \times n}$ denote a class indicator matrix, *e.g.,* 0-1 encoding, where c is the number of classes. We formulate a multi-class feature selection problem by means of a multi-task learning with a sparse least square regression model as follows:

$$\min_{\mathbf{W}} \ \frac{1}{2}\|\mathbf{Y} - \mathbf{W}^T\mathbf{X}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d \times c}$ is a regression coefficient matrix and λ is a sparsity control parameter. The $\ell_{2,1}$-norm $\|\mathbf{W}\|_{2,1}$ penalizes the coefficients in the same row of $\mathbf{W}$ together for joint selection or unselection in regressing the response variables in $\mathbf{Y}$. In Eq. (1), the optimal solution assigns a large weight to the important

features and zero or a small weight to less important features and this method has been successfully applied for a binary classification [10,12,20]. With respect to the multi-task learning, it has been shown that Eq. (1) utilizes the correlation of different classes [1] by regarding each class as one task. However, in its current form, it cannot guarantee the class-discriminative power of the selected features and the preservation of the neighborhood structure of data points, which are important characteristics for a good classification performance [3,6].

In this section, we propose a novel discriminative feature selection method that considers both the *global* data distribution and the *local* topological relation among data in a sparse least square regression framework. We first utilize a Fisher's LDA that considers the global data distribution based on the ratio between within-class-variance and between-class-variance to find the class-discriminative features. Second, we take the concept of an LPP [7] to preserve the topological relation among data.

Regarding the Fisher's criterion for discriminative feature selection, a straightforward approach is to penalize the objective function of Eq. (1) with a regularization term defined as follows:

$$R_G = \frac{\mathbf{W}^T \mathbf{\Sigma}_b \mathbf{W}}{\mathbf{W}^T \mathbf{\Sigma}_w \mathbf{W}} \tag{2}$$

where $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_b$ denote, respectively, the within-class variance and the between-class variance. However, due to the non-convexity of Eq. (2), it is not trivial to find an optimal solution of the objective function. Fortunately, Ye [15] presented that the multi-class LDA that finds a subspace by maximizing Eq. (2) can be equivalently formulated with a linear regression model by defining the class indicator matrix $\mathbf{Y} = [y_{i,k}]$ in Eq. (1) as follows:

$$y_{i,k} = \begin{cases} \sqrt{\frac{n}{n_k}} - \sqrt{\frac{n_k}{n}}, & \text{if } \; l(\mathbf{x}_i) = k \\ -\sqrt{\frac{n_k}{n}}, & \text{otherwise} \end{cases} \tag{3}$$

where $l(\mathbf{x}_i)$ denotes a class label of $\mathbf{x}_i$ and n_k is the sample size of the class k. That is, using a class indicator matrix $\mathbf{Y}$ defined as Eq. (3), we can efficiently use the global information, *i.e.,* data distribution in the original space, without changing the formulation. Importantly, we don't transform the original input feature space into a low-dimensional space, in which it is difficult to interpret or investigate the results.

As for the topological relation among data, *i.e.,* local information, we use a graph Laplacian by defining the similarity $s_{i,j}$ between every pair of data points $\mathbf{x}_i$ and $\mathbf{x}_j$ via a heat kernel[1] and define a regularization term as follows:

$$R_L = tr(\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}) \tag{4}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$ with a similarity matrix $\mathbf{S} = [s_{i,j}] \in \mathbb{R}^{n \times n}$ and a diagonal matrix $\mathbf{D} = [d_{i,i} = \sum_j s_{i,j}] \in \mathbb{R}^{n \times n}$.

[1] $H(\mathbf{x}_i, \mathbf{x}_j) = exp\left[-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}\right]$, where $\sigma \in \mathbb{R}^+$ is a parameter.

Therefore, our final objective function is formulated as follows:

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{Y} - \mathbf{W}^T\mathbf{X}\|_F^2 + \lambda_1 tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) + \lambda_2\|\mathbf{W}\|_{2,1} \tag{5}$$

where $\mathbf{Y}$ is defined as Eq. (3), and λ_1 and λ_2 are tuning parameters. Here, we should note that Eq. (5) efficiently combines the ideas of subspace learning (LDA and LPP) and feature selection in a unified framework.

Our method can be discriminated from the previous methods in the following senses: (1) Unlike the previous sparse linear regression-based feature selection methods [11,21], the proposed method finds the class-discriminative and noise-resistant regression coefficient matrix thanks to the use of the Fisher's criterion and graph Laplacian. (2) Compared to the subspace learning methods such as Principal Component Analysis (PCA), LDA, and LPP, which all have an interpretational limitation, the proposed method selects features in the original space, and thus it has an advantage of intuitive investigation of the results. (3) Unlike the conventional LDA [3] based on the criterion in Eq. (2), the proposed method uses the Fisher's criterion but still operates in the original feature space, and thus allows for an intuitive interpretation of the selected features. Furthermore, while the conventional LDA finds at most $(c-1)$-dimension features for a c-class classification task, *e.g.,* 2-D space in a three-class classification task, Eq. (5) selects at most d features (in general, $d \gg c$ in the AD study).

2.2 Optimization

Eq. (5) is a convex but non-smooth function. In this work, we solve it by designing a new accelerated proximal gradient method [9,19]. We first conduct the proximal gradient method on Eq. (5) by setting

$$f(\mathbf{W}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{W}^T\mathbf{X}\|_F^2 + \lambda_1 tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \tag{6}$$

$$\mathcal{L}(\mathbf{W}) = f(\mathbf{W}) + \lambda_2\|\mathbf{W}\|_{2,1}. \tag{7}$$

Note that $f(\mathbf{W})$ is convex and differentiable, while $\lambda_2\|\mathbf{W}\|_{2,1}$ is convex but non-smooth [9]. To optimize $\mathbf{W}$ with the proximal gradient method, we iteratively update it by means of the following optimization rule:

$$\mathbf{W}(t+1) = \arg\min_{\mathbf{W}} G_{\eta(t)}(\mathbf{W}, \mathbf{W}(t)), \tag{8}$$

where $G_{\eta(t)}(\mathbf{W}, \mathbf{W}(t)) = f(\mathbf{W}(t)) + \langle \nabla f(\mathbf{W}(t)), \mathbf{W} - \mathbf{W}(t)\rangle + \frac{\eta(t)}{2}\|\mathbf{W} - \mathbf{W}(t)\|_F^2 + \lambda_2\|\mathbf{W}\|_{2,1}$, $\nabla f(\mathbf{W}(t)) = (\mathbf{X}\mathbf{X}^T + \lambda_1\mathbf{X}\mathbf{L}\mathbf{X}^T)\mathbf{W}(t) - \mathbf{X}\mathbf{Y}^T$, and $\eta(t)$ and $\mathbf{W}(t)$ are, respectively, a tuning parameter and the value of $\mathbf{W}$ obtained at the t-iteration.

By ignoring the terms independent of $\mathbf{W}$ in Eq. (8), we can rewrite it as

$$\mathbf{W}(t+1) = \pi_{\eta(t)}(\mathbf{W}(t)) = \arg\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{W} - \mathbf{U}(t)\|_2^2 + \frac{\lambda_2}{\eta(t)}\|\mathbf{W}\|_{2,1} \tag{9}$$

where $\mathbf{U}(t) = \mathbf{W}(t) - \frac{1}{\eta(t)}\nabla f(\mathbf{W}(t))$ and $\pi_{\eta(t)}(\mathbf{W}(t))$ is the Euclidean projection of $\mathbf{W}(t)$ onto the convex set $\eta(t)$. Thanks to the separability of $\mathbf{W}(t+1)$ in each row, we can obtain the optimal $\mathbf{W}(t+1)$ by finding a closed form solution of each row [9].

Meanwhile, in order to accelerate the proximal gradient method in Eq. (8), we further introduce an auxiliary variable $\mathbf{V}(t+1)$ as:

$$\mathbf{V}(t+1) = \mathbf{W}(t) + \frac{\alpha(t)-1}{\alpha(t+1)}(\mathbf{W}(t+1) - \mathbf{W}(t)). \tag{10}$$

where the coefficient $\alpha(t+1)$ is usually set as $\alpha(t+1) = \frac{1+\sqrt{1+4\alpha(t)^2}}{2}$ [9].

3 Experimental Analysis

3.1 Dataset and Feature Extraction

We conducted performance evaluation on a subset (202 subjects: 51 AD, 43 MCI Converter: MCI-C, 56 MCI Non-Converter: MCI-NC, and 52 NC) of the ADNI dataset by comparing the proposed method with the competing methods. We considered two multi-class classification problems: AD vs. MCI (including both MCI-C and MCI-NC) vs. NC and AD vs. MCI-C vs. MCI-NC vs. NC. Regarding the feature extraction, we first sequentially performed spatial distortion, skull-stripping, and cerebellum removal for Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) images. For the MRI images, we further segmented them into three tissue types of gray matter, white matter, and cerebrospinal fluid. By warping a template into a subject's brain image, we parcellated the gray matter into 93 Region-Of-Interests (ROIs). The PET images were spatially aligned to its respective MRI images. Finally, we obtained 93 gray matter tissue volumes from an MRI image and also 93 mean intensities from a PET image. For the modality fusion of MRI and PET (MRI+PET), we concatenated their features into a long vector of 186 features.

3.2 Experimental Setting

We compared our feature selection method with the widely used methods such as Fisher Score (FS for short) [3], LPP [7], LDA [3], and PCA [3]. The FS is categorized as a feature selection method since it selects features in the original feature space based on the score ranking [3]. Meanwhile, LPP, LDA, and PCA are subspace learning methods, which aim, respectively, at preserving the local structures, the maximal variance, and the global structures of the data [3,15]. We also compared the proposed method with the state-of-the-art feature selection methods applied for AD diagnosis: Sparse Joint Classification and Regression (SJCR) [12] and Multi-Modal Multi-Task (M3T) [16]. For these two methods, we followed their papers to apply a 0-1 encoding method for the class indicator matrix.

Table 1. Comparison of classification accuracy ((mean±standard deviation)%) of two classification tasks

Method	AD/MCI/NC			AD/MCI-C/MCI-NC/NC		
	MRI	**PET**	**MRI+PET**	**MRI**	**PET**	**MRI+PET**
FS	62.33±1.56	60.11±1.54	62.88±1.31	50.87±1.73	50.44±1.49	51.76±1.58
PCA	63.71±1.30	61.49±1.58	64.61±1.60	51.05±1.64	51.51±1.62	52.20±1.60
LPP	63.21±1.91	61.03±1.22	64.35±1.29	51.72±1.42	51.39±1.58	52.60±1.37
LDA	49.01±1.71	39.02±1.23	51.85±1.66	35.25±1.65	31.82±1.40	36.32±1.64
SJCR	64.02±1.36	61.31±1.73	67.66±1.63	52.13±1.73	51.85±1.68	55.98±1.65
M3T	63.30±1.66	61.32±1.90	67.91±1.91	51.89±1.61	50.91±1.83	54.47±1.67
Proposed	**68.31±1.23**	**65.50±1.50**	**73.35±1.53**	**59.74±1.52**	**56.29±1.53**	**61.06±1.40**

3.3 Classification Results

Table 1 reports the classification accuracy of all the methods for two multi-class classification problems. The experimental results in Table 1 clearly show that the proposed method outperformed all the competing methods in all experiments. For example, in the three-class classification problem, our method improved the classification accuracy by 4.29% (MRI), 4.01% (PET), and 5.44% (MRI+PET), respectively, compared to the best performances among the competing methods. Meanwhile, in the four-class classification problem, the classification improvements were higher than the best performances among the competing methods as much as 7.61% (MRI), 4.44% (PET), and 5.08% (MRI+PET), respectively. Based on these results, we argue that the proposed discriminative and noise-resistant feature selection method helped enhance the classification performances.

Besides, we found that LDA achieved the worst classification performances among all the methods. The main reason was that LDA projected the original high dimensional feature space into only two or three dimensional subspace, respectively. Such low-dimensional space was not enough to correctly classify the neurimaging features. On the other hand, the subspace learning methods, except for LDA, outperformed the feature selection method of FS. This makes it reasonable to integrate subspace learning into the feature selection framework. Moreover, the proposed method clearly outperformed both the conventional feature selection and subspace learning methods thanks to the combination of the two approaches.

3.4 Discussions

We investigated the importance of the brain regions in discriminating among classes based on the frequency of the selected ROIs by the proposed method with MRI+PET. According to our experimental results, we can know that the commonly selected regions in two multi-class classification tasks were uncus right, hippocampal formation right, uncus left, middle temporal gyrus left, hippocampal formation left, amygdala left, middle temporal gyrus right, and amygdala right from MRI, and precuneus right, precuneus left, and angular gyrus left

from PET. These regions were also selected by the proposed method with either MRI or PET and almost all the competing methods with MRI+PET. Moreover, these regions have been also shown to be highly related to AD and MCI practical clinical diagnosis [2,8]. In this regard, we can say that these regions can be the potential biomarkers for AD diagnosis.

Meanwhile, the numbers of selected features in three- and four-class classification tasks were, respectively, 50.52 and 34.36 on average. That is, the smaller number of features were used in the classification task of considering the larger number of classes. It is also interesting that the larger number of features from MRI rather than PET was selected in both three- and four-class classification problems. This was also observed in the competing methods. Furthermore, from Table 1, we can see that in general, the MRI-based methods achieved better performance than the PET-based methods. Based on these observations, it is likely that the structural MR image provides more discriminative information in identifying the clinical status related to AD, compared to the functional PET image.

4 Conclusions

In this work, we focused on the issue of discriminative feature selection for multi-class classification in AD diagnosis. Specifically, we proposed a novel feature selection method by integrating subspace learning, which utilized both the global and the local information inherent in the data, into in a sparse least square regression framework. In our experimental results on the ADNI dataset, we validated the efficacy of the proposed method by enhancing the classification accuracies in multi-class classification problems.

Acknowledgements. This study was supported by National Institutes of Health (EB006733, EB008374, EB009634, AG041721, AG042599, and MH100217). Xiaofeng Zhu was partly supported by the National Natural Science Foundation of China under grant 61263035.

References

1. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine Learning 73(3), 243–272 (2008)
2. Chételat, G., Eustache, F., Viader, F., Sayette, V.D.L., Pélerin, A., Mézenge, F., Hannequin, D., Dupuy, B., Baron, J.C., Desgranges, B.: FDG-PET measurement is more accurate than neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment. Neurocase 11(1), 14–25 (2005)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. John Wiley & Sons (2012)
4. Fan, Y., Rao, H., Hurt, H., Giannetta, J., Korczykowski, M., Shera, D., Avants, B.B., Gee, J.C., Wang, J., Shen, D.: Multivariate examination of brain abnormality using both structural and functional MRI. NeuroImage 36(4), 1189–1199 (2007)

5. Franke, K., Ziegler, G., Klöppel, S., Gaser, C.: Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. NeuroImage 50(3), 883–892 (2010)
6. Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27(2), 83–85 (2005)
7. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: NIPS, pp. 1–8 (2005)
8. Misra, C., Fan, Y., Davatzikos, C.: Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. NeuroImage 44(4), 1415–1422 (2009)
9. Nesterov, Y.: Introductory lectures on convex optimization: a basic course, vol. 87 (2004)
10. Nie, F., Huang, H., Cai, X., Ding, C.H.Q.: Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: NIPS, pp. 1813–1821 (2010)
11. Suk, H.-I., Shen, D.: Deep learning-based feature representation for AD/MCI classification. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 583–590. Springer, Heidelberg (2013)
12. Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, A.J., Shen, L.: Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 115–123. Springer, Heidelberg (2011)
13. Wang, H., Nie, F., Huang, H., Risacher, S.L., Saykin, A.J., Shen, L., et al.: Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. Bioinformatics 28(12), i127–i136 (2012)
14. Wee, C.Y., Yap, P.T., Zhang, D., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D.: Identification of MCI individuals using structural and functional connectivity networks. Neuroimage 59(3), 2045–2056 (2012)
15. Ye, J.: Least squares linear discriminant analysis. In: ICML, pp. 1087–1093 (2007)
16. 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)
17. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55(3), 856–867 (2011)
18. Zhu, X., Huang, Z., Shen, H.T., Cheng, J., Xu, C.: Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recognition 45(8), 3003–3016 (2012)
19. Zhu, X., Huang, Z., Yang, Y., Shen, H.T., Xu, C., Luo, J.: Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recognition 46(1), 215–229 (2013)
20. Zhu, X., Suk, H.I., Shen, D.: Matrix-similarity based loss function and feature selection for Alzheimer's Disease diagnosis. In: CVPR (2014)
21. Zhu, X., Suk, H.I., Shen, D.: Multi-modality canonical feature selection for Alzheimer's disease diagnosis. In: Golland, P. (ed.) MICCAI 2014, Part II. LNCS, vol. 8674, pp. 162–169. Springer, Heidelberg (2014)
22. Zhu, X., Suk, H.I., Shen, D.: A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. NeuroImage 14, 1–30 (2014)
23. Zhu, X., Suk, H.-I., Shen, D.: A novel multi-relation regularization method for regression and classification in AD diagnosis. In: Golland, P. (ed.) MICCAI 2014, Part III. LNCS, vol. 8675, pp. 401–408. Springer, Heidelberg (2014)