High-Order Graph Matching Based Feature Selection for Alzheimer's Disease Identification

Feng Liu[1,2], Heung-Il Suk[2], Chong-Yaw Wee[2], Huafu Chen[1], and Dinggang Shen[2]

[1] Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Sichuan, China
[2] Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, NC, USA
dgshen@med.unc.edu

Abstract. One of the main limitations of l_1-norm feature selection is that it focuses on estimating the target vector for each sample individually without considering relations with other samples. However, it's believed that the geometrical relation among target vectors in the training set may provide useful information, and it would be natural to expect that the predicted vectors have similar geometric relations as the target vectors. To overcome these limitations, we formulate this as a graph-matching feature selection problem between a predicted graph and a target graph. In the predicted graph a node is represented by predicted vector that may describe regional gray matter volume or cortical thickness features, and in the target graph a node is represented by target vector that include class label and clinical scores. In particular, we devise new regularization terms in sparse representation to impose high-order graph matching between the target vectors and the predicted ones. Finally, the selected regional gray matter volume and cortical thickness features are fused in kernel space for classification. Using the ADNI dataset, we evaluate the effectiveness of the proposed method and obtain the accuracies of 92.17% and 81.57% in AD and MCI classification, respectively.

1 Introduction

The most prevalent neurodegenerative brain disease in elderly people is Alzheimer's Disease (AD), which is characterized by progressive worsening of cognitive and memory functions [1]. According to Brookmeyer *et al.*'s report [2], it is expected that more than 30 million people worldwide could be living with this disease by 2050. Its prodromal stage called Mild Cognitive Impairment (MCI) can also cause cognitive changes that have a high risk of progressing to AD within years [3]. There has been great interest in the automatic diagnosis and/or prognosis of these diseases in many scientific fields.

The main difficulty for the computer-aided brain disease diagnosis comes from the high-dimensional nature of the neuroimaging data. For the last decades, the application of machine learning techniques to the neuroimaging data made a promising improvement in brain disease classification [4, 5]. However, it remains challenging to

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 311–318, 2013.

circumvent the problem of feature selection from the noisy and redundant features to enhance the diagnostic accuracy. To address this issue, the Least Absolute Shrinkage and Selection Operator (LASSO), a sparse representation method, which penalizes a linear regression model with l_1-norm [6], has been used. While LASSO proved the efficacy in selecting features by inducing sparsity to the regression coefficients, it is limited in the sense that it estimates the target vector for each sample individually without considering the relation with other samples.

To overcome this limitation, we propose a novel feature selection method by means of high-order graph matching in sparse representation. The motivation of our approach is that the target vectors of the same class should be closer to each other, while those of different classes should be farther apart from each other. We expect that the same property should be satisfied with the vectors predicted via the sparse representation. This is formulated as a graph-matching feature selection problem between a predicted graph and a target graph. In the predicted graph a node is represented by the predicted vector that may describe regional Gray Matter Volume (GMV) or Cortical Thickness (CT) features, while in the target graph a node is represented by target vector. In short, the graph of the predicted vectors should be similar to, or ideally matched to, the graph of target vectors. We introduce two regularization terms to the conventional LASSO: the first term takes into account the similarity between two nodes in the predicted graph and the corresponding two nodes in the target graph (i.e. binary term), and the second term takes into account the geometric similarity between three nodes in the predicted graph and the corresponding three nodes in the target graph (i.e. ternary term). Unlike the conventional LASSO that considers a unary relation between a target vector and its predicted one, the proposed method takes into account high-order relations such as binary and ternary from a graph perspective. It's worth noting that our method is also different from Local Linear Embedding (LLE), where neighboring samples in the original space still stay near to each other in the dimension-reduced space. Therefore, LLE does not guarantee the separation of original nearby samples that belong to different classes. In contrast, the proposed method finds a low-dimensional space where the transformed samples are more separable between classes due to the two additional regularization terms. Using the ADNI dataset, we evaluate the performance of the proposed method, and compare with the competing methods.

2 Materials and Preprocessing

We use the MRI datasets of 594 subjects in the ADNI dataset[1]: 198 AD patients, 198 MCI patients, and 198 normal controls. Subjects from each group are randomly selected with a ratio of 1:1:1 to prevent data unbalance problem. Although the dataset includes longitudinal MRI data, we consider only the baseline data in this study. The T1-weighted MRI images are preprocessed using FreeSurfer software[2]. The preprocessed images are parcellated into 68 cortical regions based on the Desikan–Killiany

[1] http://www.loni.ucla.edu/ADNI

[2] http://surfer.nmr.mgh.harvard.edu/

Cortical Atlas [7]. In each cortical region, we compute the average regional GMV and CT as features, obtaining two 68-dimensional feature vectors for each subject. Before feature selection, GMV and CT of each region are normalized by intracranial volume and its corresponding standard deviation, respectively.

3 Proposed Method

Figure 1 illustrates an overview of the proposed framework for AD/MCI classification. From the preprocessed MRI images, we first extract the GMV and CT features from each ROI as outlined in Section 2. For each feature type, feature selection is performed using the proposed graph matching based method. Using the selected features, a GMV kernel matrix and a CT kernel matrix are constructed, respectively. The two kernel matrices are then combined and used to train the SVM classifier.

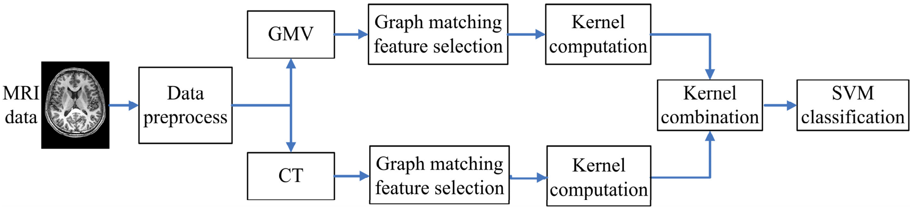

Fig. 1. Schematic diagram of the proposed classification framework

3.1 Sparse Representation with High-Order Graph Matching

Let $\mathbf{X}^f = [\mathbf{x}_1^f, \ldots, \mathbf{x}_n^f, \ldots, \mathbf{x}_N^f] \in \Re^{D \times N}$ be a feature matrix of feature type f and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n, \ldots, \mathbf{y}_N] \in \Re^{M \times N}$ be a matrix of corresponding target vectors, where N is the total number of training samples, D is the dimension of the feature vector, and M is the dimension of the target vector. In general, the target vector can include class label, Mini-Mental State Examination (MMSE) score, or other clinical scores. Assuming that the target vectors can be represented by a linear combination of the feature vectors, we minimize the regression errors between the target vectors and the predicted vectors as follows:

$$L(\mathbf{W}^f) = \min_{\mathbf{W}^f} \left\| (\mathbf{W}^f)^T \mathbf{X}^f - \mathbf{Y} \right\|_F^2 \tag{1}$$

where $\mathbf{W}^f = [\mathbf{w}_1^f, \ldots, \mathbf{w}_m^f, \ldots, \mathbf{w}_M^f] \in \Re^{D \times M}$ denotes the regression coefficient matrix and $\|\cdot\|_F$ denotes a Frobenius norm. In order to remove the unexpected noises and redundant information in data, a sparse representation method referred as LASSO [6] has been applied in previous studies by penalizing linear regression model $\mathbf{W}^f$with a l_1-norm. However, the conventional LASSO estimates the target vector of each sample independently without considering the relation among samples. Therefore, from a classification standpoint, it is limited in terms of finding discriminative features that can enhance the classification accuracy.

Geometrically, target vectors of the same class should be closer to each other, while those of different classes should be farther apart from each other. It is expected that the same property should be satisfied with the predicted vectors. That is, the geometric relations among the target vectors $\mathbf{Y}$ should be kept in the predicted vectors $(\mathbf{W}^f)^T\mathbf{X}^f$. This can be defined as a graph-matching problem, between a predicted graph and a target graph, where the graph of the predicted vectors should be similar to, or ideally match to, the graph of target vectors. To accomplish this, we devise the binary regularization term (B) as shown in Eq. (2), and the ternary regularization term (T) as shown in Eq. (3).

$$\mathrm{B} = \sum_{i,j=1}^{N} \left\| (\mathbf{y}_i - \mathbf{y}_j) - (\mathbf{W}^f)^T(\mathbf{x}_i^f - \mathbf{x}_j^f) \right\|_F^2 \tag{2}$$

$$\mathrm{T} = \sum_{i,j,k=1}^{N} \left\| (\mathbf{y}_i - \mathbf{y}_j)^T(\mathbf{y}_j - \mathbf{y}_k) - \{(\mathbf{W}^f)^T(\mathbf{x}_i^f - \mathbf{x}_j^f)\}^T\{(\mathbf{W}^f)^T(\mathbf{x}_j^f - \mathbf{x}_k^f)\} \right\|_F^2 \tag{3}$$

In particular, the binary regularization term in Eq. (2) produces a scalar value that measures the similarity between two target vectors ($\mathbf{y}_i$ and $\mathbf{y}_j$) in the target graph and the corresponding predicted ones ($(\mathbf{W}^f)^T\mathbf{x}_i^f$ and $(\mathbf{W}^f)^T\mathbf{x}_j^f$) in the predicted graph. Meanwhile, the regularization term in Eq. (3) produces a scalar value that measures the geometric (i.e., the dot product between two edges) similarity among three nodes ($\mathbf{y}_i$, $\mathbf{y}_j$ and $\mathbf{y}_k$) in the target graph and the corresponding three nodes in the predicted graph ($(\mathbf{W}^f)^T\mathbf{x}_i^f$, $(\mathbf{W}^f)^T\mathbf{x}_j^f$, and $(\mathbf{W}^f)^T\mathbf{x}_k^f$). **Figure 2** presents a conceptual illustration of these relations in a graph.

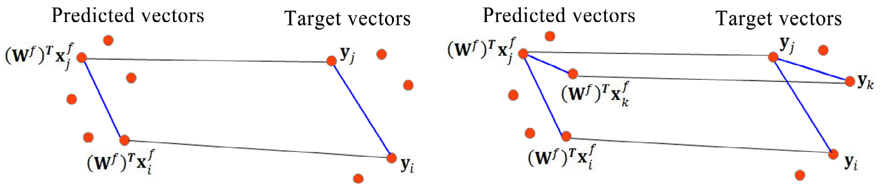
Predicted vectors
Target vectors
$(\mathbf{W}^f)^T\mathbf{x}_j^f$
$\mathbf{y}_j$
$(\mathbf{W}^f)^T\mathbf{x}_i^f$
$\mathbf{y}_i$
Predicted vectors
Target vectors
$(\mathbf{W}^f)^T\mathbf{x}_j^f$
$\mathbf{y}_j$
$\mathbf{y}_k$
$(\mathbf{W}^f)^T\mathbf{x}_k^f$
$(\mathbf{W}^f)^T\mathbf{x}_i^f$
$\mathbf{y}_i$

Fig. 2. A conceptual illustration of the proposed high-order graph matching, binary (left) and ternary (right) relation

With the introduction of these two terms into the conventional LASSO, we define our new objective function as follows:

$$L(\mathbf{W}^f) = \min_{\mathbf{W}^f} \left\| (\mathbf{W}^f)^T\mathbf{X}^f - \mathbf{Y} \right\|_F^2 + \lambda_1 \left\| \mathbf{W}^f \right\|_1 + \lambda_2 \mathrm{B} + \lambda_3 \mathrm{T} \tag{4}$$

where λ_1, λ_2 and λ_3 are regularization control parameters. By setting λ_2 and λ_3 zero, we obtain the conventional LASSO. In contradiction to the conventional LASSO, the proposed method considers the high-order relationship, i.e., the binary and ternary, from a graph perspective. Besides the binary and ternary relations, theoretically, it is possible to include higher-order of graph matching information into the objective function in Eq. (4). An Accelerated Proximal Gradient (APG) method [8] is used for the optimization of Eq. (4). Of note, we select features with non-zero regression coefficients from original feature space for final classification.

It is noteworthy that the proposed method can be used for feature selection in both classification of the clinical category and prediction of the clinical scores. If we use class label as the target vector, as done in our experiments, then the selected features can be considered as discriminative ones in classification. Meanwhile, if MMSE and other clinical scores are considered as the target vector, then the selected features are informative in predicting clinical scores in regression study. In the experiment below, we just use class label as the target vector for the purpose of classification.

3.2 Multi-kernel SVM for Classification

Using the regional GMV features and CT features selected in Section 3.1, we exploit the complementary information. Specifically, in order to fuse the information for classification, we utilize a multi-kernel SVM approach [9]. First a kernel matrix is constructed for each feature type, and then combined using a weighted linear combination as follows:

$$K[(\mathbf{x}_n^1, \mathbf{x}_n^2), (\mathbf{x}^1, \mathbf{x}^2)] = \sum_{f=1}^{2} \beta^f k^f\left(\mathbf{x}_n^f, \mathbf{x}^f\right) \tag{5}$$

where $(\mathbf{x}_n^1, \mathbf{x}_n^2)$ is the feature vectors of the n-th sample with two types of features $\mathbf{x}_n^1$ and $\mathbf{x}_n^2$, and $(\mathbf{x}^1, \mathbf{x}^2)$ is the feature vectors of a testing sample. β^f is a mixing coefficient with the constraint of $\beta^f \geq 0$ and $\sum_{f=1}^{2} \beta^f = 1$. $k^f\left(\mathbf{x}_n^f, \mathbf{x}^f\right) = \phi^f\left(\mathbf{x}_n^f\right)^T \phi^f(\mathbf{x}^f)$ is a kernel function and ϕ^f is a kernel-mapping function of the f-type feature. After constructing the combined kernel matrix, it is then straightforward to apply a linear SVM as follows:

$$l(\mathbf{x}^1, \mathbf{x}^2) = sign\left\{\sum_{n=1}^{N} c_n \alpha_n K[(\mathbf{x}_n^1, \mathbf{x}_n^2), (\mathbf{x}^1, \mathbf{x}^2)] + b\right\} \tag{6}$$

where $c_n \in \{1, -1\}$ is the class label of the n-th training sample, α_n is a Lagrangian multiplier, and b is a bias. LIBSVM toolbox[3] is used to solve the above functions.

4 Experimental Results and Analysis

In this section, we present the effectiveness of the proposed method in two binary classification problems, i.e., AD vs. NC and MCI vs. NC, on ADNI dataset. Even though the proposed method can deal with multiple target values by concatenating

[3] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

them into a vector, we use only the class label for the target value in Eq. (4) to focus on the classification problems.

4.1 Experimental Settings

A 10-fold cross-validation strategy is employed to evaluate the classification performance. Specifically, the whole samples are randomly partitioned into 10 subsets and then we choose one subset for test and use the remaining 9 for training, and this procedure is repeated 10 times. In order to determine the hyper-parameters of λ_1, λ_2 and λ_3 in Eq. (4), and β^f in Eq. (5), we further split the training samples for another round of cross-validation. The hyper-parameters that perform the best in the inner cross-validation are used to classify testing subjects in the outer loop. Due to a possible bias during dataset partitioning for cross-validation, we repeat the whole process 10 times. The final accuracy is computed by averaging of the accuracies from all experiments. We quantify classification performance using four statistical measures, i.e., accuracy, sensitivity, specificity, and Area Under receiver operating characteristic Curve (AUC).

4.2 Classification Performances

We first consider the effectiveness of the proposed regularization terms by comparing with the conventional LASSO-based feature selection. In this experiment, feature combination is not considered to validate the methodological efficacy of our method. Specifically, we consider the following four competing methods: 1) LASSO-based feature selection on the CT feature (L-CT), 2) LASSO-based feature selection on the GMV feature (L-GMV), 3) Proposed Graph Matching based feature selection method on the CT feature (GM-CT), and 4) Proposed Graph Matching based feature selection method on the GMV feature (GM-GMV).

Table 1. Performance comparison with the competing methods for AD/MCI classification

Methods	AD vs. NC				MCI vs. NC			
	ACC (%)	SEN (%)	SPE (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	AUC (%)
L-CT	82.42	81.46	83.38	89.03	72.23	67.20	77.25	76.78
L-GMV	86.57	85.00	88.13	92.20	75.65	71.70	79.60	79.30
GM-CT	85.10	82.32	87.88	92.78	75.25	74.75	75.76	81.94
GM-GMV	88.89	85.35	92.42	94.44	77.76	75.25	80.30	84.83
L-COM	88.69	88.23	89.14	93.93	77.22	73.45	81.00	80.82
Baseline	90.59	88.42	92.76	95.79	80.15	76.12	84.18	86.88
Proposed	**92.17**	**89.39**	**94.95**	**96.83**	**81.57**	**77.78**	**85.35**	**88.16**

We summarize the results in **Table 1** and present Receive Operating Characteristic curves in **Figure 3**. The proposed high-order graph matching method consistently outperforms the conventional LASSO across feature types in both classification problems. For MCI diagnosis, the proposed method result in an increase of 7.55% and

3.55% in sensitivity (i.e., correct diagnosis of MCI patients) by using CT and GMV, respectively, which is clinically important for early and proper treatment. In our second experiment, we test the validity of the proposed method by combining the feature types of GMV and CT. Specifically, we compare with the method of LASSO-based feature selection from each type of features (GMV and CT) and combine with multi-kernel SVM (L-COM).

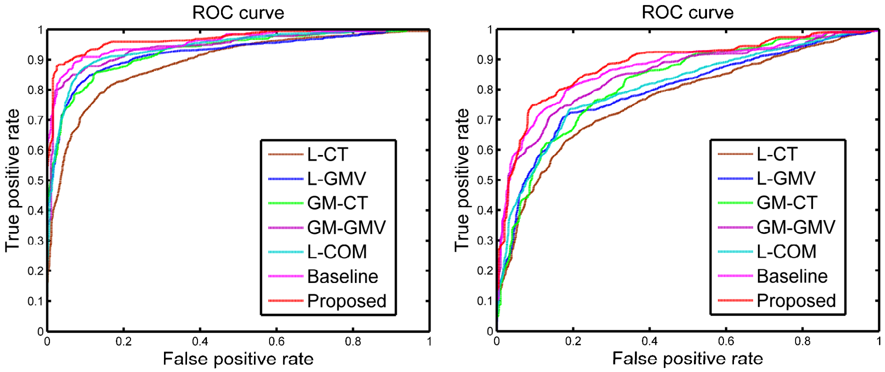

Fig. 3. Receiver Operating Characteristic (ROC) curves of the competing methods for AD vs. NC (left) and MCI vs. NC (right)

By fusing the GMV and CT features, compared to L-COM, we could improve the accuracies, showing 92.17% (with sensitivity of 89.39%, specificity of 94.95%, and AUC of 96.83%) and 81.57% (with sensitivity of 77.78%, specificity of 85.35%, and AUC of 88.16%) in AD and MCI classification, respectively. To test the validity of multi-kernel SVM, we perform AD and MCI classifications by concatenating of all types of features (Baseline, as shown in **Table 1**), i.e., equal weight for each feature type, and obtain the accuracies of 90.59% and 80.15% for AD and MCI classification, respectively. These results demonstrate the superiority of multi-kernel SVM method.

Table 2. Comparison of reconstruction error and the number of selected features (Mean ± SD)

Methods	AD vs. NC		MCI vs. NC	
	Number of features	Reconstruction error	Number of features	Reconstruction error
L-COM	35.16 ± 11.43	0.75 ± 0.22	47.77 ± 23.50	0.86 ± 0.38
Proposed	29.60 ± 15.45	0.63 ± 0.18	38.63 ± 17.24	0.73 ± 0.29

We also compare the reconstruction error (i.e., first term in Eq. (4)) and the number of selected features between L-COM and the proposed method. Since the feature selection in each fold is performed on the different training samples, the selected features and the reconstruction error can vary across cross-validation folds. To this end, we provide the statistics of them in **Table 2**. Both the number of selected features and the reconstruction error are smaller than those of L-COM, indicating the effectiveness of the proposed method. Besides, we define the most discriminative regions as the regions that are most frequently selected in all cross-validations. The most discriminative regions include precuneus, entorhinal cortex, temporal pole, fusiform gyrus, parahippocampal gyrus, insula, *etc*, which are highly associated with AD-pathology.

5 Conclusions

We propose a novel feature selection method via high-order graph matching framework. The key idea of the proposed method is that the predicted target vectors should have the same geometric properties with that of the target vectors, formulating it as a high-order graph-matching problem. We devise two new regularization terms, specifically a binary relation and a ternary relation among nodes in the target and predicted graphs. The selected regional GMV and CT features were then fused using a multi-kernel SVM. In our experiment on ADNI dataset, the proposed method outperform the competing methods, presenting the accuracies of 92.17% and 81.57% in AD and MCI classification, respectively.

References

1. Blennow, K., de Leon, M.J., Zetterberg, H.: Alzheimer's disease. Lancet 368, 387–403 (2006)
2. Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M.: Forecasting the global burden of Alzheimer's disease. Alzheimers Dement 3, 186–191 (2007)
3. Petersen, R.C., Doody, R., Kurz, A., Mohs, R.C., Morris, J.C., Rabins, P.V., Ritchie, K., Rossor, M., Thal, L., Winblad, B.: Current concepts in mild cognitive impairment. Arch. Neurol. 58, 1985–1992 (2001)
4. Casanova, R., Whitlow, C.T., Wagner, B., Williamson, J., Shumaker, S.A., Maldjian, J.A., Espeland, M.A.: High dimensional classification of structural MRI Alzheimer's disease data based on large scale regularization. Front Neuroinform. 5, 22 (2011)
5. Liu, F., Guo, W., Yu, D., Gao, Q., Gao, K., Xue, Z., Du, H., Zhang, J., Tan, C., Liu, Z., Zhao, J., Chen, H.: Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. PLoS One 7, e40968 (2012)
6. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288 (1996)
7. Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980 (2006)
8. Nesterov, Y.: Introductory lectures on convex optimization: A basic course. Springer (2003)
9. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856–867 (2011)