Published in final edited form as:

Brain Imaging Behav. 2016 March ; 10(1): 33–40. doi:10.1007/s11682-015-9360-1.

Multi-task feature selection via supervised canonical graph matching for diagnosis of autism spectrum disorder

Liye Wang,

School of Life Science, Beijing Institute of Technology, Beijing 100081, China. IDEA Lab, Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Chong-Yaw Wee,

IDEA Lab, Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Xiaoying Tang,

School of Life Science, Beijing Institute of Technology, Beijing 100081, China

Pew-Thian Yap, and

IDEA Lab, Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Dinggang Shen

IDEA Lab, Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

Dinggang Shen: dgshen@med.unc.edu

Abstract

In this paper, we propose a novel framework for ASD diagnosis using structural magnetic resonance imaging (MRI). Our method deals explicitly with the distributional differences of gray matter (GM) and white matter (WM) features extracted from MR images. We project linearly the GM and WM features onto a canonical space where their correlations are mutually maximized. In this canonical space, features that are highly correlated with the class labels are selected for ASD diagnosis. In addition, graph matching is employed to preserve the geometrical relationships between samples when projected onto the canonical space. Our evaluations based on a public ASD dataset show that the proposed method outperforms all competing methods on all clinically important measures in differentiating ASD patients from healthy individuals.

Keywords

Diagnosis of autism spectrum disorder; Magnetic resonance imaging (MRI); Multi-task feature selection

Correspondence to: Dinggang Shen, dgshen@med.unc.edu.

Conflict of interest Liye Wang, Chong-Yaw Wee, Xiaoying Tang, Pew-Thian Yap, and Dinggang Shen declare that they have no conflicts of interest.

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by impairment of social interaction, language, behavior, and cognitive functions (Geschwind and Levitt 2007; Wing and Gould 1979). According to the latest report released by the Centers for Disease Control and Prevention (CDC), 1 in 68 American children was affected by some form of ASD, an increase of 78 % compared to a decade ago, with boys outnumbering girls by a ratio of 4.5:1. Currently, diagnosis for ASD is solely behavior-based and relies on a series of clinical measures that quantify the severity of the disorder. However, an important drawback of solely behavioral-based diagnosis methods is that many behavioral phenotypes are associated with other numerous psychological and psychiatric disorders (Geschwind and Levitt 2007; Guilmatre et al. 2009; Wee et al. 2014b). Additionally, it is difficult to use a single clinical measure for diagnosis and prognosis due to potential heterogeneity in the patient group. Therefore, combining biological information with behavioral measurements can assist physicians in ASD diagnosis.

In this paper, we propose a novel machine-learning framework to distinguish autistic patients from normal controls based on structural MR. This is motivated by several neuroimaging and post-mortem studies suggesting that ASD is highly associated with neuroanatomical abnormalities (Amaral et al. 2008; Boddaert et al. 2004; Brambilla et al. 2003; Wee et al. 2014b). In our framework, MR image features of white matter (WM) tissues and gray matter (GM) tissues are treated as predictors, and clinical labels and scores, i.e., social responsiveness scale (SRS_TOTAL), are treated as target responses. Note that SRS_TOTAL is a quantitative measure that is commonly used in clinical ASD diagnosis for children and youth.

Our framework includes a novel feature selection method that considers explicitly the correlations between features via canonical correlation analysis (CCA) and the correlations between the features and labels via Pearson correlation. More specifically, we first transform the features to a new feature space spanned by the canonical bases obtained by CCA (Hardoon et al. 2004), and then perform multi-task learning (Chai et al. 2008; Evgeniou and Pontil 2007; Liu et al. 2009; Xue et al. 2007; Zhang et al. 2012a) to select the most discriminative features with information provided by the diagnostic labels and clinical scores. We treat the canonical features as regressors and penalize each regressor using a linear combination of CCA and Pearson correlations. Lastly, to preserve the geometric distribution of the data set after projecting into the canonical space, an additional graph matching term (Jie et al. 2013) is also included in our objective function. With the selected features, a support vector machine (SVM) (Cortes and Vapnik 1995) is finally trained for ASD classification.

The contribution of this paper is three-fold: (1) We develop an objective method to assist ASD diagnosis by using structural MR images; (2) We integrate CCA into graph matching based sparse group lasso learning method, extending the use of conventional multi-task learning; (3) We design a new objective function that utilizes both the correlation between

features, and also the correlation between features and clinical labels for selecting the most discriminative features.

The remainder of the paper is organized as follows. In the “Materials and preprocessing” section, we provide information on the imaging data and the preprocessing pipeline. In the “Methods” section, the mathematical detail of the proposed feature selection method is described. In the “Experimental results” section, we demonstrate the validity of the proposed method for ASD diagnosis by comparison with state-of-the-art methods. Finally, we discuss our findings and conclude our work in the “Discussion and Conclusion” sections.

Materials and preprocessing

Subjects

Data used in this study are obtained from Autism Brain Imaging Data Exchange (ABIDE) database, which is publicly available. We used MR images of 54 ASD patients and 57 normal controls under 15 years of age, scanned at New York University (NYU) Langone Medical Center. Table 1 summarizes the demographic information of the subjects used in our work.

There are two main reasons for us to select only the NYU dataset: 1) ABIDE is a collection of datasets scanned at different sites with different scanning parameters and protocols - this inconsistency will complicate the study, and 2) the NYU dataset has the largest sample size. But it is worth indicating that the proposed method is not designed specifically for a certain dataset; it can be applied to other datasets, as long as data are acquired with similar imaging protocols.

Data acquisition and preprocessing

For details on data acquisition protocols and scanning parameters, please refer to ‘http://fcon_1000.projects.nitrc.org/indi/abide/’. All the MR images were preprocessed by skull stripping (Wang et al. 2011), cerebellum removal, and tissue segmentation (into WM, GM, and cerebrospinal fluid (CSF)) (Lim and Pfefferbaum 1989). The anatomical automatic labeling (AAL) atlas, parcellated with 90 predefined regions, was registered using HAMMER (Shen and Davatzikos 2002; Wang et al. 2011) to the native space of each subject. (Note that other registration methods can also be applied here.) We then computed the WM and GM tissue volumes in each region and used them as features, i.e., obtaining 90 WM and 90 GM features, for training a classifier.

Methods

We propose a novel CCA-based graph matching sparse group lasso (GMSGSL) feature selection method for facilitating ASD diagnosis. First, the GM and WM features are linearly projected onto a joint canonical feature space. Then, we extend the ordinary sparse group lasso to the CCA space by regularizing element-wise and group-wise sparse terms with a linear combination of both canonical and Pearson correlations. Features with large correlations (i.e., the sum of canonical and Pearson correlations) are less penalized and are thus more likely to be selected. Besides, a graph matching term is included to preserve the

geometric structure of the features. Lastly, the canonical features that are strongly correlated with the class labels are used to train a SVM classifier. The flowchart of the proposed ASD diagnosis framework is shown in Fig. 1. It should be noted that, in the stage of CCA-based GMSGL, both the SRS_TOTAL scores and the class labels are used, whereas, in the stage of SVM classification, only the class labels are used as the target responses.

Throughout the paper, we denote matrices as boldface uppercase, vectors as boldface lowercase, and scalars as italic letters, respectively. A superscript T is used to denote a vector/matrix transpose.

Canonical correlation analysis

Assuming that we have the features corresponding to D ROIs for gray matter (GM) and white matter (WM) of N subjects, we can form a feature matrix $\mathbf{X} = [\mathbf{X}^{(G)}; \mathbf{X}^{(W)}] \in \mathbb{R}^{2D \times N}$,

where $\mathbf{X}^{(G)} \in \mathbb{R}^{D \times N}$ and $\mathbf{X}^{(W)} \in \mathbb{R}^{D \times N}$. Let $\Sigma = \begin{pmatrix} \Sigma_{G,G} & \Sigma_{G,W} \\ \Sigma_{W,G} & \Sigma_{W,W} \end{pmatrix}$ be the corresponding covariance matrix. CCA projects two multi-dimensional random variables onto a joint space where their correlation is maximized. Specifically, it seeks basis vectors $\mathbf{B}^{(G)} \in \mathbb{R}^{D \times D'}$ and $\mathbf{B}^{(W)} \in \mathbb{R}^{D \times D'}$ with $D' = \min\{\text{rank}(\mathbf{X}^{(G)}), \text{rank}(\mathbf{X}^{(W)})\}$ to project the features onto a new space where the correlations are mutually maximized, i.e.,

$$(\hat{\mathbf{B}}^{(G)}, \hat{\mathbf{B}}^{(W)}) = \arg \max_{(\mathbf{B}^{(G)}, \mathbf{B}^{(W)})} \text{trace} \left(\mathbf{B}^{(G)T} \Sigma_{G,W} \mathbf{B}^{(W)} \right) \quad (1)$$

subject to $\mathbf{B}^{(G)T} \Sigma_{G,G} \mathbf{B}^{(G)} = \mathbf{I}$, $\mathbf{B}^{(W)T} \Sigma_{W,W} \mathbf{B}^{(W)} = \mathbf{I}$, and $\mathbf{B}^{(G)T} \Sigma_{G,W} \mathbf{B}^{(W)}$ has zero off-diagonal elements. The optimal solution $(\hat{\mathbf{B}}^{(G)}, \hat{\mathbf{B}}^{(W)})$ can be effectively obtained by generalized eigen decomposition (Hardoon et al. 2004). The canonical features $\mathbf{Z}^{(G)} = \hat{\mathbf{B}}^{(G)T} \mathbf{X}^{(G)} \in \mathbb{R}^{D' \times N}$ and $\mathbf{Z}^{(W)} = \hat{\mathbf{B}}^{(W)T} \mathbf{X}^{(W)} \in \mathbb{R}^{D' \times N}$ can be grouped as a canonical feature matrix as $\mathbf{Z} = [\mathbf{Z}^{(G)}; \mathbf{Z}^{(W)}] \in \mathbb{R}^{2D' \times N}$. Also, the canonical representations satisfy the following properties (Kakade and Foster 2007):

- a. Orthogonality: $\mathbb{E} \left[\mathbf{z}_j^{(i)T} \mathbf{z}_k^{(i)} \right] = \delta(j, k)$,
- b. Correlation: $\mathbb{E} \left[\mathbf{z}_j^{(G)T} \mathbf{z}_k^{(W)} \right] = r_j \delta(j, k)$

where $\mathbf{z}_j^{(i)}$ is the j -th row of $\mathbf{Z}^{(i)}$, and $i \in \{G, W\}$. $\mathbb{E}[\cdot]$ denotes the expectation operator and δ denotes the Kronecker delta function. $\delta(j, k) = 1$, if $j = k$; otherwise, $\delta(j, k) = 0$. Note that $1 \quad r_1 \quad r_2 \quad \dots \quad r_{D'} \quad 0$. This CCA step is summarized in Fig. 2.

Feature selection

We employ a sparse regression method (Liu et al. 2014; Wee et al. 2014a; Zhang et al. 2012b) to deal with small sample size problem. Since the two target responses, i.e., class labels (patients: +1; normal controls: -1) $\mathbf{y}^{(C)}$ and clinical score (SRS_TOTAL) $\mathbf{y}^{(S)}$ are correlated, we apply a multi-task learning algorithm for feature selection. Here, each task is

associated with one target response. We solve the following sparse group lasso (Friedman et al. 2010) problem in the CCA domain:

$$\min_{\hat{\mathbf{W}}} L(\hat{\mathbf{W}}) = \frac{1}{2} \sum_{i \in \{C, S\}} \|\mathbf{y}^{(i)} - \mathbf{Z}^T \hat{\mathbf{w}}^{(i)}\|_2^2 + \lambda_1 \|\hat{\mathbf{W}}\|_{2,1}^{\text{CCA}} + \lambda_2 \|\hat{\mathbf{W}}\|_1^{\text{CCA}} \quad (2)$$

where $\hat{\mathbf{W}} = [\hat{\mathbf{w}}^{(C)}, \hat{\mathbf{w}}^{(S)}]$, with $\hat{\mathbf{w}}^{(C)} \in \mathbb{R}^{2D'}$ and $\hat{\mathbf{w}}^{(S)} \in \mathbb{R}^{2D'}$ denoting the weight vectors of the canonical features for the label-based task and the score-based task, respectively.

$$\|\hat{\mathbf{W}}\|_{2,1}^{\text{CCA}} = \sum_{j=1}^{D'} \frac{1-r_j'}{r_j'} \left(\|\hat{\mathbf{w}}^j\|_2 + \|\hat{\mathbf{w}}^{j+D'}\|_2 \right), \text{ and}$$

$$\|\hat{\mathbf{W}}\|_1^{\text{CCA}} = \sum_{j=1}^{D'} \frac{1-r_j'}{r_j'} \left(\|\hat{\mathbf{w}}^j\|_1 + \|\hat{\mathbf{w}}^{j+D'}\|_1 \right), \text{ where } \hat{\mathbf{w}}^j \text{ is the } j\text{-th row of } \hat{\mathbf{W}}, \text{ and } \{r_j'\}_{j=1:D'}$$

denotes a set of canonical correlation coefficients. To preserve the geometric distribution information among the data, we further introduce a graph matching term.

$$\min_{\hat{\mathbf{W}}} \sum_{i \in \{C, S\}} \sum_{m,n=1}^N s_{m,n}^{(i)} \|\hat{\mathbf{w}}^{(i)T} \mathbf{z}_{.,m} - \hat{\mathbf{w}}^{(i)T} \mathbf{z}_{.,n}\|_2^2 = 2 \sum_{i \in \{C, S\}} \hat{\mathbf{w}}^{(i)T} \mathbf{Z}^T \mathbf{L}^{(i)} \mathbf{Z} \hat{\mathbf{w}}^{(i)} \quad (3)$$

$S^{(i)} = [s_{m,n}^{(i)}]$ denotes a matrix encoding the similarity of the i -th task across different samples. $\mathbf{z}_{.,m}$ denotes the feature vector of the m -th subject, and also the m -th column in $\mathbf{Z}$. $\mathbf{L}^{(i)} = \mathbf{D}^{(i)} - \mathbf{S}^{(i)}$ represents a Laplacian matrix for the i -th task, where $\mathbf{D}^{(i)}$ is a diagonal

matrix defined as $D_{m,m}^{(i)} = \sum_{n=1}^N s_{m,n}^{(i)}$. The similarity matrix is defined as

$s_{m,n}^{(i)} = \exp \left(- \left(y_m^{(i)} - y_n^{(i)} \right)^2 \right)$. Also, we take into account the Pearson correlations between features and class labels when formulating the $L_{2,1}$ and L_1 norm terms. Hence, the objective function of the proposed CCA-based GMSGL is as follows:

$$\min_{\hat{\mathbf{W}}} L(\hat{\mathbf{W}}) = \frac{1}{2} \sum_{i \in \{C, S\}} \|\mathbf{y}^{(i)} - \mathbf{Z}^T \hat{\mathbf{w}}^{(i)}\|_2^2 + \lambda_1 \|\hat{\mathbf{W}}\|_{2,1}^{\text{CCA+Pearson}} + \lambda_2 \|\hat{\mathbf{W}}\|_1^{\text{CCA+Pearson}} + \lambda_3 \sum_{i \in \{C, S\}} \hat{\mathbf{w}}^{(i)T} \mathbf{Z}^T \mathbf{L}^{(i)} \mathbf{Z} \hat{\mathbf{w}}^{(i)} \quad (4)$$

where $\|\hat{\mathbf{W}}\|_{2,1}^{\text{CCA+Pearson}} = \sum_{j=1}^{D'} \frac{1-r_j'}{r_j'} \left(\|\hat{\mathbf{w}}^j\|_2 + \|\hat{\mathbf{w}}^{j+D'}\|_2 \right)$, and

$$\|\hat{\mathbf{W}}\|_1^{\text{CCA+Pearson}} = \sum_{j=1}^{D'} \frac{1-r_j'}{r_j'} \left(\|\hat{\mathbf{w}}^j\|_1 + \|\hat{\mathbf{w}}^{j+D'}\|_1 \right) \text{ subject to } r_j = \theta r_j' + (1-\theta) |r_j'|, \theta \in [0,$$

1], and r_j' denotes the Pearson coefficient between the j -th feature and class label. It should

be noted that $(\mathbf{y}_m^{(i)} - \mathbf{y}_n^{(i)})^2$ is the relative distance between the target responses $\mathbf{y}_m^{(i)}$ and $\mathbf{y}_n^{(i)}$, while $\|\hat{\mathbf{w}}^{(i)T} \mathbf{z}_{.,m} - \hat{\mathbf{w}}^{(i)T} \mathbf{z}_{.,n}\|_2^2$ is the respective distance between feature vectors after projection to the common space (or respective distance between predictions) (Jie et al. 2013;

Zhang et al. 2013). Basically, if $(\mathbf{y}_m^{(i)} - \mathbf{y}_n^{(i)})^2$ is small, $\|\hat{\mathbf{w}}^{(i)T} \mathbf{z}_{.,m} - \hat{\mathbf{w}}^{(i)T} \mathbf{z}_{.,n}\|_2^2$ is also required to be small. By solving the optimization problem in Eq. (4) via an accelerated

proximal gradient method (Chen et al., 2009), we can select the informative canonical features based on the non-zero entries of the weight coefficient vector $\hat{\mathbf{w}}^{(C)}$.

Support vector machine (SVM) classifier

Let $\mathbf{z}'_{:,n}$ be the finally selected feature vector of the n -th training sample and $y_n^{(C)} \in \{+1, -1\}$ be the corresponding class label (patients: +1; normal controls: -1). The primary optimization problem of SVM is given as:

$$\min_{\mathbf{H}, b, \{\xi_n\}} \frac{1}{2} \|\mathbf{H}\|^2 + c \sum_{n=1}^N \xi_n, s.t. y_n^{(C)} (\mathbf{H}^T \phi(\mathbf{z}'_{:,n}) + b) \geq 1 - \xi_n, \xi_n \geq 0, n = 1, \dots, N, \quad (5)$$

where ξ_n is the non-negative slack variable, c is the penalty parameter, ϕ is the kernel induced mapping function, and b is the bias. For a given test sample $\mathbf{z}$, the decision function of SVM for the predicted label is defined as

$$f(\mathbf{z}) = \text{sign} \left(\sum_{n=1}^N y_n^{(C)} \alpha_n \kappa(\mathbf{z}'_{:,n}, \mathbf{z}) + b \right), \quad (6)$$

where α_n is the Lagrange multiplier and $\kappa(\mathbf{z}'_{:,n}, \mathbf{z}) = \phi^T(\mathbf{z}'_{:,n}) \phi(\mathbf{z})$ is the kernel function for $\mathbf{z}'_{:,n}$ and $\mathbf{z}$.

Validation

We validate the superiority of our method by using a nested 10-fold cross-validations approach. Specifically, the dataset was randomly partitioned into 10 subsets with no overlap; 9 out of the 10 subsets were used for training and the remaining for testing. We further partitioned the training set into 10 subsets for an inner-loop cross-validation determination of model parameters, i.e., $\lambda_1, \lambda_2, \lambda_3$, and θ in Eq. (4). The parameters that produced the best performance in the inner loop were used for classification of unseen test samples. The whole process was repeated 10 times with different random partitioning and the averaged results were reported.

Experimental results

SVM classifiers in all experiments are implemented using the LIBSVM toolbox (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>), with all hyperparameters are set to default values. To validate the effectiveness of our framework for ASD diagnosis, we perform extensive experiments and compare our feature selection method with state-of-the-art methods. Specifically, we first compare the proposed method with both multi- and single-task learning based methods that use the original GM and WM features. We then compare the proposed method with the competing methods using canonical features, namely CCA group lasso (Zhu et al. 2014) (CCA GL) and CCA sparse group lasso (CCA SGL). We also evaluated the performance of the SVM classifiers when different kernels are used.

Comparison with methods using original GM and WM features

We compare the proposed method with three multi-task learning methods, including (1) group lasso (GL), (2) sparse group lasso (SGL), (3) dirty model (DM) (Jalali et al. 2010), and (4) a single task learning method (lasso) (Tibshirani 1996). Group lasso can only select features jointly across tasks, whereas sparse group lasso selects features by simultaneously imposing l_1 -norm and $l_{2,1}$ -norm on the coefficients. The dirty model separates the coefficients into two parts, and regulates them using l_1 -norm and $l_{\infty,1}$ -norm separately. The single task lasso selects features based on a single target response, i.e., the class label. SVM classifier with polynomial of second degree is used for all except the proposed method, which uses a linear SVM. This is because the polynomial kernel performs better in the original feature space, whereas the linear kernel performs better in the canonical space.

Table 2 presents the performance of all comparison methods in terms of classification accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), and the area under receiver operating characteristic (ROC) curve (AUC). The proposed method achieves the best performance with ACC of 75.4 %, followed by sparse GL (with ACC of 59.11 %), dirty model (with ACC of 58.77 %), lasso (with ACC of 58.04 %), and group lasso (with ACC of 57.49 %). In the meantime, the proposed method achieves SEN of 74.63 %, SPE of 75.96 %, PPV of 74.78 %, NPV of 76.10 %, and AUC of 0.804, which is superior to all competing methods. ROC curves for each of the comparison methods are shown in Fig. 3. The comparison results indicate better diagnosis power of the proposed method than the state-of-the-art methods using the original GM and WM features.

Comparison with methods using canonical features

We compare the proposed method with two CCA-based competing methods, namely CCA group lasso, and CCA sparse group lasso. The former regularizes the ordinary group lasso

with a squared CCA-norm (Kakade and Foster 2007) term $\sum_{j=1}^{D'} \frac{1-r_j}{r_j} \left(\|\hat{\mathbf{w}}^j\|_2^2 + \|\hat{\mathbf{w}}^{j+D'}\|_2^2 \right)$, where r_j denotes the canonical coefficient for the j -th feature. The latter utilizes only canonical correlations to penalize regularization terms of sparse group lasso Eq. (2). The proposed method that considers both canonical and Pearson correlations for feature selection Eq. (4). For both the competing methods, features corresponding to large canonical coefficients are more likely to be selected, while features corresponding to small canonical coefficients tend to be discarded. It should be noted that all comparison methods use the same linear kernel-based SVM classifier.

Table 3 presents the results of all comparison methods. The proposed method achieves the best performance on all statistical measures, followed by the CCA sparse group lasso with ACC of 68.51 %, SEN of 64.63 %, SPE of 72.11 %, PPV of 68.78 %, NPV of 68.32 %, and AUC of 0.73, and the CCA group lasso, which achieves ACC of 61.27 %, SEN of 53.52 %, SPE of 68.6 %, PPV of 61.7 %, and AUC of 0.627. These comparison results also suggest that feature selection methods perform better in the canonical feature space, compared to their counterparts in the original feature space. The ROC curves for comparison methods are displayed in Fig. 4.

Results of SVM classifiers with different kernels

We further evaluated the proposed method using SVM classifiers with different kernels, including the radial basis function (RBF) kernel, polynomial kernel, and linear kernel. Specifically, we first use the proposed method to select informative canonical features, and then, with the selected features, we performed classification using SVM classifiers with different kernels.

Figure 5 shows the results of SVM classifiers using different kernels. Linear kernel achieves the best performance as reported above. Polynomial kernel achieves ACC of 65.35 %, SEN of 60.93 %, SPE of 69.12 %, PPV of 65.47 %, NPV of 65.54 %, and AUC of 0.645, while RBF kernel achieves the worst performance compared to the linear and polynomial kernels, with ACC of 47.34 %, SEN of 36.29 %, SPE of 56.67 %, PPV of 41.09 %, NPV of 48.96 %, and AUC of 0.458.

Discussion

In this study, we proposed a novel multi-task feature selection method via canonical graph matching to assist physicians in ASD diagnosis. From the experiments, the proposed method outperformed all competing methods in both original feature space and canonical space. Also, methods using canonical features generally performed better than their counterparts using original GM and WM features.

Since we have two target responses that are correlated to each other, multi-task learning was employed in our application to fully utilize complementary information among the responses. We used sparse group lasso so that both joint information and task-specific information can be utilized during the learning process. Note that, although SRS_TOTAL score is a clinical measure for ASD diagnosis, features for predicting SRS_TOTAL scores might be different from those for predicting class labels, e.g., SRS_TOTAL scores may be affected by environmental factors. Thus, sparse learning that selects both joint and task-specific features is more flexible for handling multi-task problem in this paper. This is demonstrated by the result in the first experiment that both sparse group lasso and a similar method based on the dirty model outperformed group lasso. From the other hand, group lasso performed even worse than the single-task learning based lasso, demonstrating too strong use of group constraint.

By linearly projecting the original features into a canonical space, the GM and WM features can be represented by canonical coefficients, despite their initially differing distributions. This can be demonstrated by the result in the second experiment that all methods using the canonical features performed better than methods using the original features. Both CCA sparse group lasso and the proposed method used sparse group lasso for multi-task learning. The superiority of the proposed method over CCA sparse group lasso can be attributed to the facts that (1) canonical correlation might not be the only indicator representing the significance of features - Pearson correlation between features and class labels provides additional information, and (2) the application of graph matching among samples preserves the geometric relationships between features, thus improving the performance of the proposed method. It should be noted that CCA group lasso uses a squared CCA norm that

encourages shrinkage of weighting coefficients, instead of sparsity (Zhu et al. 2014). From a supplementary experiment where different types of kernels were used to make prediction based on the features selected by the proposed method, we conclude that nonlinearly projecting features into a space with higher dimension is not helpful for classification.

The performance of the current framework can potentially be improved by first partitioning the population into several homogeneous subpopulations using clustering method. However, we note that the ability of our current framework to select discriminative features provides greater robustness to such heterogeneity. On the other hand, our current study uses only the structural T1-weighted MRI. To further improve performance, we can also include information from other imaging modalities, such as DTI and functional MRI, if these imaging modalities are available in the future studies.

Conclusion

In summary, we have proposed a novel feature selection method for effective ASD diagnosis based on structural MRI. We project GM and WM features onto a canonical space where their correlation is maximized. Then, both canonical correlation between GM and WM features and the Pearson correlation between feature and label are utilized for feature selection from the canonical space, with explicit consideration of the geometric relationships between the samples. To use both joint and task-specific information given by class labels and clinical scores, we also extend sparse group lasso into canonical space for feature selection. The experimental results demonstrated the superiority of the proposed method over state-of-the-art methods.

Acknowledgments

This work was supported partially by NIH grant (EB006733, EB008374, EB009634, MH100217, AG041721, AG042599), and National Natural Science Foundation of China (NSFC) Grants (61473190, 81471743).

References

- Amaral DG, Schumann CM, Nordahl CW. Neuroanatomy of autism. *Trends in Neurosciences*. 2008; 31:137–145. [PubMed: 18258309]
- Boddaert N, Chabane N, Gervais H, Good C, Bourgeois M, Plumet M, Barthelemy C, Mouren M, Artiges E, Samson Y. Superior temporal sulcus anatomical abnormalities in childhood autism: a voxel-based morphometry MRI study. *NeuroImage*. 2004; 23:364–369. [PubMed: 15325384]
- Brambilla P, Hardan A, di Nemi SU, Perez J, Soares JC, Barale F. Brain anatomy and development in autism: review of structural MRI studies. *Brain Research Bulletin*. 2003; 61:557–569. [PubMed: 14519452]
- Chai KMA, Williams CK, Klanke S, Vijayakumar S. Multi-task gaussian process learning of robot inverse dynamics. *Advances in Neural Information Processing Systems*. 2008:265–272.
- Chen, X.; Pan, W.; Kwok, JT.; Carbonell, JG. Accelerated gradient method for multi-task sparse learning problem. 2009. ICDM'09. Ninth IEEE International Conference on Data Mining. IEEE; 2009. p. 746-751.
- Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995; 20:273–297.
- Evgeniou A, Pontil M. Multi-task feature learning. *Advances in Neural Information Processing Systems*. 2007; 19:41.
- Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso. 2010 arXiv preprint arXiv:1001.0736.

- Geschwind DH, Levitt P. Autism spectrum disorders: developmental disconnection syndromes. *Current Opinion in Neurobiology*. 2007; 17:103–111. [PubMed: 17275283]
- Guilmatre A, Dubourg C, Mosca A-L, Legallie S, Goldenberg A, Drouin-Garraud V, Layet V, Rosier A, Briault S, Bonnet-Brilhault F. Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Archives of General Psychiatry*. 2009; 66:947–956. [PubMed: 19736351]
- Hardoon D, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*. 2004; 16:2639–2664. [PubMed: 15516276]
- Jalali A, Sanghavi S, Ruan C, Ravikumar PK. A dirty model for multi-task learning. *Advances in Neural Information Processing Systems*. 2010:964–972.
- Jie, B.; Zhang, D.; Cheng, B.; Shen, D. Manifold regularized multi-task feature selection for multi-modality classification in Alzheimer's disease. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. Springer; 2013. p. 275-283.
- Kakade, SM.; Foster, DP. Multi-view regression via canonical correlation analysis. *Learning Theory*. Springer; 2007. p. 82-96.
- Lim KO, Pfefferbaum A. Segmentation of MR brain images into cerebrospinal fluid spaces, white and gray matter. *Journal of Computer Assisted Tomography*. 1989; 13:588–593. [PubMed: 2745775]
- Liu F, Wee C-Y, Chen H, Shen D. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. *NeuroImage*. 2014; 84:466–475. [PubMed: 24045077]
- Liu, J.; Ji, S.; Ye, J. Multi-task feature learning via efficient $l_{2,1}$ -norm minimization. *Proceedings of the Twenty-Fifth conference on uncertainty in artificial intelligence*. AUAI Press; 2009. p. 339-348.
- Shen D, Davatzikos C. HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging*. 2002; 21:1421–1439. [PubMed: 12575879]
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B Methodological*. 1996:267–288.
- Wang, Y.; Nie, J.; Yap, P-T.; Shi, F.; Guo, L.; Shen, D. Robust deformable-surface-based skull-stripping for large-scale studies. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*. Springer; 2011. p. 635-642.
- Wee C-Y, Yap P-T, Zhang D, Wang L, Shen D. Group-constrained sparse fMRI connectivity modeling for mild cognitive impairment identification. *Brain Structure and Function*. 2014a; 219:641–656. [PubMed: 23468090]
- Wee CY, Wang L, Shi F, Yap PT, Shen D. Diagnosis of autism spectrum disorders using regional and interregional morphological features. *Human Brain Mapping*. 2014b; 35:3414–3430. [PubMed: 25050428]
- Wing L, Gould J. Severe impairments of social interaction and associated abnormalities in children: epidemiology and classification. *Journal of Autism and Developmental Disorders*. 1979; 9:11–29. [PubMed: 155684]
- Xue Y, Liao X, Carin L, Krishnapuram B. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research*. 2007; 8:35–63.
- Zhang D, Shen D. Alzheimer's disease Neuroimaging I. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*. 2012a; 59:895–907. [PubMed: 21992749]
- Zhang D, Shen D. Alzheimer's Disease Neuroimaging I. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PloS One*. 2012b; 7:e33182. [PubMed: 22457741]
- Zhang T, Ghanem B, Liu S, Ahuja N. Robust visual tracking via structured multi-task sparse learning. *International Journal of Computer Vision*. 2013; 101:367–383.
- Zhu, X.; Suk, H-I.; Shen, D. Multi-modality canonical feature selection for Alzheimer's disease diagnosis. *17th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2014*; September 14 2014 – September 18, 2014; Boston: Springer Verlag; 2014. p. 162-169.

Fig. 1.

Flowchart of the proposed ASD diagnosis framework. In the CCA-based GMSGL, both the SRS_TOTAL scores and the class labels are used as target responses. In SVM classification, only the class labels are used as the target responses

Fig. 2. Canonical correlation analysis. The decreasing line widths are indicative of the decreasing correlations between the canonical features (i.e., the rows of $\mathbf{Z}^{(G)}$ and $\mathbf{Z}^{(W)}$)

Fig. 3.

The ROC curves for the proposed method and various methods that use the original features, including dirty model (DM), sparse group lasso (SGL), group lasso (GL), and lasso

Fig. 4. The ROC curves of different methods using canonical features. CCA GL: CCA group lasso; CCA SGL: CCA sparse group lasso

Fig. 5. Effects of different SVM kernels. Note that the AUC value is represented as percentage

Table 1

Demographic information. FIQ: full-scale intelligence quotient; SRS_TOTAL: social responsiveness

	Age (Mean $\pm$ SD)	FIQ score (Mean $\pm$ SD)	SRS_TOTAL (Mean $\pm$ SD)	Male/Female
Autism	10.8 $\pm$ 2.2	107.9 $\pm$ 18.2	91.2 $\pm$ 30.9	47/7
Control	11.3 $\pm$ 2.3	114.4 $\pm$ 13.7	20.5 $\pm$ 12.9	40/17

Table 2

Comparison between various methods in original feature spaces

Method	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	AUC
Lasso	58.04 ± 2.44	51.67 ± 5.20	64.04 ± 4.17	57.66 ± 2.86	58.38 ± 2.43	0.605 ± 0.030
Group lasso	57.49 ± 2.62	54.63 ± 6.37	60.18 ± 3.68	56.43 ± 2.64	58.49 ± 2.96	0.603 ± 0.017
Sparse GL	59.11 ± 4.20	55.56 ± 6.98	62.46 ± 5.30	58.58 ± 3.71	59.70 ± 5.01	0.616 ± 0.035
Dirty model	58.77 ± 3.66	56.67 ± 4.48	60.88 ± 6.66	58.07 ± 4.11	59.67 ± 3.33	0.600 ± 0.038
Proposed method	75.40 ± 1.23	74.63 ± 3.98	75.96 ± 4.00	74.78 ± 2.25	76.10 ± 2.13	0.804 ± 0.022

Table 3

Comparison between the proposed method and two competing methods (CCA group lasso, and CCA sparse group lasso) in canonical feature spaces

Method	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	AUC
CCA GL	61.27 ± 2.33	53.52 ± 6.492	68.60 ± 3.548	61.70 ± 2.074	61.07 ± 2.64	0.627 ± 0.042
CCA SGL	68.51 ± 1.49	64.63 ± 3.15	72.11 ± 3.18	68.78 ± 2.15	68.32 ± 1.64	0.730 ± 0.025
Proposed method	75.40 ± 1.23	74.63 ± 3.98	75.96 ± 4.00	74.78 ± 2.25	76.10 ± 2.13	0.804 ± 0.022