HHS Public Access

Author manuscript

Med Image Comput Comput Assist Interv. Author manuscript; available in PMC 2015 June 13.

Published in final edited form as:

Med Image Comput Comput Assist Interv. 2014 ; 17(0 2): 162–169.

Multi-modality Canonical Feature Selection for Alzheimer's Disease Diagnosis

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

Department of Radiology and BRIC, University of North Carolina at Chapel Hill, USA

Dinggang Shen: dgshen@med.unc.edu

Abstract

Feature selection has been commonly regarded as an effective method to lessen the problem of high dimension and low sample size in medical image analysis. In this paper, we propose a novel multimodality canonical feature selection method. Unlike the conventional sparse Multi-Task Learning (MTL) based feature selection method that mostly considered only the relationship between target response variables, we further consider the correlations between features of different modalities by projecting them into a canonical space determined by canonical correlation analysis. We call the projections as *canonical representations*. By setting the canonical representations as regressors in a sparse least square regression framework and by further penalizing the objective function with a new canonical regularizer on the weight coefficient matrix, we formulate a multi-modality canonical feature selection method. With the help of the canonical information of canonical representations and also a canonical regularizer, the proposed method selects canonical-cross-modality features that are useful for the tasks of clinical scores regression and multi-class disease identification. In our experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, we combine Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) images to jointly predict clinical scores of Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) and Mini-Mental State Examination (MMSE) and also identify multiclass disease status for Alzheimer's disease diagnosis.

1 Introduction

In medical image analysis, the feature dimensionality usually overwhelms the number of available samples. Thus, it is common to apply dimension reduction methods to tackle this so-called 'High Dimension, Low Sample Size' (HDLSS) problem. Among various methods in the literature, the sparse least square regression model has proven its effectiveness for the HDLSS problem in many applications. For example, Liu *et al.* proposed an $\ell_{2,1}$-norm regularized regression model for joint feature selection via Multi-Task Learning (MTL) [4] and Zhang *et al.* recently applied the method for the tasks of clinical status identification and clinical scores prediction with respect to Alzheimer's Disease (AD) diagnosis [10]. Mathematically, the sparse MTL model can be formulated as follows:

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

$$\min_{\mathbf{W}} \frac{1}{2}\|\boldsymbol{Y} - \mathbf{W}^T\mathbf{X}\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} \quad (1)$$

where **Y, X**, and **W** denote, respectively, a response matrix, a regressor matrix, and a weight coefficient matrix, and α is a sparsity control parameter. The $\ell_{2,1}$-norm $\|\mathbf{W}\|_{2,1}$ leads to the sparsity with respect to the rows in **W**, *i.e.*, discarding features when their respective weight coefficients in the rows are equal to or close to zeros [12]. The beauty of the MTL is that it effectively considers the relation among tasks in selecting features that can be jointly used across the tasks. However, because the data distributions of different modalities in the original spaces can be complex and heterogenous, it is limited for the conventional sparse MTL to properly fuse multiple modalities, *e.g.*, Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), which has been shown to be useful in AD diagnosis [6,9,13], since it has no way of using the correlation between modalities.

In this paper, we propose a novel canonical feature selection method that can efficiently integrate the correlational information between modalities into a sparse MTL along with a new regularizer. Specifically, we first transform the features in an ambient space into a canonical feature space spanned by the canonical bases obtained by Canonical Correlation Analysis (CCA) and then perform the sparse MTL with an additional canonical regularizer by having the canonical features as new regressors. The rationale for imposing the correlation information into our method is that the structural and functional images of a subject are highly correlated to each other [5,7,10] and by explicitly projecting the features of these modalities into a common space, where their correlation becomes maximized via CCA, we can help select more task-related features. We justify the effectiveness of the proposed method by applying it to the tasks of regressing clinical scores of Alzheimer's Disease Assessment Scale Cognitive (ADAS-Cog) and Mini-Mental State Examination (MMSE), and identifying a multi-stage status, *i.e.*, AD, Mild Cognitive Impairment (MCI), or Normal Control (NC), on ADNI dataset.

2 Method

In this section, we describe a novel canonical feature selection method that integrates the ideas of CCA and a sparse MTL into a unified framework.

2.1 Canonical Correlation Analysis (CCA)

Assume that we have *d*-dimensional samples from two different modalities of *n* samples: $\mathbf{X}^{(1)} \in \mathbb{R}^{d\times n}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{d\times n}$. Let $\mathbf{X} = [\mathbf{X}^{(1)},\mathbf{X}^{(2)}] \in \mathbb{R}^{d\times 2n}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$ denote a multi-modal feature matrix and its covariance matrix, respectively. CCA is a classical method to find correlations between two multivariate random variables, *i.e.*, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Specifically, it seeks two sets of basis vectors $\mathbf{B}^{(1)} \in \mathbb{R}^{d\times d}$ and $\mathbf{B}^{(2)} \in \mathbb{R}^{d\times d}$ such that the correlations between the projections of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ onto the new space spanned by these basis vectors are mutually maximized [2,11] as follows:

$$(\hat{\mathbf{B}}^{(1)}, \hat{\mathbf{B}}^{(2)}) = \arg \max_{(\mathbf{B}^{(1)}, \mathbf{B}^{(2)})} \frac{\mathbf{B}^{(1)^T} \boldsymbol{\Sigma}_{12} \mathbf{B}^{(2)}}{\sqrt{\mathbf{B}^{(1)^T} \boldsymbol{\Sigma}_{11} \mathbf{B}^{(1)}} \sqrt{\mathbf{B}^{(2)^T} \boldsymbol{\Sigma}_{22} \mathbf{B}^{(2)}}}. \quad (2)$$

The optimal solution ($\hat{\mathbf{B}}^{(1)}$, $\hat{\mathbf{B}}^{(2)}$) can be effectively obtained by a generalized eigen-decomposition [1,11]. The projections of the original features onto their respective canonical bases can be considered as new representations, which we call '*canonical representations*':

$$\boldsymbol{Z}^{(m)} = \hat{\mathbf{B}}^{(m)^T} \mathbf{X}^{(m)}, \quad (3)$$

where $\boldsymbol{Z}^{(m)} = \left[\mathbf{z}_j^{(m)^T}\right]_{j=1:d} \in \mathbb{R}^{d \times n}, \mathbf{z}_j^{(m)} \in \mathbb{R}^n$, and $m \in \{1, 2\}$.

2.2 Canonical Feature Selection

CCA ensures the projections of the original features, *i.e.*, canonical representations, to be maximally correlated across modalities in a new space. Moreover, according to the recent work by Kakade and Foster [3], it was shown that a model can more precisely fit data with the guidance of the canonical information between modalities. Inspired by these favorable characteristics of canonical representations, we propose a new feature selection method by exploring the correlations of multimodal features in a canonical space and defining a new canonical regularizer.

Let $\mathbf{Y} \in \mathbb{R}^{c \times n}$ denote a response matrix, where c is the number of the response variables[1]. We first formulate a sparse multi-class linear regression model in an MTL framework by setting the regressors with our canonical representations as follows:

$$\min_{\mathbf{W}} \frac{1}{2} \|\boldsymbol{Y} - \mathbf{W}^T \boldsymbol{Z}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} \quad (4)$$

where $\mathbf{Z} = [\mathbf{Z}^{(1)}; \mathbf{Z}^{(2)}] \in \mathbb{R}^{2d \times n}$, $\mathbf{W} \in \mathbb{R}^{2d \times c}$ is a regression coefficient matrix, and β is a tuning parameter controlling the row-wise sparsity of $\mathbf{W}$. It should be emphasized that Eq. (4) considers not only the relationships among response variables, thanks to the $\ell_{2,1}$ norm of the weight coefficient matrix, but also the correlations across modalities by means of the canonical representations $\mathbf{Z}$.

A canonical norm over a vector $\mathbf{p} = [p_j] \in \mathbb{R}^d$ is defined [3]:

$$\|\mathbf{p}\|_{CCA} = \sqrt{\sum_{j=1}^{d} \frac{1 - \lambda_j}{\lambda_j} (p_j)^2}. \quad (5)$$

where $\{\lambda_j\}_{j=1:d}$ denotes a set of canonical correlation coefficients.

[1] In our work, the response variables correspond to ADAS-Cog, MMSE, and a multiclass label. For the class labels, we use a 0/1 encoding.

Based on this definition, we devise a new canonical regularizer over a weight coefficient matrix **W** as follows:

$$\|\mathbf{W}\|_{CCA}^2=\sum_{i=1}^{2d}\|\mathbf{w}^i\|_{CCA}^2=\sum_{i=1}^{d}\frac{1-\lambda_i}{\lambda_i}\sum_{j=1}^{c}\left\{(w_{i,j})^2+(w_{i+d,j})^2\right\} \quad (6)$$

where $\mathbf{w}^i$ denotes the *i*-th row vector in **W**.

It is noteworthy that our canonical regularizer in Eq. (6) enforces the highly correlated canonical representations across modalities, *i.e.*, large canonical correlation coefficients, to be selected while the merely or uncorrelated canonical representations across modalities, *i.e.*, small canonical correlation coefficients, to be unselected. Specifically, the large λ_i leads to the large weight on $\mathbf{w}^i$ in the optimization process.

By further penalizing the objective function in Eq. (4) with our canonical regularizer, we finally define the proposed feature selection model as follows:

$$\min_{\mathbf{W}}\frac{1}{2}\|\boldsymbol{Y}-\mathbf{W}^T\boldsymbol{Z}\|_F^2+\beta\|\mathbf{W}\|_{2,1}+\gamma\|\mathbf{W}\|_{CCA}^2 \quad (7)$$

where β and γ are tuning parameters. To find the optimal solution of Eq. (7), which is convex but non-smooth, we use the accelerated proximal gradient method [4].

The main contributions on this method can be as follows: Unlike the conventional sparse MTL-based feature selection method [10] that employed a least square loss function with regressors of the original features in an ambient space, we use a canonical loss function by using canonical representations as regressors. Thus, we can say that our loss function is more predictive than the conventional sparse MTL by achieving relatively smaller estimation error [3]. In this regard, we argue that the selected features by the proposed method are more powerful in predicting the target response variables than the conventional sparse MTL-based feature selection method. Besides, thanks to the flexibility of CCA, it is also possible to apply the proposed method to consider the correlation between features and the response matrix. Later, in our experiments, we apply this supervised approach for a single modality.

3 Experimental Results

We performed experiments on a subset of the ADNI dataset (‘www.adni-info.org’) to justify the effectiveness of the proposed method.

3.1 Preprocessing and Feature Extraction

We used baseline MRI and PET images of 202 subjects including 51 AD, 43 MCI Converter (MCI-C), 56 MCI Non-Converter (MCI-NC), and 52 NC. We preprocessed the MRI and PET images by applying spatial distortion, skull-stripping, and cerebellum removal, sequentially. We then segmented MRI images into gray matter, white matter, and cerebrospinal fluid, and warped into a template that dissected a human brain into 93 regions. For the PET images, we aligned them to their respective MRI images. We obtained 93 gray

matter volumes from a MRI image and 93 mean intensities from a PET image and used them as features.

3.2 Experimental Setting

In our experiments, we considered the Joint clinical scores Regression and Multiclass AD status Identification (JRMI) problem. Depending on the separation of the MCI status, *i.e.*, MCI-C and MCI-NC, two different JRMI problems were performed: AD vs. MCI vs. NC (3-JRMI) and AD vs. MCI-C vs. MCI-NC vs. NC (4-JRMI). In the 3-JRMI problem, the samples of MCI-C and MCI-NC were regarded as MCI. To show the validity of the proposed method, we compared with the dimension reduction methods of Fisher Score (FS) [1], Principal Component Analysis (PCA) [1], CCA [2], Sparse Joint Classification and Regression (SJCR) [8], and Multi-Modal Multi-Task (M3T) [10].

For each JRMI problem, we followed the steps of (1) selecting features by each of the competing methods, (2) learning two support vector regression models for ADAS-Cog and MMSE, respectively, and a support vector classification model for disease status identification using an LIBSVM toolbox publicly available at 'http://www.csie.ntu.edu.tw/~cjlin/libsvm/', and (3) evaluating the performance of the methods with the metrics of correlation coefficient in regression and the classification accuracy in classification.

3.3 Results and Discussions

Multi-Modality—For fusing the modalities of MRI and PET, we concatenated the feature vectors of two modalities into a single long vector for the methods of FS, PCA, and CCA. Meanwhile, for M3T and SJCR, we followed the corresponding literature to regard each modality as a group. Note that unlike these competing methods, our method operated with canonical representations rather than the original feature vectors. Table 1 shows the performances of all the methods. The proposed method achieved the best performances for regression and classification in both 3-JRMI and 4-JRMI problems. For example, in the 3-JRMI problem, the proposed method achieved improvements of 10.1% on ACC, 0.045 on CC-A, and 0.081 on CC-M, respectively, compared to the FS method, and 5.3% (ACC), 0.024 (CC-A), and 0.02 (CC-M), compared to SJCR that achieved the overall best performance among the competing methods. For the 4-JRMI task, the performance improvement by the proposed method is 10.2% (ACC), 0.062 (CC-A), and 0.127 (CC-M), respectively, compared to the worst competing method FS, and 7.4% (ACC), 0.032 (CC-A), and 0.039 (CC-M), compared to the best competing method M3T. These experimental results demonstrated that the use of canonical information, *i.e.*, canonical representations and the canonical regularizer, in the proposed method helped improve the performances in the JRMI problems. It is also noteworthy that in Table 1, CCA, in which we obtained the canonical representations based on the correlation between fused multi-modal feature vectors and the response matrix, outperformed the unsupervised methods of FS and PCA.

Single-Modality—We also applied the proposed multi-modality feature selection method for single-modality feature selection by performing CCA between single-modality features and the response matrix, and then using the canonical representations of the feature and response matrices in Eq. (7). We should note that there is no mathematical change in the

proposed method for this single-modality experiment. We conducted feature selection with all the methods on the single-modality data (MRI or PET) for the JRMI problem and reported the experimental results in Table 2 and Table 3.

Again, the proposed method achieved the best performance in both JRMI problems on single-modality data. Specifically, in the 3-JRMI problem, our proposed method with MRI improved by 1.7% (ACC), 0.011 (CC-A), and 0.012 (CC-M), compared to the best performance of all the competing methods with MRI. In the 4-JRMI problem, our proposed method with PET improved by 3.3% (ACC), 0.019 (CC-A), and 0.013 (CC-M), compared to the best performance of all the competing methods with PET. Based on these experimental results, we can see that the proposed multi-modality feature selection method can also work well for single-modality data with the canonical information between features and the response matrix by achieving better performance than the competing methods.

Discussions—From Table 2 and Table 3, the methods with MRI showed better performances than those with PET. For example, in the 4-JRMI problem, the proposed method with MRI improved by 2.0% (ACC), 0.053 (CC-A), and 0.041 (CC-M), respectively, compared to that with PET. Meanwhile, the performances of all the methods with bi-modality were better than those of all the methods with an MRI or PET. For example, in the 3-JRMI problem, the proposed method on bi-modality data made improvements by 8.0% (ACC), 0.039 (CC-A), and 0.086 (CC-M), compared to our method with PET, and by 8.8% (ACC), 0.031 (CC-A), 0.038 (CC-M), compared to the best performances of the competing methods with MRI or PET. Moreover, the proposed method improved by about 6.74%, 4.02%, and 5.70% compared to the method of conducting feature selection on the features concatenating MRI and PET, the method of separately conducting PCA on MRI and PET before classification, and the method without the canonical norm term, *i.e.*, Eq. (7) without the last term, respectively.

4 Conclusions

In this work, we proposed a novel feature selection method for joint regression and multi-class classification. To circumvent the conventional sparse MTL-based feature selection method, we proposed a canonical feature selection method by explicitly using the correlation between modalities. Specifically, we discovered canonical representations of the original inputs by projecting them into a new space spanned by the canonical bases obtained by CCA. In a sparse MTL framework, we set the regressors with our canonical representations and further penalized it with a newly defined canonical regularizer of the weight coefficient matrix. In our experiments on the ADNI dataset, we achieved the best performances for the joint clinical scores regression and multi-class clinical status identification. We should note that the proposed method is limited to bi-modality fusion. By fusion of more than two modalities, we may be able to further enhance the diagnostic accuracy. This will be our forthcoming research topic, *i.e.*, by integrating the multi-view CCA [2] to our framework.

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. Duda, RO.; Hart, PE.; Stork, DG. Pattern classification. John Wiley & Sons; 2012.
2. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods. Neural Computation. 2004; 16(12):2639–2664. [PubMed: 15516276]
3. Kakade, SM.; Foster, DP. Multi-view regression via canonical correlation analysis. In: Bshouty, NH.; Gentile, C., editors. COLT. LNCS (LNAI). Vol. 4539. Heidelberg: Springer; 2007. p. 82-96.
4. Liu J, Ji S, Ye J. Multi-task feature learning via efficient l2, 1-norm minimization. UAI. 2009:339–348.
5. May A, Ashburner J, Büchel C, McGonigle D, Friston K, Frackowiak R, Goadsby P. Correlation between structural and functional changes in brain in an idiopathic headache syndrome. Nature Medicine. 1999; 5(7):836–838.
6. Perrin RJ, Fagan AM, Holtzman DM. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. Nature. 2009; 461:916–922. [PubMed: 19829371]
7. 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., editors. MICCAI 2013, Part II. LNCS. Vol. 8150. Heidelberg: Springer; 2013. p. 583-590.
8. Wang, H.; Nie, F.; Huang, H.; Risacher, S.; Saykin, AJ.; Shen, L. Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In: Fichtinger, G.; Martel, A.; Peters, T., editors. MICCAI 2011, Part III. LNCS. Vol. 6893. Heidelberg: Springer; 2011. p. 115-123.
9. Westman E, Muehlboeck JS, Simmons A. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. NeuroImage. 2012; 62(1):229–238. [PubMed: 22580170]
10. Zhang D, Shen D. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage. 2012; 59(2):895–907. [PubMed: 21992749]
11. Zhu X, Huang Z, Shen HT, Cheng J, Xu C. Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recognition. 2012; 45(8):3003–3016.
12. Zhu X, Huang Z, Yang Y, Shen HT, Xu C, Luo J. Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recognition. 2013; 46(1):215–229.
13. Zhu X, Suk HI, Shen D. Matrix-similarity based loss function and feature selection for Alzheimer's disease diagnosis. CVPR. 2014

Table 1

Comparison of the performances of all methods with bi-modality data on two classification problems (3-JRMI and 4-JRMI).

Method	3-JRMI			4-JRMI		
	ACC	CC-A	CC-M	ACC	CC-A	CC-M
FS	0.628±1.30	0.695±0.16	0.594±0.14	0.517±1.54	0.491±0.34	0.440±0.16
PCA	0.646±1.59	0.698±0.12	0.599±0.13	0.525±2.21	0.510±0.56	0.453±0.47
CCA	0.648±1.81	0.702±0.22	0.602±0.19	0.536±1.29	0.522±0.25	0.462±0.53
SJCR	0.676±1.68	0.716±0.38	0.655±0.37	0.559±1.64	0.538±0.85	0.493±0.33
M3T	0.679±1.67	0.709±0.92	0.647±0.28	0.545±1.61	0.521±0.71	0.528±0.32
Proposed	**0.729±1.38**	**0.740±0.18**	**0.675±0.23**	**0.619±1.54**	**0.553±0.34**	**0.567±0.23**

(ACC: classification ACCuracy, CC-A: Correlation Coefficient for ADAS-Cog, and CC-M: Correlation Coefficient for MMSE)

Table 2

Comparison of the performances of all methods with MRI on two classification problems (3-JRMI and 4-JRMI)

Method	3-JRMI			4-JRMI		
	ACC	CC-A	CC-M	ACC	CC-M	CC-M
FS	0.621±1.53	0.682±0.37	0.572±0.16	0.508±1.70	0.483±0.37	0.437±0.90
PCA	0.637±4.29	0.690±0.30	0.578±0.19	0.510±1.23	0.491±0.24	0.444±0.18
CCA	0.640±2.31	0.699±0.53	0.589±0.21	0.527±2.43	0.504±0.25	0.452±0.21
SJCR	0.641±1.36	0.709±0.38	0.637±0.75	0.531±1.67	0.514±0.27	0.473±0.75
M3T	0.639±1.57	0.705±0.42	0.626±0.39	0.529±1.62	0.513±0.55	0.489±0.83
Proposed	**0.658±1.31**	**0.720±0.45**	**0.649±0.26**	**0.563±0.89**	**0.523±0.34**	**0.503±0.26**

Table 3

Comparison of the performances of all methods with PET on two classification problems (3-JRMI and 4-JRMI)

Method	3-JRMI			4-JRMI		
	ACC	CC-A	CC-M	ACC	CC-M	CC-M
FS	0.593±1.54	0.671±0.40	0.560±0.84	0.494±1.49	0.410±0.14	0.414±0.33
PCA	0.617±1.58	0.688±0.72	0.562±0.33	0.495±1.61	0.401±0.75	0.423±0.29
CCA	0.620±1.22	0.696±0.30	0.574±0.96	0.498±1.38	0.439±0.19	0.438±0.64
SJCR	0.621±1.73	0.670±0.57	0.578±0.46	0.508±1.69	0.438±0.32	0.437±0.32
M3T	0.624±1.48	0.690±0.64	0.582±0.77	0.510±1.67	0.451±0.35	0.449±0.43
Proposed	**0.649±1.36**	**0.701±0.19**	**0.589±0.30**	**0.543±1.21**	**0.470±0.27**	**0.462±0.21**