View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IUPUIScholarWorks

HHS Public Access

Author manuscript

Inf Process Med Imaging. Author manuscript; available in PMC 2018 June 01.

Published in final edited form as:

Inf Process Med Imaging. 2017 June ; 10265: 198–209. doi:10.1007/978-3-319-59050-9_16.

Predicting Interrelated Alzheimer's Disease Outcomes via New Self-Learned Structured Low-Rank Model

Xiaoqian Wang[1], **Kefei Liu**[2,3], **Jingwen Yan**[2,3], **Shannon L. Risacher**[2], **Andrew J. Saykin**[2], **Li Shen**[2], **Heng Huang**[1,*], and **ADNI**[**]

[1]Computer Science & Engineering, University of Texas at Arlington, TX, 76019, USA

[2]Radiology & Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, 46202, USA

[3]BioHealth, Indiana University School of Informatics & Computing, Indianapolis, IN, 46202, USA

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder. As the prodromal stage of AD, Mild Cognitive Impairment (MCI) maintains a good chance of converting to AD. How to efficaciously detect this conversion from MCI to AD is significant in AD diagnosis. Different from standard classification problems where the distributions of classes are independent, the AD outcomes are usually interrelated (their distributions have certain overlaps). Most of existing methods failed to examine the interrelations among different classes, such as AD, MCI conversion and MCI non-conversion. In this paper, we proposed a novel self-learned low-rank structured learning model to automatically uncover the interrelations among different classes and utilized such interrelated structures to enhance classification. We conducted experiments on the ADNI cohort data. Empirical results demonstrated advantages of our model.

Keywords

Alzheimer's Disease; MCI Conversion Prediction; Structured Low-Rank Model

1 Introduction

Alzheimer's Disease (AD) usually progresses along a temporal continuum, initially from a preclinical stage, subsequently to mild cognitive impairment (MCI) and ultimately deteriorating to AD [19]. As the transitional step between normal aging and dementia, MCI has attracted high attention since it provides promising opportunities for early detection of AD. MCI is recognized as a clinical state of individuals who are memory impaired but functioning well otherwise, which does not meet the clinical criteria for dementia [13].

[*]To whom correspondence should be addressed. At UTA, this work was partially supported by NIH R01 AG049371, NSF IIS 1302675, NSF IIS 1344152, NSF DBI 1356628, NSF IIS 1619308, NSF IIS 1633753. At IU, this work was partially supported by NIH R01 EB022574, R01 LM011360, U01 AG024904, P30 AG10133, and R01 AG19771.

[**]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_AcknowledgementList.pdf.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

According to [11], MCI patients preserve a conversion-to-AD rate of approximately 15% per year, thus it is of great importance to distinguish MCI patients with high potential of AD conversion from those not years before dementia.

Recent advances in neuroimaging have offered a helping hand for exploring associations between brain structure and behavior, which have provided effective features for early detection of AD[7,8]. In the past few years, several machine learning techniques have been applied to predict MCI conversion by means of neuroimaging data [12]. Researches utilized various classification models to identify MCI converters from other classes, *e.g.*, health control samples and MCI non-converters by adopting neuroimaging data only in baseline time, which indicated a promising approach of "forecasting" stage changes of MCI patients several years before the conversion happens. As successful early detection of MCI conversion can boost therapeutic intervention of AD to a large extent, studies on this topic have attracted high attention in recent time.

However, most existing models hold a simple and common assumption that the neuroimaging data is drawn from an unimodal distribution [11,12,18,17,16], which is not applicable for all occasions. In AD research, since MCI converters and AD eventually evolve to AD with certain common biological mechanism, it is reasonable to assume that these subjects share similar distribution patterns, but their distributions are distinct from that of health control samples. That is to say, the brain data may come from multimodal distribution, *e.g.*, mixture of Gaussian. Thus, it is natural to assume latent group structure exists among different classes. Discovery of such subspace structure can enhance MCI conversion prediction and improve image biomarker discovery.

The most straightforward way to discover such groupwise interrelations is to cluster different data into groups before classification. However, since the clustering step is detached with the classification model, the learned interrelation structures are not associated to the prediction results. Such separated steps usually lead to suboptimal result. Here, we propose a novel structured low-rank learning model to simultaneously uncover the interrelations among different diagnostic stages and employ such interrelated structures to enhance the prediction of MCI conversion. We adopt Schatten p-norm to identify the shared low-rank subspace. Our new model is applied to the ADNI cohort for MCI conversion prediction. All empirical results show that the proposed classification model is capable of predicting MCI conversion with better performance.

2 Self-Learned Low-Rank Structured Classification Model

Multi-class classification problem with c classes can be seen as a multi-task learning problem with c tasks, where each task is to classify one class from all others via the one-vs-rest technique. Suppose these c tasks come from g groups, where tasks in each group are mutually related. We introduce and optimize a group index matrix set $Q = \{Q_1, Q_2, \ldots Q_g\}$ to discover this group structure. Each Q_i is a diagonal matrix with $Q_i \in \{0, 1\}^{c \times c}$ showing the assignment of tasks to the i-th group. For the (k, k)-th element of Q_i, $(Q_i)_{kk} = 1$ means the k-th task belongs to the i-th group while $(Q_i)_{kk} = 0$ means not. To avoid overlap of groups, we have $\sum_{i=1}^{g} Q_i = I$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Since each group of tasks share correlative dependence, we reasonably assume the latent subspace of each group maintains a low-rank structure. Schatten-p norm [10] can be used as a low-rank regularization for uncovering common subspaces shared by tasks.

For a matrix $A \in \mathbb{R}^{d\times n}$, suppose σ_i is its i-th singular value, then the rank of A can be written as $\mathrm{rank}(A)=\sum_{i=1}^{\min\{d,n\}}\sigma_i^0$, where $0^0 = 0$. The definition of p-th power Schatten p-norm ($0 < p < \infty$) is:

$$\|A\|_{S_p}^p = Tr((A^TA)^{\frac{p}{2}}) = \sum_{i=1}^{\min\{d,n\}}\sigma_i^p.$$

The well-known trace norm (*a.k.a.* nuclear norm) is a special case of Schatten p-norm with $p = 1$: $\|A\|_* = \|A\|_{S_1} = Tr((A^TA)^{\frac{1}{2}}) = \sum_{i=1}^{\min\{d,n\}}\sigma_i$.

Obviously, when $0 < p < 1$, Schatten p-norm makes a better approximation of $rank(A)$ thus a more strict low-rank constraint than trace norm. The more closer p is 0, the more strict low-rank constraint the regularization term $\|A\|_{S_p}^p$ imposes.

According to the above analysis, we can formulate our novel self-learned structured low-rank classification model as follows:

$$\min_{W,\mathbf{b},Q_i|_{i=1}^g\in\{0,1\}^{c\times c},\sum_{i=1}^g Q_i=I} \mathscr{L}(Y;X,W,\mathbf{b})+\gamma\sum_{i=1}^g(\|WQ_i\|_{S_p}^p)^k \quad (1)$$

In Problem (1), we use a general classification loss $\mathscr{L}(Y; X, W, \mathbf{b})$, which can be any loss function, *e.g.*, logistic regression, hinge loss, *etc.* $W \in \mathbb{R}^{d\times c}$ is the weight matrix for classification, $\mathbf{b} \in \in^{c\times 1}$ is the bias, and $Y \in \mathbb{R}^{n\times c}$ is the label matrix. Moreover, we add a power parameter k to the Schatten p-norm regularization term for robustness of Problem (1), whose influence will be elaborately discussed in Section 4.

When $0 < p < 1$, it is apparent that the new objective is non-convex thus difficult for optimization. In the next section, we adopt an efficient re-weighted optimization algorithm.

3 Optimization Algorithm

Here, we first introduce a re-weighted algorithm to solve a general problem where Problem (1) is a special case, and then talk about the detailed optimization of (1).

3.1 Optimization Algorithm for A General Problem

Lemma 1. Let $g_i(x)$ denote a general function over x, where x can be a scalar, vector or matrix, $\mathscr{C}$ denotes the constraints on x, then we can claim:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

When $\delta \to 0$, The optimization problem

$$\min_{x \in C} f(x) + \sum_i Tr((g_i^T(x) g_i(x))^{\frac{p}{2}}),$$

is equivalent to

$$\min_{x \in C} f(x) + \sum_i Tr(g_i^T(x) g_i(x) D_i), \quad \text{where} \quad D_i = \frac{p}{2}(g_i^T(x) g_i(x) + \delta I)^{\frac{p-2}{2}}.$$

Proof: When $\delta \to 0$, it's apparent that the optimization problem

$$\min_{x \in C} f(x) + \sum_i Tr((g_i^T(x) g_i(x) + \delta I)^{\frac{p}{2}}), \tag{2}$$

will reduce to

$$\min_{x \in C} f(x) + \sum_i Tr((g_i^T(x) g_i(x))^{\frac{p}{2}}). \tag{3}$$

So with a fairly small parameter δ, we turn the non-smooth Problem (3) to the smooth Problem (2).

The Lagrangian function of Problem (2) is:

$$f(x) + \sum_i Tr((g_i^T(x) g_i(x) + \delta I)^{\frac{p}{2}}) - \lambda \tilde{I}_C(x),$$

where $\tilde{I}_C(x)$ equals 0 if $x \in C$ and ∞ otherwise [4]. Take derivative w.r.t. x and set it to zero. Based on the chain rule [2], we have:

$$\sum_i \frac{Tr\left(2\frac{p}{2}(g_i^T(x) g_i(x) + \delta I)^{\frac{p-2}{2}} g_i^T(x) \partial g_i(x)\right)}{\partial x} + f'(x) - \lambda \frac{\partial \tilde{I}_C(x)}{\partial x} = 0. \tag{4}$$

According to the Karush-Kuhn-Tucker conditions [4], if we can find a solution x that satisfies Eq. (4), then we usually find a local/global optimal solution to Problem (2). However, it is intractable to directly find the solution x that satisfies Eq. (4). Here we came up with a strategy as follows:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

If $D_i=\frac{p}{2}(g_i^T(x)g_i(x)+\delta I)^{\frac{p-2}{2}}$ is a given constant, Eq. (4) can be reduced to

$$f'(x)+\sum_i \frac{Tr(2Dg_i^T(x)\partial g_i(x))}{\partial x} - \lambda \frac{\partial \overset{\grave{E}ij}{I}_C(x)}{\partial x}=0. \quad (5)$$

Based on the chain rule [2], the optimal solution x^* of Eq. (5) is also an optimal solution to the following problem:

$$\min_{x\in C} f(x)+\sum_i Tr(g_i^T(x)g_i(x)D_i). \quad (6)$$

Based on this observation, we can first guess a solution x, next calculate D_i based on the current solution x, and then update the current solution x by the optimal solution of Problem (6) on the basis of the calculated D_i. We can iteratively perform this procedure until it converges.

3.2 Optimization of Problem (1)

It is obvious that Problem (1) can be optimized via Lemma 1. Noticing that $Q_iQ_i^T=Q_i$, our objective becomes:

$$\min_{W\mathbf{b},Q_i|_{i=1}^g\in\{0,1\}^{c\times c},\sum_{i=1}^g Q_i=I} \mathscr{L}(Y;X,W,\mathbf{b})+\gamma\sum_{i=1}^g Tr(WQ_iW^TD_i), \quad (7)$$

where D_i is defined as:

$$D_i=\frac{kp}{2}(\|WQ_i\|_{S_p}^p)^{k-1}(WQ_iW^T+\delta I)^{\frac{p-2}{2}}, \quad (8)$$

with δ being a fairly small parameter close to zero.

We can solve Problem (7) by means of the alternating optimization method.

The first step is fixing W and solving Q, then Problem (7) becomes:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\min_{Q_i|_{i=1}^{g}\in\{0,1\}^{c\times c},\sum_{i=1}^{g}Q_i=I}\sum_{i=1}^{g}Tr((W^TD_iW)Q_i), \tag{9}$$

Let $A_i = W^TD_iW$, then the solution of each Q_i is evident as follows:

$$Q_i(l,l)=\begin{cases}1, & i=\arg\min_j A_j(l,l)\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

The second step is fixing Q and solving W, **b**, then Problem (7) becomes:

$$\min_{W,\mathbf{b}}\mathscr{L}(Y;X,W,\mathbf{b})+\gamma\sum_{i=1}^{g}Tr(WQ_iW^TD_i). \tag{11}$$

Problem (11) can be solved according to the choice of the classification loss $\mathscr{L}(Y; X, W, \mathbf{b})$.

Here, we take an example to illustrate the optimization steps of Problem (11) when we adopt hinge loss for $\mathscr{L}(Y; X, W, \mathbf{b})$. Problem (11) can be written as:

$$\min_{W,\mathbf{b}}C\sum_{i=1}^{n}\sum_{j=1}^{c}h_{ij}\left(1-y_{ij}(\mathbf{w}_j^T\mathbf{x}_i+b_j)\right)+\frac{1}{2}\|W\|_F^2+\gamma\sum_{i=1}^{g}Tr(WQ_iW^TD_i). \tag{12}$$

where $H\in\mathbb{R}^{n\times c}$ is a slack variable defined as follow:

$$h_{ij}=\begin{cases}1, & y_{ij}(\mathbf{w}_j^T\mathbf{x}_i+b_j)\le 1\\ 0, & \text{otherwise}\end{cases} \tag{13}$$

Take derivative of Problem (12) w.r.t. **b** and set it to zero, then we get: $\sum_{i=1}^{n}h_{ij}y_{ij}=0$. which indicates that **b** can be updated according to the support vectors.

Take derivative of Problem (12) w.r.t. $\mathbf{w}_j$ and set it to zero, then we get:

$$\mathbf{w}_j=C(I+2\gamma(\sum_{i=1}^{g}Q_i(j,j)D_i))^{-1}\sum_{i=1}^{n}h_{ij}y_{ij}\mathbf{x}_i.$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We can iteratively update D, Q, W and **b** with the alternating steps mentioned above and the algorithm of Problem (7) is summarized in Algorithm 1.

Convergence and time analysis—Our algorithm as a whole employs the alternating optimization method to update variables, whose convergence has already been proved in [3]. Our model usually converges in 15 iterations. In our experiments on the ADNI data, the runtime for five-fold cross validation is around 3 seconds.

Algorithm 1 Algorithm to solve problem (7).

Input:

Imaging feature data $X \in \mathbb{R}^{d \times n}$, label matrix $Y \in \mathbb{R}^{n \times c}$, parameter p, k, γ, group number g.

Output:

Weight matrix $W \in \mathbb{R}^{d \times c}$ and g different group matrices $Q_i|_{i=1}^{g} \in \mathbb{R}^{c \times c}$ which groups the c classes into g subspaces.

Initialize W by the optimal solution to the ridge regression problem.

Initialize Q randomly.

while not converge **do**

1. Update D according to the definition in Eq. (8)
2. Update Q according to the solution in Eq. (10)
3. Update W and **b** by solving Problem (11). The solution differs w.r.t. the choice of loss function $\mathcal{L}(Y, X, W, \mathbf{b})$.

end while

4 Discussion of Parameters

We introduced several hyper-parameters to make our model more general and adaptive to various circumstances. Here, we analyze the functionality of each parameter in detail.

In Problem (1), the parameter p is the norm parameter for the low-rank regularization term. It adjusts the stringency of the low-rank penalty. As is analyzed in previous section, Schatten p-norm makes a more strict low-rank constraint than trace norm when $0 < p < 1$. The closer p is to 0, the more rigorous low-rank constraint the regularization term $\|M\|_{S_p}^{p}$ imposes. But empirically we don't set p to a too small value since it makes the model contain too many local-minima thus is sensitive to noise and outliers.

The parameter k in Problem (1) is proposed to guarantee the robustness of our model. When p is small, the number of local solutions becomes more thus lead the model to be more sensitive to outliers. Under this condition, a larger k value will render the model more robust to outliers.

Let's take an intuitive example for understanding p and k: Suppose $p \to 0$, then in Problem (1) our regularization term approximates to $\sum_{i=1}^{g} (\mathrm{rank}(WQ_i))^k$. Assume $rank(W) = 10$ and there exist two latent low-rank subspaces that $rank(WQ_1) = 5$ and $rank(WQ_2) = 5$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

If $k = 1$, the real latent low-rank subspaces WQ_1 and WQ_2 give the minimum regularization value as 5 + 5 = 10. However, if we find some non-low-rank subspaces WQ_3 and WQ_4 instead, the regularization value is at most 10 + 10 = 20, which maintains no big difference from 10.

If $k = 2$, the optimal regularization value is $5^2 + 5^2 = 50$, while WQ_3 and WQ_4 cause a value of $10^2 + 10^2 = 200$, which is much larger than 50. In this case, the algorithm will favor the real low-rank subspaces. This is how a larger k value makes our model robust when p is small. According to our pre-experiments, we usually set k value in the range of [2, 3].

Parameter γ is use to balance the role of the low-rank penalty, which can be adjusted to accommodate different cases. γ can be set to any positive value.

When conducting the experiments, we did not spend too much time tuning the parameters. On the contrary, in order to fairly compare all methods, we simply set each parameter to a reasonable value, which is discussed in the next section. While these parameters introduced significant challenges in optimizing our objective, they make our model more flexible and adapt to different situations.

5 Experimental Results

5.1 Experimental Settings

In the classification experiment, we employed hinge loss in Problem (1). We compared with the following methods: Support Vector Machine with ℓ_1-norm loss (L1SVM) as baseline, k-Nearest Neighbors algorithm (KNN), Least Square SVM (LSSVM) [15] and SVM with with ℓ_2-norm loss (L2SVM). To apply SVM model to the multi-class classification problem, we adopted 1-vs-all mechanism. Besides, we compared with one state-of-art method conducting structured multi-task learning via trace norm regularization (TMTL) [9]. In TMTL model, we also used hinge loss to conduct classification. It is notable that TMTL makes a special case of our model (1) with $p = 1$ and $k = 1$.

In our experiments, we exploited the toolbox of LIBSVM [6] to implement both L1SVM and L2SVM. All participating data sets were normalized to the range of [0, 1] and randomly divided using 5-fold cross validation. We excavated the classification result in each fold and recorded the average in these 5 times repetition.

The evaluation of different methods was based on the percentage of correctly classified samples, *i.e.*, classification accuracy. For KNN, we set $k = 1$. For all other methods using the hinge loss, we tuned the C parameter in the range of $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ on training and validation data and recorded the performance on testing data using the best parameter *w.r.t.* each method.

Our model consists of several other parameters such as p, γ, k and δ. In our pre-experiments, we use cross-validation to find a reasonable range for each parameter. We found the performance of our model relatively stable within the reasonable range of parameters (data not shown). Indeed, we can further improve the performance with fine-tuning the parameters. Instead, we simply fix $p = 0.25$, $\gamma = 1$, $k = 3$ and $\delta = 10^{-12}$ in the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

experiments. Unless specified otherwise, we set the number of groups as $g = 2$. The values of these parameters were determined according to the theoretical reasonable range discussed in Section 4 and empirical convention.

5.2 Description of ADNI Data

Data used in the preparation of this paper was obtained from the ADNI database (http://adni.loni.usc.edu). One goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, we refer interested readers to visit www.adni-info.org. We downloaded baseline 1.5 T MRI scans and demographic information for 818 ADNI-1 participants. For each baseline MRI scan, FreeSurfer [14] was employed for brain segmentation and cortical parcellation, and extracted 90 thickness and volume measures, which were pre-adjusted by intracranial volume (ICV) using the regression weights derived from the healthy control (HC) participants. Besides, we performed voxel-based morphometry (VBM) [14] on the MRI data, and extracted mean gray matter (GM) density measures for 90 target regions of interest (ROIs). The time points examined in this study for both imaging markers and diagnostic status included baseline (BL) and Month 36 (M36). All participants with no missing BL/M36 MRI measurements and diagnostic status were involved in this study. All in all, we include 516 sample subjects in our study, including 105 AD samples, and 237 MCI samples and 174 health control (HC) samples. Among the 237 MCI samples, 9 of them become HC in M36, 95 become AD in M36 while the rest 133 remain as MCI along this three-year continuum.

5.3 Performance Comparison on ADNI Cohort

We labeled the ADNI data according to a three-year clinical observation to five different classes, which are are: 1. health control (HC), 2. MCI(baseline)-HC(M36), 3. MCI(baseline)-MCI(M36), 4. MCI(baseline)-AD(M36) and 5. AD. Classification experiments were performed only on the baseline neuroimaging data so as to compare the "forecasting" ability of different methods. Our goal is to classify these different classes using baseline data, *i.e.*, detect MCI stage changes three years before the clinical diagnosis, which will make a contribution to therapeutic intervention of AD in the most effective stage. The comparison results are summarized in Table 1.

From Table 1, we found that our new method performs better than the counterparts in classifying the different classes using merely baseline data. Besides, we get two other interesting observations: 1) SVM methods outperforms KNN on the ADNI data; 2) L1SVM and L2SVM perform equal or better than LSSVM. The reason may go as follows: For KNN, it is a method focused more on the local data structure, while SVM model is meant to effectively find the separating hyperplanes, which is more suitable for classification. The unilateral loss is more interpretable and robust than bilateral loss for classification, thus we notice that L1SVM and L2SVM perform equal or better than LSSVM method. As for our proposed method, we utilized the unilateral hinge loss to be adaptive for classification and also automatically discovered the groupwise structure among different classes, which strengthened the classification performance. To compare our method with TMTL, even

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

though both methods attempted to detect the groupwise structure among different tasks, our model is more general and robust. The use of Schatten p-norm and the power parameter k make our model better approximate the low-rank structure of the latent subspaces thus perform better.

It is also worth mentioning that in this classification, we only use neuroimaging data but not cognitive test information as previous papers do, *e.g.*, [12]. In [12], the classification accuracy is over 70% by adding the cognitive test information to prediction. However, cognitive assessment is a direct diagnostic criterion of MCI and AD [1]. Predicting MCI with cognitive scores is like classifying with label information, which will definitely boost the performance. But using the cognitive test scores as features, the classification is no longer “forecasing” but just a classification of existing information.

Moreover, we present the detected groupwise structure from TMTL and our method on VBM analysis in Fig. 1. It seems that TMTL fails to detect the appropriate group structure among the five classes, but put them all together in one group. On the contrary, our method successfully finds the intrinsic group information among different classes. Fig. 1 shows an interesting phenomenon that no matter what the group number g we set, our model always groups the five classes into two clusters. This illustrates that our model is able to find the intrinsic group structure regardless of g parameter settings. Also, according to the detected structure, we know that even though three different types of MCI patients *i.e.*, class 2, 3 and 4, end up with 3 different stages in month 36, they adopt a similar pattern in the baseline. As a subdivision, MCI-AD shows a potential similarity with AD while the other two types of MCI obtain patterns like HC. Such detected group information may help with the diagnosis of MCI and AD.

5.4 Discussion on Top Ranked Features

In this section, let's take an insight into the results. We use heat maps and brain maps to intuitively indicate the degree of influence imposed by each imaging feature, such that important features in classification can be determined.

Shown in Fig. 3 and Fig. 2 are the heat maps of sorted neuroimaging feature weights and corresponding brain maps. The figures demonstrate the capture of a small set of features that are predominant for classification. Among the selected features, we found LHippoCampus and LPostCentral on the top, whose impact on AD have already been proved in the previous papers [5,20]. These identified imaging disease associations warrant further investigation in independent cohorts. If replicated, these findings can potentially contribute to biomarker discovery for diagnosis and drug design.

5.5 Experiments of Convergence Analysis

In this subsection, we empirically analyze the convergence of our algorithm with respect to the two parameters p and k in Eq. (1). We apply our method to the entire data with two different p values (*i.e.*, 0.25 and 0.75) and two different k values (*i.e.*, 2 and 3), then record the objective value of our model in each iteration.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We use the results on FreeSurfer as an example. The convergence plots are shown in Fig. 4. We notice that the number of iterations need before convergence is fairly stable with respect to the settings of p and k parameters. No matter what p and k values are, our model usually converges within 15 iterations.

6 Conclusions

In this paper, we proposed a novel low-rank structured classification model to predict MCI conversion using neuroimaging data in the baseline time. Our model simultaneously uncovered the interrelation structures existing in different classes and employed such structure to enhance the classification model. Moreover, we utilized Schatten p-norm to extract the common low-rank subspace shared by different patient classes. We conducted experiments on ADNI cohort. Empirical results validated the effectiveness of our model by demonstrating improved classification performance compared with competing methods. In addition, the top ranked biomarkers in our method were verified by previous literature, which indicated the potential contribution of our model to AD diagnosis and drug design.

References

1. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, et al. The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimer's & Dementia. 2011; 7(3):270–279.
2. Bentler P, Lee SY. Matrix derivatives with chain rule and rules for simple, hadamard, and kronecker products. Journal of Mathematical Psychology. 1978; 17(3):255–262.
3. Bezdek JC, Hathaway RJ. Convergence of alternating optimization. Neural, Parallel & Scientific Computations. 2003; 11(4):351–368.
4. Boyd, S., Vandenberghe, L. Convex optimization. Cambridge university press; 2004.
5. Cacabelos R, Yamatodani A, Niigawa H, Hariguchi S, Tada K, Nishimura T, Wada H, Brandeis L, Pearson J. Brain histamine in alzheimer's disease. Methods and findings in experimental and clinical pharmacology. 1989; 11(5):353–360. [PubMed: 2755282]
6. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2:27:1–27:27. software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
7. Devanand D, Pradhaban G, Liu X, Khandji A, De Santi S, Segal S, Rusinek H, Pelton G, Honig L, Mayeux R, et al. Hippocampal and entorhinal atrophy in mild cognitive impairment prediction of alzheimer disease. Neurology. 2007; 68(11):828–836. [PubMed: 17353470]
8. Hua X, Leow AD, Parikshak N, Lee S, Chiang MC, Toga AW, Jack CR, Weiner MW, Thompson PM, Initiative ADN, et al. Tensor-based morphometry as a neuroimaging biomarker for alzheimer's disease: an mri study of 676 ad, mci, and normal subjects. Neuroimage. 2008; 43(3):458–469. [PubMed: 18691658]
9. Kang, Z., Grauman, K., Sha, F. Learning with whom to share in multi-task feature learning. Proceedings of the 28th International Conference on Machine Learning (ICML-11); 2011. p. 521-528.
10. Kittaneh F. Inequalities for the schatten p-norm. Glasgow Mathematical Journal. 1985; 26(02): 141–143.
11. 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. 2009; 44(4):1415–1422. [PubMed: 19027862]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

12. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, Initiative ADN, et al. Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. NeuroImage. 2015; 104:398–412. [PubMed: 25312773]
13. Petersen R, Stevens J, Ganguli M, Tangalos E, Cummings J, DeKosky S. Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review) report of the quality standards subcommittee of the american academy of neurology. Neurology. 2001; 56(9):1133–1142. [PubMed: 11342677]
14. Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud T, Pankratz N, Moore JH, Sloan CD, et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in mci and ad: A study of the adni cohort. Neuroimage. 2010; 53(3):1051–1063. [PubMed: 20100581]
15. Suykens, JA., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., Suykens, J., Van Gestel, T. Least squares support vector machines. Vol. 4. World Scientific; 2002.
16. Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, AJ., Shen, L. ADNI: Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. IEEE Conference on Computer Vision; 2011. p. 557-562.
17. Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, AJ., Shen, L. ADNI: Joint classification and regression for identifying ad-sensitive and cognition-relevant imaging biomarkers. The 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI); 2011. p. 115-123.
18. Wang H, Nie F, Huang H, Risacher SL, Saykin AJ, Shen L, ADNI. Identifying disease sensitive and quantitative trait relevant biomarkers from multi-dimensional heterogeneous imaging genetics data via sparse multi-modal multi-task learning. Bioinformatics. 2012; 28(12):i127–i136. [PubMed: 22689752]
19. Wenk GL, et al. Neuropathologic changes in alzheimer's disease. Journal of Clinical Psychiatry. 2003; 64:7–10.
20. West MJ, Coleman PD, Flood DG, Troncoso JC. Differences in the pattern of hippocampal neuronal loss in normal ageing and alzheimer's disease. The Lancet. 1994; 344(8925):769–772.

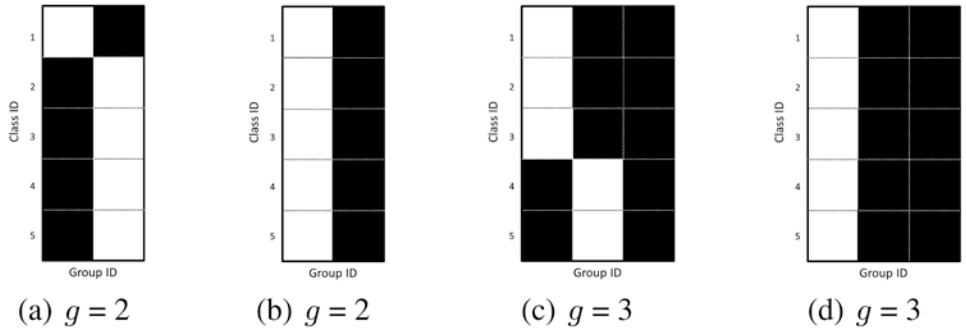


Fig. 1.

Illustration of the detected group structure among different classes in our method ((a) and (c)) and TMTL ((b) and (d)) in the VBM analysis. We set the number of groups to be 2 and 3, respectively. White blocks denote that a class belongs to a certain group while black block denote otherwise. The five classes are: 1. health control (HC), 2. MCI(baseline)-HC(M36), 3. MCI(baseline)-MCI(M36), 4. MCI(baseline)-AD(M36) and 5. AD.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

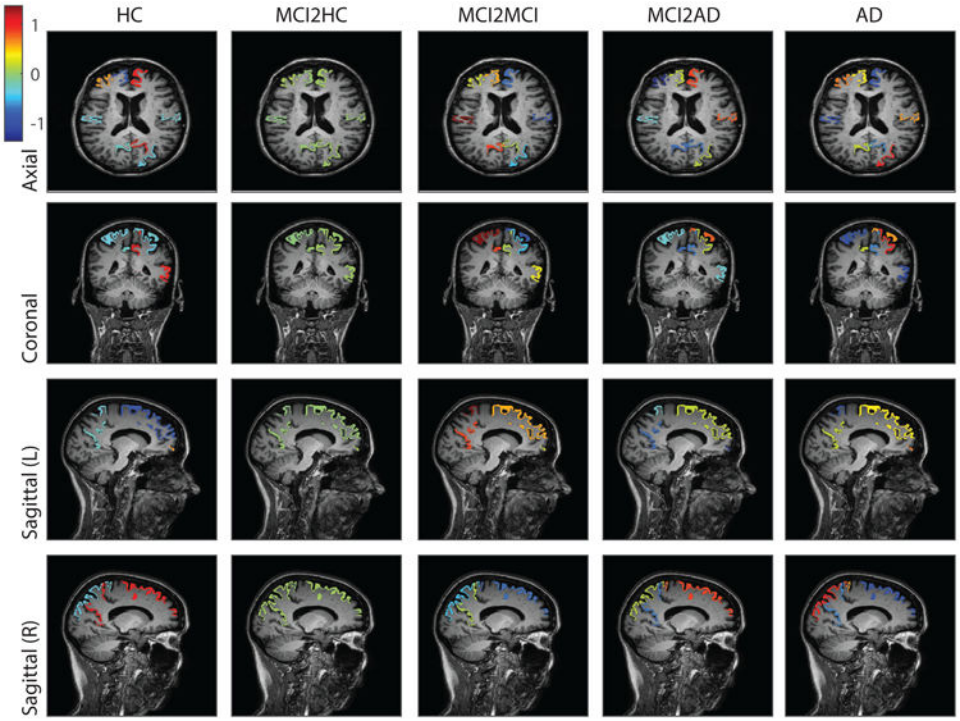


Fig. 2.
Neuroimaging features mapped on the brain for the FreeSurfer analysis.

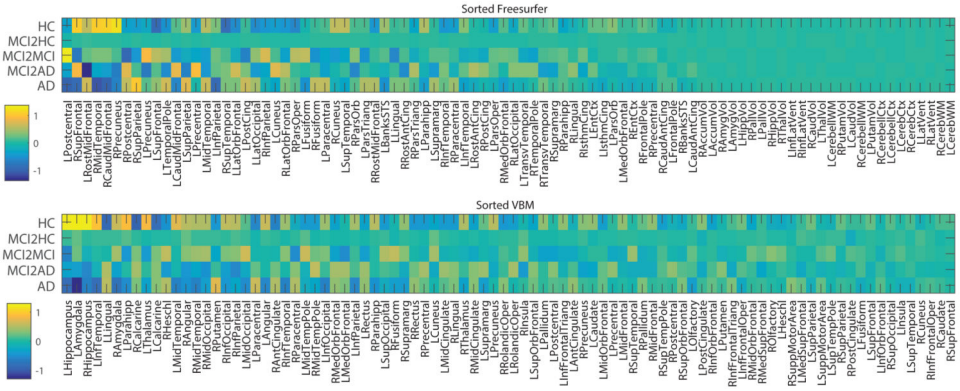
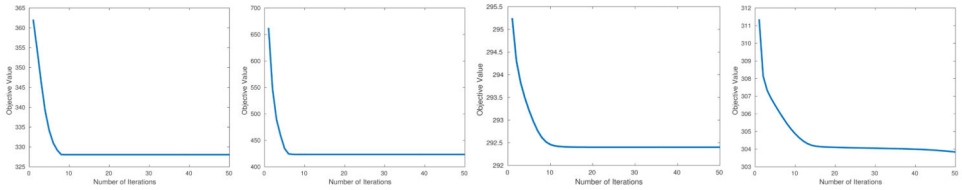


Fig. 3.

Heat maps of sorted neuroimaging feature weights in our method in descending order from left to right. The feature weight matrix is learned on the entire data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



(a) $p = 0.25, k = 2$ (b) $p = 0.25, k = 3$ (c) $p = 0.75, k = 2$ (d) $p = 0.75, k = 3$

Fig. 4.

Objective function value of Eq. (1) with different k and p parameters in each iteration on the FreeSurfer data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Classification accuracy (%) comparison using "FreeSurfer" and "VBM" data

	KNN	L1SVM	LSSVM	L2SVM	TMTL	OURS
VBM	33.21±8.00	43.66±1.73	44.14±5.64	44.15±5.61	42.55±3.37	**44.79±4.46**
FreeSurfer	37.77±3.74	45.78±3.71	44.23±3.60	45.02±4.97	44.83±4.79	**48.48±3.25**