Identifying Genetic Associations with MRI-derived Measures via Tree-Guided Sparse Learning

Xiaoke Hao[1], Jintai Yu[2], Daoqiang Zhang[1,*]

[1] Department of Computer Science and Engineering,
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[2] Department of Neurology, Qingdao Municipal Hospital,
Nanjing Medical University, Nanjing, 210029, China
dqzhang@nuaa.edu.cn

Abstract. In recent imaging genetic studies, much work has been focused on regression analysis that treats large-scale single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) as association variables. To deal with the weak detection and high-throughput data problem, feature selection methods such as the least absolute shrinkage and selection operator (Lasso) are often used for selecting the most relevant SNPs associated with QTs. However, one problem of Lasso as well as many other feature selection methods for imaging genetics is that some useful prior information, i.e., the hierarchical structure among SNPs throughout the whole genome, are rarely used for designing more powerful model. In this paper, we propose to identify the associations between candidate genetic features (i.e., SNPs) and magnetic resonance imaging (MRI)-derived measures using a tree-guided sparse learning (TGSL) method. The advantage of our method is that it explicitly models the priori hierarchical grouping structure among the SNPs in the objective function for feature selection. Specifically, two kinds of hierarchical structures, i.e., group by gene and group by linkage disequilibrium (LD) clusters, are imposed as a tree-guided regularization term in our sparse learning model. Experimental results on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database show that our method not only achieves better predictions on the two MRI measures (i.e., left and right hippocampal formation), but also identifies the informative SNPs to guide the disease-induced interpretation compared with other reference methods.

1 Introduction

Imaging genetics is the study of how individual genetic differences lead to differences in brain wiring, structure and intellectual function. Compared to case-control status, quantitative brain imaging measures are considered to be intermediate or endophenotypes that are closer to the underlying biological mechanisms of the disease.

Genome-wide association studies (GWAS) are increasingly being used to identify the associations between the high-throughput single nucleotide polymorphisms (SNPs) and

* Corresponding author

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 757–764, 2014.

the quantitative traits (QTs) of imaging data [1]. To our knowledge, most existing GWAS-based methods focus on univariate analysis, which ignores the underlying interacting relationship among SNPs and thus easily leads to a weak detection of associations. To address that problem, feature selection methods such as the least absolute shrinkage and selection operator (Lasso) [2] have been proposed to identify a subset of features (i.e., SNPs) for subsequent association analysis [3]. In Lasso, an L1-regularized term is used to enforce the 'sparsity' on the individual features, without considering the structural information among SNPs that exist throughout the whole genome. Recently, based on group Lasso method that extends Lasso by imposing the 'group sparsity' with L1/L2 norm-based regularization [4], an excellent method has been proposed in [5] to consider the group structure among SNPs. However, in that method [5], the hierarchical structure among SNPs that are different from flat group structure, are still not used for designing more powerful model.

On the other hand, in machine learning community sparse learning methods with tree-structured regularizations have been proposed to consider the underlying multi-level tree (i.e., hierarchical) structures among the inputs or outputs [6, 7]. The hierarchical structured sparsity has been implemented with hierarchical agglomerative clustering technique for multi-scale mining of functional magnetic resonance imaging (MRI) data [8]. Recently, those tree structure-based method have been successfully used for neuroimaging-based brain disease classification [9]. Motivated by the above works, in this paper, we propose to identify the associations between SNPs and MRI-derived measures using a tree-guided sparse learning (TGSL) method, which explicitly models the priori hierarchical tree structure among the SNPs in the objective function for feature selection. Here, the hierarchical tree structure is constructed based on the following priori knowledge, i.e., each tree node is for one feature group and different tree heights represent different levels of groups. Specifically, some SNPs are naturally connected via different pathways, and multiple SNPs located in one gene often jointly express certain genetic functionalities. Also, another genetic biology phenomenon, i.e., linkage disequilibrium (LD) [10], describes the non-random distributions between alleles at different loci. Inspired by the above prior knowledge, the spatial gene and LD relationships among SNPs can be encoded into the tree regularization to guide the selection of relevant features for subsequent prediction.

We apply the proposed TGSL method to the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort for predicting well known disease-related hippocampal MRI-derived measures with pre-selected candidate SNPs. The empirical results show that our method not only yield improved prediction performance, but also detect a compact set of SNP predictors relevant to the imaging phenotypes.

2 Method

2.1 L1-Regularized Sparse Coding (Lasso)

Assume we have M training subjects, with each represented by a N-dimensional feature vector (i.e. SNPs) and a response value (i.e. MRI-derived measure). Let X be a $M \times N$ feature matrix with the m-th row $x^m = (x_1^m, \ldots, x_n^m, \ldots, x_N^m) \in R^N$ denoting the

m-th subject's feature vector, and *y* be the corresponding MRI-derived measures of *M* subjects. A linear regression model can be formulated as follows:

$$y = X\alpha + \varepsilon \tag{1}$$

where α is a vector of coefficients assigned to the respective features, and ε is an error term. To encourage the 'sparsity' among features, in the Lasso method a L1-norm regularization is imposed on the coefficients as follows [2]:

$$\alpha = \text{argmin}_{\alpha} ||y - X\alpha||^2 + \lambda||\alpha||_1 \tag{2}$$

where λ is a regularization parameter that controls the sparsity in the solution. The non-zero elements of α indicate that the corresponding input features are relevant to the regression outputs.

2.2 Tree-Guided Sparse Learning

It's known that in feature selection it is promising to consider the grouping structure among features instead of treating them as individual units. In order to address the group-wise association among the features, sparsity can be enforced at the group level by a L1/L2 regularization, where the L2-norm is applied for the input features within the same group, while the L1-norm penalty is applied over the groups of input features [4], as have done in [6] for imaging genetic study. However, it is not enough to capture the relationship among SNPs via simple group-wise association as a flat manner, because there exist more complex structures (e.g., hierarchical tree structure) among SNPs. In this section, we introduce a tree-guided sparse learning (TGSL) method [6] for identifying the association between SNPs and imaging measures. In TGSL, a tree structure is used to represent the hierarchical spatial relationship among SNPs, with leaf nodes denoting SNPs and internal nodes denoting the groups of SNPs. Such hierarchical tree structures are shown in Fig.1.

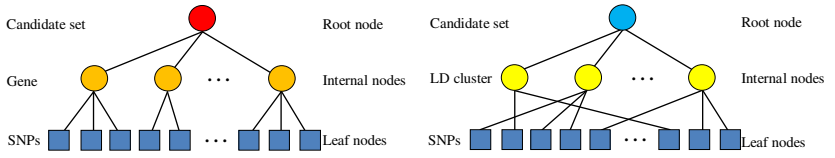

Fig. 1. Illustration on the tree-structured hierarchical relationship among SNPs: **(a)** group by gene, **(b)** group by linkage disequilibrium (LD) cluster

As can be seen from Fig.1, two kinds of methods are used for the tree construction, i.e., (a) gene-based method, and (b) linkage disequilibrium (LD) cluster-based method. Specifically, in the gene-based method, some SNPs are naturally connected via different pathways and multiple SNPs located on one gene often jointly express certain genetics functionalities. So, SNPs are naturally divided into groups upon their

belonging genes. On the other hand, in LD cluster-based method, we generate the tree structure through estimating non-random association of alleles at different loci, and the relationship among SNPs in terms of genetic linkage are hence established [5]. Both gene and LD cluster groups could be encoded into the tree structure as shown in Fig.1.

Assume that a hierarchical tree T has d depth levels, and there are n_i nodes organized as $T_i = \{G_1^i, \ldots, G_j^i, \ldots, G_{n_i}^i\}$ in the i-th level ($0 \le i \le d$). Different depth levels indicate the variant scales of feature groups. The index sets of the nodes at the same level have no overlapping while the index sets of a child node is a subset of its parent node. The TGSL method [6] can be formulated as:

$$\alpha = \arg\min_\alpha ||y - X\alpha||^2 + \lambda \sum_{i=0}^{d} \sum_{j=1}^{n_i} w_j^i ||\alpha_{G_j^i}||_2 \quad (3)$$

where $\alpha_{G_j^i}$ is the set of coefficients assigned to the features within node G_j^i, w_j^i is a predefined weight for node G_j^i and is usually set to be the same for each group at the same level, and the number of depth levels d is set to 3 in our experiments. A regularization predefined by the tree structure can be imposed on the sparse learning optimization problem to encourage a joint selection of structured relevant SNPs.

2.3 Imaging Phenotype Prediction

We consider each SNP as a feature and each QT as a response variable, and formulate a regression model including multiple features (SNPs) and single response (MRI-derived measures). Our goal is to reveal the relationship between these genetic features and imaging phenotypes. Fig.2 shows the flowchart of the proposed method. First, to capture the hierarchical relationship of the SNPs in our candidate set, we construct a tree structure by naturally agglomerating related SNPs into gene groups or LD cluster groups, by using a hierarchical clustering technique. Then, the constructed tree structure is imposed on the regularization of tree-guided sparse learning (TGSL) model to select the relevant features. Finally, support vector regression (SVR) is used to predict the MRI-derived measures using the selected SNPs features.

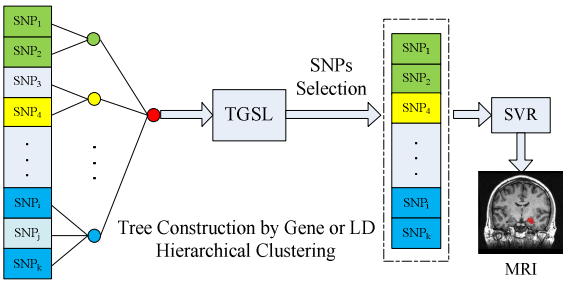

Fig. 2. The flowchart of the proposed method

3 Experiments

In this section, we evaluate the effectiveness of the proposed method on the ADNI database (http://adni.loni.usc.edu/), where candidate SNPs are examined and selected to predict the response of the MR imaging phenotypes. In our experiments, baseline 1.5 Tesla MRI scans and SNP data are included. This yields a total of 734 subjects, including 173 AD patients, 360 MCI patients and 210 healthy controls.

3.1 Imaging Data and Pre-processing

Standard image pre-processing is performed for all MR images. With atlas warping, we can partition each subject image into 93 regions of interests (ROIs). For each of the 93 ROIs, we compute the GM tissue volume from the subject's MR image. A detailed description on acquiring MRI data from ADNI as used in this paper can be found in [11]. For identifying QTs, two well-known MRI phenotypes, i.e., left and right hippocampal formation, are used in our experiments.

3.2 SNP Genotyping and Pre-processing

Genome-wide genotyping data are available for the full set of ADNI subjects. The 620901 SNPs data, used in this study, are genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) [12]. We ignore the SNPs defining the APOE e4 variant which are not included in the original genotyping chip. Only SNPs, belonging to the top 11 AD candidate genes listed on reference [13,14] and the Alz-Gene database (www.alzgene.org) as of Jan 10, 2014, are selected after the standard quality control (QC) and imputation steps. Additionally, we apply filter rules to the genotype data to remove rare SNPs (minor allele frequency < 0.05), violations of Hardy-Weinberg Equilibrium (HWE $p < 10^{-6}$), and poor call rate (< 90%). Data are further "phased" to impute any missing individual genotypes after filtering using the MaCH program [15]. After that, candidate SNPs on the genes listed at 1000Genomes website (http://browser.1000genomes.org/) are used to select a subset of SNPs.

As we introduced before, we form two kinds of tree structures of SNPs: 1) SNPs annotated within the same gene (yielding 107 SNPs from 11 genes); 2) SNPs within the LD hierarchical cluster. For group by gene, since all SNPs had been divided into different genes naturally we use this natural groups to construct tree. For group by LD, We first compute the correlation (i.e., r^2) between paired SNPs by Plink tool, and then perform agglomerative hierarchical clustering based on pairwise distances (i.e., $1-r^2$) among SNPs to get the hierarchical tree.

3.3 Experimental Settings

To reduce the computational cost, we constrain that the group weights are set to be the same for each group at the same level and the values are tuned by nested cross-validation. As for Lambda, we determine its values corresponding to the number of selected SNPs from 10 to 100 with approximate step of 10. In our experiment, we use Lasso, TGSL-gene (denoting TGSL with gene grouped tree structure) and TGSL-LD (denoting TGSL with LD cluster grouped tree structure) methods to select a subset of features (i.e., SNPs) to predict the regression responses for the test data. The performance of each trial is

assessed with root mean squared error (RMSE), a widely used criterion in regression analysis. Average RMSE result is calculated based on 10-fold cross validation.

3.4 Results

We compare our proposed methods (including TGSL-gene and TGSL-LD) against standard Lasso and group Lasso (including GroupLasso-gene and GroupLasso-LD) feature selection method. For testing the regression performance with respect to different level of selected features in all five methods, we adjust the regularization parameter to control the sparsity. Fig.3 reports the RMSE for regression on left hippocampal formation and right hippocampal formation by adopting a polynomial model to fit all the data obtained by different regularization parameters. As can be seen from Fig. 3, the proposed TGSL-gene and TGSL-LD methods outperform the Lasso and group Lasso methods. TGSL methods can get the best RMSEs at top 10 SNPs selection in the quantitative assessment.

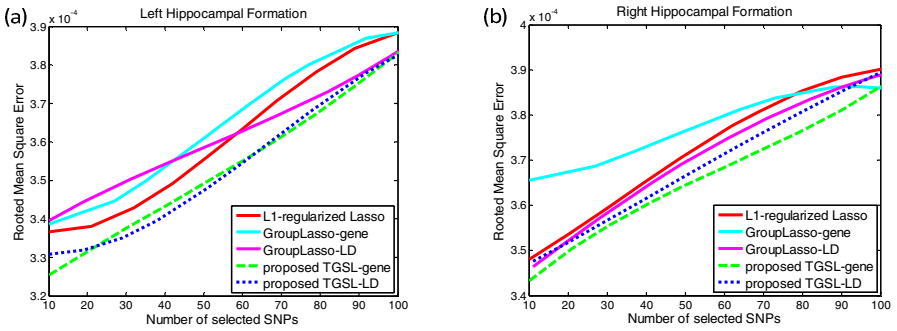

Fig. 3. Comparison of RMSE with respect to different number of selected SNPs by L1-regularized Lasso, GroupLasso-gene, GroupLasso-LD, the proposed TGSL-gene and TGSL-LD in prediction on (a) left hippocampal formation, (b) right hippocampal formation.

The regression coefficients for the top 10 selected SNPs by each approach on the MRI-derived measures (including left and right hippocampal formation volume) are plotted in Fig.4. The group Lasso methods are trend to select the entire gene or LD clusters which pick up most SNPs with an excessive constraint, e.g., the only two SNPs are selected with the GroupLasso-gene method for the best performances. Thus, the group Lasso methods are much more sensitive to the definition of their groups.

As illustrated in the Fig.4, the PICALM-rs11234532are significantly associated with the predictions on left and right hippocampal formation with the proposed methods in the experiment. As can be seen from Fig. 4, TGSL prefer selecting more SNPs on PICLAM. As PICLAM is a new Aβ toxicity modifier gene, the more SNPs on PICLAM have been detected by the proposed TGSL method are significantly associated with risk of late-onset Alzheimer disease (LOAD) [16]. The rs10501608 in PICALM is also associated with LOAD risk in the genome-wide SNP linkage and association studies [17]. Among the top 10 SNPs selected by TGSL-gene, PICALM-

rs11234495, PICALM-rs713346 and PICALM-rs10792820 are detected on left hippocampal formation, while PICALM-rs2077815, PICALM-rs11234495 and PICALM-rs527162 show the strongest association with the right hippocampal formation. It's worth noting that these SNPs are also reported in other related heritable neurodevelopmental disorders [18].

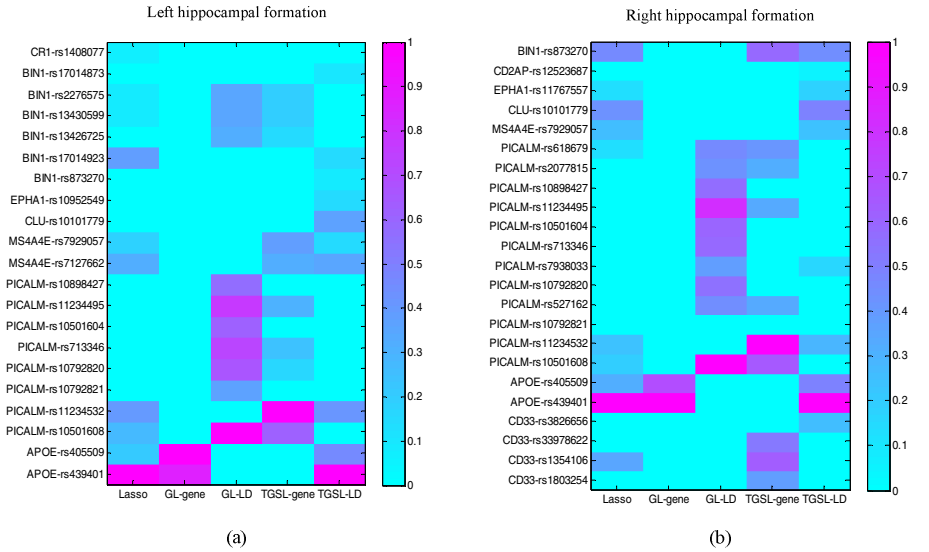

Fig. 4. Regression coefficients for the top 10 selected SNPs on (a) left hippocampal formation and (b) right hippocampal formation prediction by L1-regularized Lasso, GroupLasso-gene (GL-gene for short), GroupLasso-LD (GL-LD for short), TGSL-gene and TGSL-LD.

4 Conclusion

In this paper, we investigate the potential of exploiting tree-guided sparse learning (TGSL) method for identifying the associations between SNPs and MRI-derived measures, given hierarchical tree structure among SNPs. Specifically, two kinds of methods, i.e., TGSL-gene and TGSL-LD are developed based on grouping by the gene and linkage disequilibrium (LD) clusters, respectively. Experimental results on the ADNI database show that our method not only achieves better prediction performances on the MRI-derived hippocampal formation volume measures, but also identifies informative SNPs biomarkers to guide the disease interpretation compared with other reference methods.

Acknowledgments. This work was supported in part by the Jiangsu Natural Science Foundation for Distinguished Young Scholar (No. BK20130034), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20123218110009), the NUAA Fundamental Research Funds (No. NE2013105).

References

1. Shen, L., 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 53, 1051–1063 (2010)
2. Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society Series B-Statistical Methodology 73, 273–282 (2011)
3. Kohannim, O., et al.: Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. Front. Neurosci. 6, 115 (2012)
4. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B-Statistical Methodology 68, 49–67 (2006)
5. Wang, H., et al.: Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics 28, 229–237 (2012)
6. Liu, J., Ye, J.: Moreau-Yosida regularization for grouped tree structure learning. In: Lafferty, J., Williams, C., Shawe-Taylorand, J., Zemel, R., Culotta, A. (eds.) NIPS 2010, vol. 23, pp. 1459–1467. Curran Associates, Inc. (2010)
7. Kim, S., Xing, E.P.: Tree-Guided Group Lasso for Multi-Response Regression with Structured Sparsity, with an Application to Eqtl Mapping. Annals of Applied Statistics 6, 1095–1117 (2012)
8. Jenatton, R., et al.: Multiscale mining of fMRI data with hierarchical structured sparsity. Siam J. Imaging Sci. 5, 835–856 (2012)
9. Liu, M., Zhang, D., Yap, P.-T., Shen, D.: Tree-guided sparse coding for brain disease classification. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 239–247. Springer, Heidelberg (2012)
10. Barrett, J., et al.: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263–265 (2005)
11. Zhang, D., et al.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856–867 (2011)
12. Saykin, A., et al.: Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. Alzheimers & Dementia 6, 265–273 (2010)
13. Shi, H., et al.: Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS). Neurobiology of Aging 33(8), 1849.e5–1849.e18 (2012)
14. Bertram, L., et al.: Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nature Genetics 39, 17–23 (2007)
15. Li, Y., et al.: MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. Genetic Epidemiology 34, 816–834 (2010)
16. Rosenthal, S., et al.: Beta-amyloid toxicity modifier genes and the risk of Alzheimer's disease. Am J. Neurodegener. Dis. 1, 191–198 (2012)
17. Cummings, A., et al.: Genome-wide association and linkage study in the Amish detects a novel candidate late-onset Alzheimer disease gene. Ann. Hum. Genet. 76, 342–351 (2012)
18. Xia, K., et al.: Common genetic variants on 1p13.2 associate with risk of autism. Mol. Psychiatry (November 5, 2013)