Feature Selection via Sparse Regression for Classification of Functional Brain Networks

Yilun Wang[1], Guorong Wu[2], Zhiliang Long[2], Jingwei Sheng[2], Jiang Zhang[3], and Huafu Chen[2,]*

[1] School of Mathematical Sciences,
University of Electronic Science and Technology of China, Sichuan, China
yilun.wang@gmail.com

[2] Key laboratory for Neuroinformation of Ministry of Education,
School of Life Science and Technology,
University of Electronic Science and Technology of China, Chengdu, P.R. China
chenhf@uestc.edu.cn

[3] Information Research Institute, Southwest Jiaotong University,
Chengdu 610031, China

Abstract. Despite the ongoing progress to chart the differences between the healthy controls and patients at the group level, the pattern classification of functional brain networks across individuals is still a challenging task. The difficulties include the very high dimensional feature space and very small sample size, as well as the probably high noise level. In this paper, we apply the stable sparse regression to pick the very few most discriminant features (edges) for the following classification. We considered different noise to signal ratios and sparsity controlling parameters and numerical experiments based on simulated data demonstrate the much better classification performance via the feature selection based on the sparse regression.

Keywords: sparse regression, feature selection, stability selection, classification.

1 Introduction

The human brain is among the most complex network systems in the world, considering that it comprises about one hundred billion neurons, with thousands of trillions of connections between them. The anatomical and physiological studies in past few decades provided a significant body of evidence for the important role of structural connectivity in shaping physiological responses. Meanwhile, functional connections that describe statistical dependencies are derived from observations of neural time series, reflecting functional segregation and localization of function in neuroscience. That is to say, the human brain can be considering as a large-scale network, with nodes being distinct brain regions and edges representing functional or structural connectivity among them [1, 2].

* Corresponding author.

C. Sun et al. (Eds.): IScIDE 2013, LNCS 8261, pp. 554–560, 2013.

In this paper, we are focusing on pattern recognition and classification based on the brain network. Most existing research on brain networks simply focuses on describing group differences between subject classes (knowing the label of each subject) and cannot classify or predict the brain behavior across individuals, due to the relatively small number of subjects, very high dimensional feature space (consisting of the network edges) and probably high level noise, leading to the over-fitting training data and curse of dimensionality problem. Therefore, in this paper, we are focusing on selecting a small number of most discriminative features, to significantly reduce the dimension of the feature space and correspondingly enhance the classification performance. It is quite reasonable to perform feature (edge) selection, because usually only a small proportion of the pathways in the brain might be responsible for the dysfunction or certain task of the brain network.

Recently, sparse modeling, as a rapidly developing area at the intersection of statistics, machine-learning and signal processing [3, 4], can find out a small number of the most relevant variables in a high-dimensional feature space and therefore is most appealing for practical feature selection. It has been applied to many problems, including the voxel selection to localize brain activation patterns corresponding to different stimulus classes or brain states [5–7].

In this paper, we study the application of sparse regression to the feature (edge) selection in brain network with an aim to identify a small proportion of the discriminative functional pathways and brain regions. While there have existed some related work [8, 9], the deep study of sparse regression for feature selection on the brain network is still very limited. For example, the effects of the different signal and noise ratios and the discussion of the different sparsity levels have not been considered. In this study, we will deepen the existing study in the above two aspects. Notice that while our study is not limited to a specific type of network, we are mainly focusing on the network based on the functional MRI, and will mainly use the simulated data, since our main goal is to evaluate the methodology of the stable sparse regression.

2 Methods

2.1 Subjects

In total, the fMRI data of 100 subjects is generated and equally divided into 2 groups, i.e. Group 1 and Group 2, respectively. Group 1 and Group 2 differs with each other mainly in terms of the strength of functional connectivity between certain regions and we will explain it in more details later. As we know, the simulation data is usually designed to facilitate the deep understanding and testing of a variety of analytic or computational methods before they can applied to the real data.

We adopt a data generation model that is consistent with the spatiotemporal separability assumptions of independent component analysis (ICA), that is, data can be expressed as the product of time courses (TCs) and spatial maps (SMs). For each subject, the spatial map is the same, because we are considering the connectivity between same brain regions. Here we are considered 256 brain regions, and the 256 regions are further divided into two categories, i.e. 87 active regions and 169 non-active regions. They differ in terms of the definition of the corresponding time courses. For each region, we define its average time course with T time points in length. Its construction is under the assumptions that component activations result from underlying neural events as well as noise. In this simulation, each time course has $T = 160$ time points. For the active regions, the time course is divided into several blocks. The signal value of each task block is set to be a positive value between 0.9 to 1, while the signal value for the resting blocks is set to be 0. As the non-active regions, the time course is defined using random time series,defined by normal distribution with mean being 1 and the standard deviation being 1. In order to test the robustness of the classifiers, we add different levels of Gaussian noise.

The functional connectivity is established by calculating the covariances of the time courses between different regions. We have 256 regions and correspondingly the generated network of each subjects has $256 \times 255/2 = 32640$ edges. Groups 1 differs with Groups 2 in terms of the functional connectivity strength of 6 edges.

2.2 Sparse Logistic Regression

In this paper, we are considering the functional connectivity of each subject and correspondingly the feature vector consists of all the edges. The number of edges is typically very big in practice and it is 32640 in our simulated data. That is to say, each subject is defined in a 32640-dimensional feature space. The high dimension of the feature space often bring difficulty for the following classifiers no matter in terms of computational cost or classification accuracy. Therefore we want to perform dimensional reduction by selecting the small number of most discriminative edges, which are used to form a much lower dimensional new feature space.

We use sparse Sparse Logistic Regression based supervised learning to perform the feature selection. The basic formulation of the sparse logistic regression is

$$\min_x \|Ax - y\|_2^2 + \lambda \|x\|_1, \tag{1}$$

where A is the training data matrix, where each row is a 32640 dimensional vector representing one subject, and each column represents a feature; x is the unknown regression coefficients; y is the known label vector for the training data, with 1 representing the subject in Groups 1 and -1 for the Groups 2. The ℓ_1 norm $\|x\|_1$ is the sparsity regularization term, and λ is the regularization parameter that controls the degree of sparsity. A larger λ leads to x with more zeros, i.e. a more sparse coefficient vector. There have existed many different solvers for the above sparse logistic regression model and in the paper we use the solver "SLEP: A Sparse Learning Package " developed by the Arizona State University [10]. By solving the sparse regression problem, we obtain the regression coefficients x whose absolute value indicates the contribution of the corresponding edges to discriminating these two groups.

Randomization for Stability Selection. Due to the presence of the correlation in the training data matrix A, we adopt the randomized sparse regression called stability selection [13] to improve the performance of the sparse feature selection method. The randomized sparse regression is to repeat solving many similar sparse regression problems as (1). Each problem is generated by randomly perturbing the data matrix A via taking only a fraction of the training samples and randomly scaling each feature, in our case each edge. By counting how often each edge is selected across the repetitions, each edge can be assigned a score. Higher scores denote variables likely to belong to the set of the true discriminative edges.

2.3 Support Vector Machine

There have existed many classifiers and in this paper we take the widely used Support Vector Machine (SVM) as an example to demonstrate that our feature selection can help improve the classification accuracy of SVM [11].

Support Vector Machine (SVM) is a specific type of supervised machine learning method that aims to classify data points by maximising the margin between classes in a high-dimensional space and it adopts the ℓ_2-norm regularization to avoid over-fitting. This optimization problem belongs to quadratic programming and can be efficiently solved by many specific solvers such as sequential minimal optimization. Like most of classification methods, SVM involves the training stage and the testing stage. The goal of SVM is to produce a model based on the training data to predicts the target values of the testing data. The traditional SVM performs linear classification, but non-linear classification can also be performed by incorporating the so called the kernel trick, which implicitly maps the inputs into high-dimensional feature spaces. Due to its outstanding practical performance and solid theory guarantee based on the statistical learning, SVM becomes one of the most popular classifiers and therefore in this paper we use it to demonstrate the performance of our feature selection scheme based on sparse optimization.

Consider a training data set $D = \{(x_i, y_i), i = 1, \ldots, n\}$ where $x_i \in R^d$ are data points and y_i are labels. The problem of learning a linear classifier, $y = sign(\omega^T x + b)$, where $y = \{1, -1\}$ or a linear function $y = w^T + b$ where y is a scalar can be understood as estimating $\{\omega, b\}$ from D. Over the years Support Vector Machines(SVMs) have emerged as powerful tools for estimating such functions.

To develop notation we briefly discuss the problem of training linear classifiers. The SVM formulation for linearly separable datasets is given by

3 Numerical Results

In this paper, the classification accuracy is measured by the commonly used quantities, such as generalization rate (GR), sensitivity (SS) and specificity (SC). Here The proportion of all subjects that were correctly predicted is evaluated by the GR; SS is defined as the proportion of correctly predicted Group 1 subjects, while SC represents the proportion of correctly predicted Group 2 subjects. Their formulations are shown below:

$$\begin{aligned}\mathtt{GR} &= (\mathtt{TP}+\mathtt{TN})/(\mathtt{TP}+\mathtt{FN}+\mathtt{TN}+\mathtt{FP})\\ \mathtt{SS} &= \mathtt{TP}/(\mathtt{TP}+\mathtt{FN})\\ \mathtt{SC} &= \mathtt{TN}/(\mathtt{TN}+\mathtt{FP}),\end{aligned}$$

where TP is the number of the Group 1 subjects correctly predicted; FN is the number of the Group 1 subjects classified as in Group 2; TN is the number of the Group 2 correctly predicted; FP is the number of the Group 2 subjects classified as in Group 1 [12]. In this paper, we are using the leave-one-out cross-validation, which use a single subject as the test data and all the remaining as the training data. Each of the 100 subjects is chosen as the test data in turn without missing or repetition, and finally we calculate the value of TP, FN TN, FP, where $\mathtt{TP}+\mathtt{FN}+\mathtt{TN}+\mathtt{FP}=100$.

We use the Support Vector Machine (SVM) as the classifier, which is provided by the toolbox of MATLAB 2012b and the default parameters are used. We first do not perform the feature selection and use all the 32640 edges as the input of SVM, run SVM and record the classification results. Then we first perform edge selection by sparse regression, and then only use the information of the very few number of selected edges as the input of SVM, run SVM, and record the classification result. We compare these two classification methods and demonstrate the significant role of selection of discriminative edges for the improvement of classification performance, in cases of adding different levels of noises.

The performance of the classifier was estimated using leave-1-out validation test with an 100 times repetition. We carried out a simulation study to assess (i) the classification performance under various Noise to Signal Ratio (NSR, for short), which is the reciprocal of the Signal to Noise Ratio (SNR, for short); (ii) the effect of different choices of the penalty parameter λ on our ability to detect the most discriminative interaction. The classification results are summarized in Table 1, Table 2, Table 3 and Table 4. "W/" represent the classification results of SVM together with the feature selection via sparse regression while the "W/O" represent the classification results of SVM without the feature selection via sparse regression. As for the regularization parameter λ, Table 1 and Table 2 are for $\lambda = 1.5$ while Table 3 and Table 4 are for $\lambda = 0.015$. As for the number of selected features (edges, here), Table 1 and Table 3 are for 6 selected most discriminative edges while Table 2 and Table 4 are for 12 selected edges. Here we note that in practice, ones might not know the exact number of the most discriminative edges as we did for simulation data. However, in many simulations ones have a rough estimation based on their experiences and performing feature (edge) selection on this number is still helpful. We will see that selecting 12 edges instead of 6 edges still brings great improvement of classification accuracy.

From the results of Table 1, 2, 3 and 4, we have several preliminary observations. 1) The performance of feature selection in terms of classification is not

Table 1. Classification results where $\lambda = 1.5$ and number of selected features is 6

NSR	GR		SS		SC	
	W/	W/O	W/	W/O	W/	W/O
0.2	1.00	0.89	1.00	0.99	1.00	0.80
0.4	1.00	0.87	1.00	0.94	1.00	0.80
0.6	0.99	0.82	0.99	0.84	1.00	0.80
0.8	0.96	0.73	0.98	0.82	0.94	0.64
1.0	0.95	0.76	0.92	0.84	0.98	0.68
1.2	0.88	0.67	0.90	0.72	0.86	0.62
1.6	0.75	0.53	0.72	0.52	0.78	0.54

Table 2. Classification results where $\lambda = 1.5$ and number of selected features is 12

NSR	GR		SS		SC	
	W/	W/O	W/	W/O	W/	W/O
0.2	1.00	0.89	1.00	0.99	1.00	0.80
0.4	1.00	0.87	1.00	0.94	1.00	0.80
0.6	1.00	0.82	1.00	0.84	1.00	0.80
0.8	0.96	0.73	0.98	0.82	0.94	0.64
1.0	0.92	0.76	0.88	0.84	0.96	0.68
1.2	0.78	0.67	0.74	0.72	0.82	0.62
1.6	0.69	0.53	0.72	0.52	0.66	0.54

Table 3. Classification results where $\lambda = 0.015$ and number of selected features is 6

NSR	GR		SS		SC	
	W/	W/O	W/	W/O	W/	W/O
0.2	1.00	0.89	1.00	0.99	1.00	0.80
0.4	1.00	0.87	1.00	0.94	1.00	0.80
0.6	0.99	0.82	0.98	0.84	1.00	0.80
0.8	0.97	0.73	1.00	0.82	0.94	0.64
1.0	0.89	0.76	0.90	0.84	0.88	0.68
1.2	0.88	0.67	0.90	0.72	0.86	0.62
1.6	0.81	0.53	0.82	0.52	0.80	0.54

Table 4. Classification results where $\lambda = 0.015$ and number of selected features is 12

NSR	GR		SS		SC	
	W/	W/O	W/	W/O	W/	W/O
0.2	1.00	0.89	1.00	0.99	1.00	0.80
0.4	1.00	0.87	1.00	0.94	1.00	0.80
0.6	0.96	0.82	0.96	0.84	0.96	0.80
0.8	0.97	0.73	0.96	0.82	0.98	0.64
1.0	0.94	0.76	0.94	0.84	0.94	0.68
1.2	0.83	0.67	0.82	0.72	0.84	0.62
1.6	0.76	0.53	0.80	0.52	0.72	0.54

strongly dependent on the choice of λ and therefore the sparse regression is reliable in practice, though the high noise level might prefer smaller λ while the low noise level might prefer larger one. 2) As the noise-to-signal ratio increases, the recognition performance deteriorate as expected, but the feature selection via sparse regression always brings significant better recognition accuracy. 3) The performance of feature selection in terms of classification is not strongly dependent on the prescribed number of selected features, if the adopted number is not far away from the true number of significantly discriminative features.

Acknowledgement. This work was supported by the Natural Science Foundation of China, Grant Nos. 61035006, 61125304, 11201054, 61273361 and by the Specialized Research Fund for the Doctoral Program of Higher Education of China 20120185110028, and by the Fundamental Research Funds for the Central Universities ZYGX2012J118, and by the Key Technology R&D Program of Sichuan Province 2012SZ0159.

References

1. Friston, K.J.: Functional and Effective Connectivity: A Review. Brain Connectivity 1(1), 13–36 (2011), doi:10.1089/brain.2011.0008.
2. Sporns, O.: The Human Connectome: Origins and Challenges (to appear 2013)
3. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust Face Recognition via Sparse Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 227(2), 210 (2009)
4. Wang, Y., Yang, J., Yin, W., Zhang, Y.: A new alternating minimization algorithm for total variation image reconstruction. SIAM Journal on Imaging Sciences 1(3), 248–272 (2008)
5. Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., Brovelli, A.: Multivoxel Pattern Analysis for fMRI Data: A Review. Computational and Mathematical Methods in Medicine, Article ID 961257, 14 (2012)
6. Li, Y., Namburi, P., Yu, Z., Guan, C., Feng, J., Gu, Z.: Voxel Selection in fMRI Data Analysis Based on Sparse Representation. IEEE Transaction on Biomedical Engineering 56(10) (2009)
7. Li, Y., Long, J., He, L., Lu, H., Gu, Z., et al.: A Sparse Representation-Based Algorithm for Pattern Localization in Brain Imaging Data Analysis. PLoS ONE 7(12) (2012)
8. Zhang, J., Cheng, W., Wang, Z., Zhang, Z., Lu, W., Lu, G., Feng, J.: Pattern Classification of Large-Scale Functional Brain Networks: Identification of Informative Neuroimaging Markers for Epileps. PLoS ONES 7(5) (2012)
9. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44, 112 (2009)
10. Liu, J., Ji, S., Ye, J.: SLEP: Sparse Learning with Efficient Projections, Arizona State University (2009), `http://www.public.asu.edu/~jye02/Software/SLEP`
11. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27:1–27:27 (2011), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
12. Su, L., Wang, L., Chen, F., Shen, H., Li, B., et al.: Sparse Representation of Brain Aging: Extracting Covariance Patterns from Structural MRI. PLoS ONE 7(5) e36147 (2012), doi:10.1371/journal.pone.0036147
13. Meinshausen, N., Bühlmann, P.: Stability selection. J. Roy. Statistical Society B 72, 417 (2010)