Automatic Classification of Schizophrenia via Adaptive Learning Algorithm using Resting-state Functional Language Network

[1]Maohu Zhu, [1]Nanfeng Jie, [2]Yuanchao Zhang and [1,2]Tianzi Jiang

[1]LIAMA Center for Computational Medicine, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2]Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 625014, P. R. China

{mhzhu,nfjie}@nlpr.ia.ac.cn, yuanchao.zhang8@gmail.com and jiangtz@nlpr.ia.ac.cn

Abstract. A reliable and precise classification of schizophrenia is significant for its diagnosis and treatment of schizophrenia. Functional magnetic resonance imaging (fMRI) is a novel tool increasingly used in schizophrenia research. Recent advances in statistical learning theory have led to applying pattern classification algorithms to access the diagnostic value of functional brain networks, discovered from resting state fMRI data. The aim of this study was to propose an adaptive learning algorithm to distinguish schizophrenia patients from normal controls using resting-state functional language network. Furthermore, here the classification of schizophrenia was regarded as a sample selection problem where a sparse subset of samples was chosen from the labeled training set. Using these selected samples, which we call informative vectors, a classifier for the clinic diagnosis of schizophrenia was established. We experimentally demonstrated that the proposed algorithm incorporating resting-state functional language network achieved 83.6% leave-one-out accuracy on resting-state fMRI data of 27 schizophrenia patients and 28 normal controls. In contrast with K-Nearest-Neighbor (KNN), Support Vector Machine (SVM) and l_1-norm, our method yielded better classification performance. Moreover, our results suggested that a dysfunction of resting-state functional language network plays an important role in the clinic diagnosis of schizophrenia.

Keywords: Schizophrenia, fMRI, resting-state, language network, pattern classification, sample selection, informative vector.

1 Introduction

Previous studies have shown that schizophrenia is associated with deficits in language function, as well as structural and functional abnormalities in brain regions that are involved with language perception and processing[1-4]. Individuals with high genetic risk for schizophrenia also have structural and functional deficits in brain pathways for language processing[1]. In this paper, we investigated resting-state functional disintegration in language network to classify the schizophrenia patients and normal controls.

Given a training set X of a number of samples $\{x_1, x_2, \ldots, x_n\} \in \mathbf{R}^{m\times n}$ with known labels $\{y_1, y_2, \ldots, y_n\}$, the goal of a classification algorithm is to infer a decision function $y = D(x)$ from the labeled training set. The decision function should predict the correct output value for any valid input object *x*. In order to measure the quality of the decision function, a loss function $L(x)$ is defined. In the current paper, we limit ourselves to the least square function:

$$L(x, X) = \operatorname*{Min}_{w \in R^n} \|X * w - x\| \qquad (1)$$

$$= \operatorname*{Min}_{w_i \in w} \sum_{i=1}^{n} \|w_i{}^* x_i - x\|$$

Where w=[w_1,…,w_n], is the weight vector of all training samples in the sample space and *x is* a test sample. It is known that the least square often yields a poor generalization performance because the solution *w* overfits the data. To solve this issue, a small group of samples could be selected from the training set to build a sparse decision function. A celebrated instantiation is in learning the prediction function of Support Vector Machine (SVM) [5], which only utilizes a limited subset of support vectors to characterize the decision boundary between two classes, rather than directly use all training examples. However, since we may not always be able to unravel the essence of every model, sometimes it is difficult to establish a sparse decision function from training examples.

The standard remedy for this problem is to impose a regularization condition of w to obtain a well posed problem. A good regularization method is l_0-norm regularization, which corresponds to the non-convex function, where we let $\|w\|_0 = |\{i : w_i \neq 0\}|$:

$$L(x, X) = \operatorname*{Min}_{w_i \in w} \sum_{i=1}^{n} \|w_i{}^* x_i - x\| + \lambda \|w\|_0 \qquad (2)$$

However, a fundamental issue with this method is the computational cost, as the number of subsets of $\{1, 2, \ldots, n\}$ of cardinality k (corresponding to the nonzero components of w) is exponential to k, It can be shown that the solution of this method is NP-hard[6], where no efficient algorithms are present. Due to computational difficulty, l_1-norm regularization, the closest convex approximation which often leads to sparse solutions, is proposed. A promising technique called LASSO was introduced by Tibshirani [7] as follow:

$$L(x,X) = \underset{w_i \in w}{Min} \sum_{i=1}^{n} ||w_i{}^{*} x_i - x|| + \lambda ||w||_1 \tag{3}$$

L_1-norm regularization is often exploited for feature selection. John *et al.* employed l_1-norm regularization to select the relevant training samples for the recognition of face images[8]. In order to generate a sparse solution, a large regularization parameter is required. However, the l_1 penalty not only shrinks the irrelevant variable to zero, but shrink relevant variables to zero[9]. Instead, greedy search strategies are known by experimentalists to be computationally advantageous and less prone to overfit [10]. In the current paper, we designed an adaptive learning algorithm incorporating resting-state functional language network to select a sparse subset of informative vectors that together were used to build a reliable classifier to classify schizophrenia patients from normal controls.

2 Materials

2.1 Subjects

Twenty-seven schizophrenia patients and 28 normal controls participated in this study. The control participants were group matched to the patients on age, handedness and sex (see Table 1). All schizophrenic patients were recruited from Peking University Sixth Hospital, China, and diagnosed with Diagnostic and Statistical Manual-IV criteria. Patients were free of any concurrent psychiatric disorders and had no history of major neurological or physical disorders leading to altered mental state. All patients accepted atypical psychotropic drugs at the time of scanning. Twenty-nine healthy subjects were recruited by advertisements as control group.

Table 1. Demographic and clinical details of the subjects

	Schizophrenia (N=27)	Control (N=28)	*P* value
Gender(Male/Femal)	12/15	12/16	>0.99**
Age	22.9±3.2	22.3±3.8	0.80*
Handdeness (Right/Left)	27/0	28/0	
Education	9.8±5.0	10.1±4.1	
PNASS	64.6		

*no significant between-group difference confirmed by chi square test (*p*>0.05).

**no significant between-group difference confirmed by two sample t test (p>0.05).

2.2 fMRI data acquisition and prepossessing

Imaging data was collected on a 3-Tesla SIEMENS scanner. Echo planar imaging blood oxygen level-dependent images of the whole brain were acquired in 30 axial slices (TR/TE = 2000/30 ms, flip angle = 90°, FOV = 22 cm, Slice Thickness = 4 mm and resolution = 3.44 × 3.44 × 4.8 mm3). The fMRI scanning was carried out in darkness, and the participants were instructed explicitly to keep their eyes closed and move as little as possible. For each participant, the fMRI scanning lasted for 6 minutes, during which 210 volumes were obtained.

We discarded the first 10 images and performed motion correction by rigid body alignment and slice timing correction using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). In the next, the realigned images are spatially normalized to the standard echo-planar imaging template and resampled to 3 × 3 × 3 mm3. Subsequently, the functional images were spatially smoothed with a Gaussian kernel of 6 × 6 × 6 mm3 full-width at half maximum to decrease spatial noise, temporally band-pass filtered (0.01 Hz < f < 0.1 Hz) using the AFNI (http://afni.nimh.nih.gov/), and motion corrected via linear regression. Finally, we removed global contributions to the time courses from the white matter, ventricles and the whole brain.

3 Methods

3.1 Adaptive learning algorithm

The selected samples, called informative vectors henceforth, are used to establish a classifier. Based on square error, we designed an adaptive learning algorithm that combines forward searching steps and backward adjusting steps. Unlike SVM, instead of choosing support vectors for all the test samples at once, a group of informative vectors for each test example were drawn in the current algorithm.

Starting with the null model without any training example, a pattern x_i, for which $L(F \cup \{x_i\})$ is the smallest (i.e., x_i decreases squared error the greatest), is added to the current set F by *forward step* in order to aggressively reduce the squared error at each step. This procedure keeps going on until the decrement of squared error falls below a given threshold ε (0.001 in this study). However, such procedure has a main shortcoming, that the selected subsets of samples are nested, where the subset F_k selected in *k*th step is always included by the subset F_{k+1}. This implies that the errors caused in earlier forward steps would never have a chance to be removed. Consequently, *backward steps* that aim to rectify these errors should be carried out. The key design of this combination is to balance the forward and backward steps. The backward steps should not only fix the errors induced by earlier forward steps, but also keep as many achievements as possible. The pseudocode of the adaptive algorithm is listed as follows.

```
Input:        $X = [x_1, \dots, x_n] \in R^{m\times n}$ for $l$ classess
              a test sample $x$
Initialize:  Each attribute of all dataset was linearly scaled to [0, 1]
```

```
                          $S = [1, \dots, n], F = \emptyset, w = \emptyset, k = 0,$
                          $\varepsilon = 0.001$ and $J_0 = \infty$
Output: $F, w$
while lengh($S$) $> 0$
{
  $k = k + 1;$
    $[i_k, w_k, J_k] = \text{argmin}_{i \in S} ||x, X(:, F \cup \{i\})||_2;$
    $\delta^+ = J_k - J_{k-1};$
    if $(\delta^+ < \varepsilon)$
    {
        $k = k - 1;$
        break;
    }
    $F = F \cup \{i_k\};$
    $S = S - \{i_k\};$
    $w = w_k;$
    while $(k > 1)$
    {
        $[j_k, w_k^-, J_k^-] = \text{argmin}_{j \in F} ||x, X(:, F - \{j\})||_2$
        $\delta^- = J_k^- - J_k;$
        if $(\delta^- < 0.5 * \delta^+)$
        {
            $S = S \cup \{j_k\};$
            $F = F - \{j_k\};$
            $w = w_k^-;$
            $k = k - 1;$
        } else
            break;
  }
}
```

Note that backward steps were only carried out when the squared error increment δ^- is no more than half of the squared error decrement in the earlier corresponding forward step δ^+. This means that as long as n forward steps have been performed, no matter how many backward steps were involved, the square error will always decrease by at least $n\varepsilon/2$, suggesting that the algorithm will automatically terminate after finite forward steps.

The selected informative vectors F=[F_1 ,..., F_k] and weight w=[w_1,...,w_k] were used to build a decision function $D(x)$ for each test sample x as follow:

$$y = D(x) = argmin_{j \in \{1,\dots,k\}} \left\| x, F_j{}^* w_j \right\| \quad (4)$$

Where F_j is the subset of informative vectors that belong to the jth class, and w_j corresponds to their weights, respectively. The decision function D(x) assigned a test sample x into the object class that minimizes the residual between x and all informative vectors from this class.

3.1 Feature extraction

To build a classifier that could distinguish schizophrenia patients and normal controls, blood oxygen level–dependent (BOLD) time courses were generated for 23 regions of interest (ROIs) that are believed to play an important role in language comprehension and production[12]. We defined the cubic region of 125 voxels (volume = 3.375 cm^3) centered at the coordinates of each ROI. The mean time series of each of 23 regions was obtained by simply averaging the fMRI time series over all voxels in this region. Correlation coefficients were then computed between each pair of these regions. For each subject, a 23-node, undirected graph of the functional language network with 253 edges was constructed.

Table 2. 23 ROIs MNI coordinates (BA: Brodmann area)

ROI	BA	x[mm]	y[mm]	z[mm]
Wernicke's area	39/40	-51	-51	30
Inferior parietal	40	57	-51	36
Broca's area	45	-51	27	18
Pars triangularis	45	51	30	18
Middle frontal	46	-39	18	45
Pars opercularis	44	42	21	42
Pars orbitalis	47	-45	39	-12
Pars orbitalis	47	45	39	-15
Inferior temporal	21/20	-57	-30	-15
Inferior temporal	21/20	63	-30	-12
Superior frontal	8	-3	36	45
Caudate		-12	9	15
Caudate		12	12	12
Putamen/globus pallidus		-18	0	9
Ventral thalamus		-9	-9	0
Cerebellum crus		15	-81	-30
Striate	17	6	-75	-6
Extrastriate	18	21	-69	-15
Posterior parietal	7	6	-81	45
Superior parietal	5	3	-51	57
Superior temporal	42	-63	-18	9

Superior temporal	42	60	-21	12
Cingulate	24	0	0	48

4 Experiment Results

A number of classification experiments were implemented with resting-state fMRI data to estimate the efficacy of the proposed classification algorithm and meanwhile compared with other machine learning algorithms.

Separating data into training and testing sets is crucial for evaluating prediction models. Typically, when partitioning a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. In order to avoid the possible bias introduced by relying on any one particular division into test and train components, a leave-one-out cross-validation (LOOCV) is used to split the p patterns into a training set of size $(p-1)$ and a test of size 1, and average the classification error on the left-out pattern over the p possible ways of obtaining such a partition. The merit of LOOCV is that all the data can be used for training—none has to be held back in a separate test set. The results in Table 3 demonstrated that our method outperformed SVMs and k-nearest-neighbor algorithm (KNN).

Table 3. Comparison of different learning algorithm

Methods	Classification accuracy	Sensitivity	Specificity
KNN	58.2%	59.3%	57.1%
SVM (linear kernel)	74.5%	70.4%	78.6%
SVM (RBF kernel)	80%	77.8%	82.1%
l_1-norm	74.5%	74.1%	75%
Our method	**83.6%**	**81.5%**	**85.7%**

Dataset scaling, which avoids attributes in greater numeric ranges dominate those in smaller numeric ranges, is quite crucial before subsequent procedures. Scaling also minimizes the numerical difficulties involved in the algorithm, where the least square errors heavily depend on the inner products of feature vectors[12]. Therefore, each attribute of all dataset were linearly scaled to the range [0, 1].

5 Conclusion

In this study, we demonstrated that an adaptive learning algorithm incorporating resting-state functional language network dramatically increased positive predictive power for the clinical diagnosis of schizophrenia. Different from SVMs, the proposed algorithm is instance-based learning that, instead of performing explicit generalization, compares new problem instances with instances seen in training, which have been stored in memory. One advantage of this algorithm is that it has zero empirical risk and infinite VC dimension. Compared with KNN that requires the orthogonality assumptions about samples, our algorithm massively utilizes mutual information between samples. Experimental results have suggested that taking into account of interactions among examples in the informative vectors selection process could have a great impact on classification performance. Our results also implied that a dysfunction of the language network plays a cardinal role in the clinic diagnosis of schizophrenia. Beyond classifying schizophrenia from control, an intriguing question for future work is whether this model can be applied for pinpointing robust differences in functional connectivity between a control and a clinical population.

Acknowledgments. This work was partially supported by the National Key Basic Research and Development Program (973) (Grant No. 2011CB707805), the Natural Science Foundation of China(Grant No. 91132301, No. 81101000 and No. 30900487), the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2011J098).

References

1. Li X, Branch CA, DeLisi LE (2009) Language pathway abnormalities in schizophrenia: a review of fMRI and other imaging studies. Current Opinion in Psychiatry 22: 131-139.
2. Sabb FW, van Erp TG, Hardt ME, Dapretto M, Caplan R, et al. (2010) Language network dysfunction as a predictor of outcome in youth at clinical high risk for psychosis. Schizophrenia research 116: 173-183.
3. Sommer I, Ramsey N, Kahn R (2001) Language lateralization in schizophrenia, an fMRI study. Schizophrenia research 52: 57-67.
4. Bleich-Cohen M, Hendler T, Kotler M, Strous RD (2009) Reduced language lateralization in first-episode schizophrenia: an fMRI index of functional asymmetry. Psychiatry Research: Neuroimaging 171: 82-93.
5. Cortes C, Vapnik V (1995) Support-Vector Networks. Machine Learning 20: 273-297.
6. Amaldi E, Kann V (1998) On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoretical Computer Science 209: 237-260.
7. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological): 267-288.
8. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31: 210-227.
9. Zhang T (2011) Adaptive forward-backward greedy algorithm for learning sparse representations. Information Theory, IEEE Transactions on 57: 4689-4708.
10. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Machine learning 46: 389-422.

11. Tomasi D, Volkow ND (2012) Resting functional connectivity of language networks: characterization and reproducibility. Molecular psychiatry 17: 841-854.

12. Hsu C-W, Chang C-C, Lin C-J (2003) A practical guide to support vector classification.