NeuroImage 117 (2015) 170–183

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Randomized structural sparsity via constrained block subsampling for improved sensitivity of discriminative voxel identification

Yilun Wang [a,b,c], Junjie Zheng [b], Sheng Zhang [a], Xunjuan Duan [b], Huafu Chen [b,*]

[a] School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731 PR China
[b] Key laboratory for Neuroinformation of Ministry of Education, School of Life Science and Technology, Center for Information in Biomedicine, University of Electronic Science and Technology of China, Chengdu, Sichuan 611054, PR China
[c] Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, USA

ARTICLE INFO

Article history:
Received 10 October 2014
Accepted 19 May 2015
Available online 28 May 2015

Keywords:
Voxel selection
Structural sparsity
Stability selection
Randomized structural sparsity
Constrained block subsampling
fMRI
Feature selection
Pattern recognition

ABSTRACT

In this paper, we consider voxel selection for functional Magnetic Resonance Imaging (fMRI) brain data with the aim of finding a more complete set of probably correlated discriminative voxels, thus improving interpretation of the discovered potential biomarkers. The main difficulty in doing this is an extremely high dimensional voxel space and few training samples, resulting in unreliable feature selection. In order to deal with the difficulty, stability selection has received a great deal of attention lately, especially due to its finite sample control of false discoveries and transparent principle for choosing a proper amount of regularization. However, it fails to make explicit use of the correlation property or structural information of these discriminative features and leads to large false negative rates. In other words, many relevant but probably correlated discriminative voxels are missed. Thus, we propose a new variant on stability selection "randomized structural sparsity", which incorporates the idea of structural sparsity. Numerical experiments demonstrate that our method can be superior in controlling for false negatives while also keeping the control of false positives inherited from stability selection.

1. Introduction

1.1. Problem statement

Decoding neuroimaging data, also called brain reading, is a kind of pattern recognition that has led to impressive results, such as guessing which image a subject is looking at from his brain activity (Haxby et al., 2001), as well as in medical diagnosis, e.g., finding out whether a person is a healthy control or a patient.

This pattern recognition typically consists of two important components: feature selection and classifier design. While the predictive or classification accuracy of these designed classifiers has received most attention in most existing literature, feature selection is an even more important goal in many practical applications including medical diagnosis where selected voxels can be used as biomarker candidates (Guyon and Elisseeff, 2003).

However, most traditional feature selection methods fail to discover in a stable manner the "complete" discriminative features accurately. They mainly aim to construct a concise classifier and they often select only a minimum subset of features, ignoring those correlated or redundant but informative features (Guyon and Elisseeff, 2003; Blum and Langley, 1997). In addition, the stability of the selected features is often ignored (Bühlmann and Van De Geer, 2011; Cover, 1965), because the inclusion of some noisy features or the exclusion of some informative features may not affect the prediction accuracy (Yu et al., 2008), which is their main objective. Therefore, a large number of uninformative, noisy voxels that do not carry useful information about the category label, could be included in the final feature detection results (Langs et al., 2011), while some informative, possibly redundant features might be missed.

In this paper, we focus on feature selection on functional MRI (fMRI) data where each voxel is considered as a feature. These features are often correlated or redundant. We focus on the "completeness" and "stability" of feature selection, i.e. aim to discover as many as possible informative but possibly redundant features accurately and stably, in contrast to most of the existing methods which mainly aim to find a subset of discriminative features which are expected to be uncorrelated. This way, potential biomarkers revealed by the discovered discriminative voxels, in both cognitive tasks and medical diagnoses are expected to be more credible.

1.2. Advantages and limitations of sparse priors in multivariate neuroimaging modeling

There are in general three main categories of supervised feature selection algorithms: filters, embedded methods, and wrappers (Guyon and Elisseeff, 2003). The filter methods usually separate feature section

[*] Corresponding author.
E-mail address: chenhf@uestc.edu.cn (H. Chen).

http://dx.doi.org/10.1016/j.neuroimage.2015.05.057
1053-8119/© 2015 Elsevier Inc. All rights reserved.

from classifier development. For example, Fisher Score (Duda et al., 2000), is among the most representative algorithms in this category. The wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set, and the features are scored according to their predictive power. The embedded models perform feature selection during learning. In other words, they achieve model fitting and feature selection simultaneously.

The following sparsity related feature selection models are all typical embedded methods, which we will mainly focus on in this paper.

In this paper, we consider commonly used supervised learning to identify the discriminative brain voxels from given training fMRI data.

While the classification problem is considered most often, the regression problem can be treated in a similar way. We consider the following linear model.

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{n \times 1}$ is the binary classification information and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the given training fMRI data and $\mathbf{w} \in \mathbb{R}^{p \times 1}$ is the unknown weights reflecting the degree of importance of each voxel. As a multivariate inverse inference problem, identification of discriminative voxels is based on the values of the weight vector $\mathbf{w}$ and their importance is proportional to the absolute values of the components. Therefore, feature selection is also called support identification in this context, because the features corresponding to the nonzero $\mathbf{w}$ components are considered as the relevant features.

Considering that the common challenge in this field is the curse of dimensionality p » n, we are focusing on sparsity-based voxel selection methods, because sparsity is motivated by the prior knowledge that the most discriminative voxels are only a small portion of the whole brain voxels (Yamashita et al., 2008).

However, sparsity alone is not sufficient for making reasonable and stable inferences. Plain sparse learning models often provide overly sparse and hard-to-interpret solutions where the selected voxels are often scattered (Rasmussen et al., 2012), though they might be useful if a concise classifier is expected. Specifically, if there is a set of highly correlated features, then only a small portion of representative voxels are selected, resulting into a large false negative rate and a potential biomarker that is hard to trust.

In addition, let denote the support of the true sparse vector $\overline{\mathbf{w}}$ as S and the number of its nonzeros as ℓ. For the success of finite sample recovery by the plain ℓ_1 norm regularized model, ℓ should be smaller than n. Let subsets of the columns of the design matrix $\mathbf{X}$ larger than ℓ must be well conditioned. In particular, the design matrix $\mathbf{X}_S$ should be sufficiently well conditioned and should not be too correlated to the columns of $\mathbf{X}$ corresponding to the noisy subspace $\mathbf{X}_{\bar{S}}$ (Varoquaux and Alexandre Gramfort, 2012).

Thus we have to extend the plain sparse learning model to incorporate important structural features of brain imaging data, such as brain segregation and integration, in order to achieve stable, reliable and interpretable results.

1.3. Existing extensions of the plain sparse model

As mentioned above, two common hypotheses have been made for fMRI data analysis: sparsity and compact structure. In sparsity, few relevant and highly discriminative voxels are implied in the classification task; in compact structure, relevant discriminative voxels are grouped into several distributed clusters, and the voxels within a cluster have similar behaviors and are, correspondingly, strongly correlated. Thus making use of these two hypotheses is very important, and we will review some state-of-the-art existing works in this direction.

Elastic net regression (Zou and Hastie, 2005) tries to make use of the voxel correlation by adding a ℓ_2 regularization, also called the Tikhonov regularization, to the classical ℓ_1 penalty (Ryali et al., 2012a) to deal with highly correlated features. Recently, other penalties have been added to consider the correlated features besides the Tikhonov regularization (Dubois et al., 2014). For example, both ℓ_1 penalization and Total-Variation (TV) penalization are used simultaneously for voxel selection (Gramfort et al., 2013), where the TV penalization is used to make use of the assumption that the activations are spatially correlated and the weights of the voxels are close to piece-wise constant. In addition, ℓ_2-fusion penalty can be used if successive regression coefficients are known to vary slowly and can also be interpreted in terms of correlations between successive features in some cases (Hebiri and van de Geer, 2011). While these models based on both ℓ_1 norm and other certain smoothing penalty, might achieve improved sensitivity over the plain ℓ_1 model, they do not make use of any explicit prior grouping or other structural information of the features (Xia et al., 2010).

Correspondingly, another class of methods to make more explicit use of the segregation and integration of the brain, is based on structured sparsity models (Bach et al., 2012b; Schmidt et al., 2011; Chen et al., 2012), which have been proposed to extend the well-known plain ℓ_1 models by enforcing more structured constraints on the solution. For example, the discriminative voxels are grouped together into few clusters (Baldassarre et al., 2012; Michel et al., 2011), where the (possibly overlapping) groups have often been known as a prior information (Xiang et al., 2012; Liu and Ye, 2010; Yuan et al., 2013; Jacob et al., 2009; Liu et al., 2009a; Ng and Abugharbieh, 2011). However, in many cases, the grouping information is not available beforehand, and one can use either the anatomical regions as an approximation (Batmanghelich et al., 2012), or the data driven methods to obtain the grouping information such as hierarchical agglomerative clustering (Ward hierarchical clustering, for example) and a top-down step to prune the generated tree of hierarchical clusters in order to obtain the grouping information (Michel et al., 2012; Jenatton et al., 2012).

While structural sparsity helps select the correlated discriminative voxels and is necessary for the “completeness” of the selected discriminative voxels, the result of feature selection may not be stable and is likely to include many noisy and uninformative voxels. For years, the idea of ensemble has been applied to reduce the variance of feature selection result (Hastie et al., 2009; Mota et al., 2014). Among them, one important class of methods for high dimensional data analysis is stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013). It is an effective way for voxel selection and structure estimation, based on subsamplings (bootstrapping would behave similarly).

It aims to alleviate the disadvantage of the plain ℓ_1 model, which either selected by chance non-informative regions, or even worse, neglected relevant regions that provide duplicate or redundant classification information (Mitchell et al., 2004; Li et al., 2012). This is due in part to the worrying instability and potential deceptiveness of the most informative voxel sets when information is non-local or distributed (Anderson and Oates, 2010; Poldrack, 2006). Correspondingly, one major advantage of stability selection is the control of false positives, i.e. it is able to obtain the selection probability threshold based on the theoretical boundary on the expected number of false positives. In addition, stability selection is not very sensitive to the choice of the sparsity panalty parameter, and stability selection has been applied to the pattern recognition based on brain fMRI data and achieved better results than plain ℓ_1 models (Ye et al., 2012; Cao et al., 2014; Ryali et al., 2012b; Mairal and Yu, 2013a; Meinshausen, 2013; Rondina et al., 2014). For example, SCoRS (Rondina et al., 2014) is an application of stability selection designed for the particular characteristics of neuroimaging data. Notice that we are focusing on the feature selection here. As for the prediction or classification accuracy, this ensemble or averaging idea has already been applied to reduce the prediction variance, and the examples include the bagging methods and forests of randomized trees (Breiman, 1996, 2001).