Journal of Neuroscience Methods 247 (2015) 1–12

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Computational Neuroscience

A statistical approach for detecting common features

Xinjun Gan[a], Bing Xu[a], Xiaoxi Ji[a], Wenlian Lu[a,b], David Waxman[a], Jianfeng Feng[a,b,c,d,*]

[a] Centre for Computational Systems Biology and School of Mathematical Sciences, Fudan University, Shanghai 200433, PR China
[b] Department of Computer Science, Warwick University, Coventry CV2 7AL, UK
[c] Shanghai Center for Mathematical Sciences, Fudan University, Shanghai 200433, PR China
[d] School of Life Science and the Collaborative Innovation Center for Brain Science, Fudan University, Shanghai 200433, PR China

HIGHLIGHTS

- We propose a new method to detect common features of multiple datasets.
- Simulation studies and theoretical comparison are performed to ascertain the advantage.
- We apply our approach to clinical data for different situations and the conclusions are consistent with literature.

ARTICLE INFO

Article history:
Received 26 November 2014
Received in revised form 10 February 2015
Accepted 11 February 2015
Available online 13 March 2015

Keywords:
Resting-state functional connectivity
Local false discovery rate
Common
FMRI

ABSTRACT

Background: With increasing numbers of datasets in neuroimaging studies, it has become an important task to pool information, in order to increase the statistical power of tests and for cross validation. However, no robust global approach unambiguously identifies the common biological abnormalities in, for example, resting-state functional magnetic resonance imaging in a number of mental disorders, where there are multiple datasets/attributes. Here we propose a novel and efficient statistical approach to this problem that finds common features in multiple datasets.
New method: By collecting the statistics of each dataset into a vector, our method uses a 'multi-dimensional local false discovery' rate to pool information and make full use of the joint distribution of datasets.
Results: We have tested our approach extensively on both simulated and clinical datasets. By conducting simulation studies, we find that our approach has a higher statistical power than existing approaches, especially on correlated datasets. Employing our approach on clinical data yields findings that are consistent with the existing literature.
Comparison with existing methods: Conventional methods cannot determine the false discovery rate underlying multiple datasets/attributes. Our approach can effectively handle these datasets. It has a solid Bayesian interpretation, and a higher power than other approaches in numerical simulations. This can be explained by the incorporation of correlations, between different attributes, into the new method.
Conclusions: In this work, we present a natural, novel and powerful statistical approach to tackle situations involving multiple datasets or attributes. This new method has significant advantages over existing approaches and wide applications.

1. Introduction

A challenging objective in neuroimaging is to identify reliable biomarkers, at the imaging and genetic levels, for mental disorders such as schizophrenia, bipolar disorder and depression (Guo et al., 2014; Phillips and Vieta, 2007; Tao et al., 2013). Despite the latest diagnostic manual (DSM-5) still clearly classifying mental illnesses into discrete categories, such as major-depressive disorder, bipolar disorder, schizophrenia and obsessive compulsive disorder (Adam, 2013), there is growing evidence that suggests that mental illnesses lie in continuum. For example, the classic theoretical Kraepelian division between schizophrenia and bipolar disorder has long been bridged by a pragmatic hybrid illness called schizoaffective disorder, in which experiences of positive symptoms (such as delusions

* Corresponding author at: Centre for Computational Systems Biology and School of Mathematical Sciences, School of Life Science and the Collaborative Innovation Centerfor Brain Science, Fudan University, Shanghai, China.
E-mail address: jffeng@fudan.edu.cn (J. Feng).

http://dx.doi.org/10.1016/j.jneumeth.2015.02.010
0165-0270/

and disordered thoughts) are shared by both disorders (Adam, 2013). A more detailed exploration is summarized by Buckholtz and Meyer-Lindenberg (2012) under the heading 'trans-diagnostic model approach'. To fulfill a trans-diagnostic model approach requires finding shared and distinctive risk factors in imaging or genetic data which could explain the comorbidity among different mental disorders. At the genetic level, recent successful work was reported in Smoller et al. (2013), where SNP's within two L-type voltage-gated calcium channel subunits, CACNA1C and CACNB2, were identified as risk loci, with shared effects on five psychiatric disorders (autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia). At the brain circuit level, there are a number of mental disorders where no robust global approach has been established that can unambiguously identify the common biological abnormalities in resting-state functional magnetic resonance imaging (fMRI) studies, although there are many published investigations that use a seed-based analysis, which is a biased approach that lacks a global and independent view. For example of a seed-based analysis, the decoupling of the dorsal lateral prefrontal cortex and the medial prefrontal cortex in both bipolar disorder and schizophrenia has been found in Lichtenstein et al. (2009).

We note that even for a single mental disorder, the reliability of the biomarkers that are identified in one dataset is often questionable and currently impossible to fully cross-validate in another experiment dataset. More generally, the hard and pervasive question is how to find common biomarkers with multiple datasets (a detailed description is given in Section 2). Meta-analysis methods (Borenstein et al., 2009) can be applied to deal with this situation when different datasets are focused on the same variable, for instance, functional connectivity. Nevertheless, meta-analysis approaches ignore correlations between datasets. Moreover, there is no statistical method to handle cases involving different attributes. As a typical example, consider the following two attributes: (i) the distance between altered functional connectivity networks of healthy controls and patients; (ii) the correlation between altered functional connectivity networks and the symptom of severity in patients (which is measured using Positive and Negative Syndrome Scale (PANSS) scores (Kay et al., 1987)). If we analyze the two attributes separately, it is quite possible that when the sample size is not large enough, the two significant subsets of features have no intersection, i.e. no common features. This is the typical issue we currently face, and it naturally leads us to analyze the two attributes simultaneously, by 'pooling' all of the information. In the current paper we present a natural, novel and powerful statistical approach to tackle the above issue.

We note that the traditional threshold method, in large-scale hypothesis testing, involves control of the family-wise error rate (FWER) using the Bonferroni procedure (Hochberg and Tamhane, 1987). After the landmark paper of Benjamini and Hochberg (1995), false discovery rate (FDR) methods for large-scale inference have been widely applied in bioinformatics (Storey and Tibshirani, 2003) and neuroimaging (Genovese et al., 2002; Logan and Rowe, 2004). The difference between FWER and FDR is that FWER corrections control the rate of any false positives in all *hypotheses*, while FDR corrections control the rate of false positives among all *positives* (Nichols and Hayasaka, 2003). The research in Tusher et al. (2001) developed a nonparametric method, named significance analysis of microarrays (SAM), to control FDR, which is a variation of the Benjamini and Hochberg (B-H) algorithm. Local false discovery rates (fdr) used empirical Bayesian techniques to provide both size and power calculations for large-scale studies (Efron et al., 2001).

Although FDR and local fdr have exhibited great power and have wide applications, when considering one dataset (i.e., a single attribute), these established statistical large-scale methods cannot solve the neuroimaging problems mentioned above. In particular, with multiple datasets (i.e., multiple attributes), the pressing aim is to identify the significant features of these datasets which are common (for an exact definition of *common significant features*, see Definition 2). To the best of our knowledge, there is no method that can handle all of these situations effectively. Traditional meta-analysis methods can deal with multiple datasets simultaneously (Borenstein et al., 2009), however, these do not consider the correlation between datasets. Additionally, we note that (Price and Friston, 1997; Friston et al., 1999) introduced a method called conjunction analysis, for estimating when two or more tasks activate the same region of the brain. This method calculates the minimum t-statistic from different datasets, ignoring the dependence among different attributes, with the result that it has low power (McNamee and Lazar, 2004; Heller et al., 2007). In the present paper we focus on developing a statistical approach to solve the trans-diagnostic problem in resting-state fMRI datasets. We achieve this by calculating the statistics of each dataset, combining them into a multi-dimensional statistical vector, and then defining a multi-dimensional fdr (especially a 2-dimensional (2-D) fdr), which is the extension of a one-dimensional local false discovery rate. Calculating a multi-dimensional fdr allows us to identify the common significant features that are shared by all datasets. Moreover, the quantity represented by a multi-dimensional fdr has a Bayesian interpretation, i.e., as the posterior probability that a feature is null, given its observed multi-dimensional statistical vector. To compare the statistical power with conventional methods, we have conducted extensive simulations. Furthermore, we have applied the new approach, presented in this work, to clinical data, and our findings are consistent with the existing literature, suggesting the validity of our approach. The proposed methods show various advantages over existing approaches, such as traditional meta analysis, and suggest extensive applications in neuroimaging research.

This paper proceeds as follows. In Section 2, we demonstrate the approach of multi-dimensional fdr, focussing on the 2-D case. In Section 3, we present simulations that illustrate the performance of the 2-D fdr procedure. To verify robustness and sensitivity, we applied the 2-D fdr method to three cases of clinical data, which covered the scenarios mentioned above (of finding common biomarkers with multiple datasets and attributes). In Section 4, we discuss advantages and disadvantages of the 2-D fdr method and suggest extensions of our work. In Appendix A we list the basics of false discovery rate and local fdr, along with properties of multi-dimensional fdr. In Appendix B we give a theoretical comparison of our approach with conventional methods. In Appendix C we present a useful transformation, the so-called *Fisher-transformation*, for handling correlation coefficients.

Our new algorithm is termed 'COMMON' (because it can identify common features of multiple datasets (i.e., multiple attributes)) and a user friendly Matlab interface may be downloaded from http://www.dcs.warwick.ac.uk/~feng/COMMON.html.

2. Methods

Before we describe the central aspect of this work, we shall introduce some of the basics of large scale testing of a single dataset. Suppose we have a dataset consisting of two groups of individuals (say patients and controls). Each individual of this dataset has m features. The hypothesis we are aiming to test is that the distribution of a particular feature, in both patients and controls, has the same distribution. If they have the same distribution, we shall say the feature is null; otherwise, we say the feature is non-null, or statistically significant. Given that there are m features, there are m hypotheses to simultaneously test. Throughout this paper, we always assume the m features within a dataset are statistically independent.