NeuroImage 76 (2013) 345–361

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Variational Bayesian mixed-effects inference for classification studies

Kay H. Brodersen [a,b,c,*], Jean Daunizeau [c,d], Christoph Mathys [a,c], Justin R. Chumbley [c], Joachim M. Buhmann [b], Klaas E. Stephan [a,c,e]

[a] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Switzerland
[b] Machine Learning Laboratory, Department of Computer Science, ETH Zurich, Switzerland
[c] Laboratory for Social and Neural Systems Research (SNS), Department of Economics, University of Zurich, Switzerland
[d] Institut du Cerveau et de la Moelle Épinière (ICM), Hôpital Pitié Salpêtrière, Paris, France
[e] Wellcome Trust Centre for Neuroimaging, University College London, UK

ARTICLE INFO

Article history:
Accepted 9 March 2013
Available online 16 March 2013

Keywords:
Variational Bayes
Fixed effects
Random effects
Normal-binomial
Balanced accuracy
Bayesian inference
Group studies

ABSTRACT

Multivariate classification algorithms are powerful tools for predicting cognitive or pathophysiological states from neuroimaging data. Assessing the utility of a classifier in application domains such as cognitive neuroscience, brain–computer interfaces, or clinical diagnostics necessitates inference on classification performance at more than one level, i.e., both in individual subjects and in the population from which these subjects were sampled. Such inference requires models that explicitly account for both fixed-effects (within-subjects) and random-effects (between-subjects) variance components. While models of this sort are standard in mass-univariate analyses of fMRI data, they have not yet received much attention in multivariate classification studies of neuroimaging data, presumably because of the high computational costs they entail. This paper extends a recently developed hierarchical model for mixed-effects inference in multivariate classification studies and introduces an efficient variational Bayes approach to inference. Using both synthetic and empirical fMRI data, we show that this approach is equally simple to use as, yet more powerful than, a conventional *t*-test on subject-specific sample accuracies, and computationally much more efficient than previous sampling algorithms and permutation tests. Our approach is independent of the type of underlying classifier and thus widely applicable. The present framework may help establish mixed-effects inference as a future standard for classification group analyses.

Introduction

Multivariate classification algorithms have emerged from the field of machine learning as powerful tools for predicting cognitive or pathophysiological states from neuroimaging data (Haynes and Rees, 2006). Classifiers are based on decoding models that differ in two ways from conventional mass-univariate encoding analyses based on the general linear model (GLM; Friston et al., 1995). First, multivariate approaches explicitly account for dependencies among voxels. Second, they reverse the direction of inference, predicting a contextual variable from brain activity (decoding) rather than the other way around (encoding). There are three related areas of application in which these two characteristics have sparked most interest.

In cognitive neuroscience, and in particular neuroimaging, classifiers have been employed to decode subject-specific cognitive or perceptual states from multivariate measures of brain activity, such as those obtained by fMRI (Brodersen et al., 2012b; Cox and Savoy, 2003; Haynes and Rees, 2006; Norman et al., 2006; Tong and Pratte, 2012). A second area is the design of brain–machine interfaces which aim at decoding subjective cognitive states (e.g., intentions or decisions) from trial-wise measurements of neuronal activity in individual subjects (Blankertz et al., 2011; Sitaram et al., 2008). A third important domain concerns clinical applications that explore the utility of multivariate decoding approaches for diagnostic purposes (Davatzikos et al., 2008; Klöppel et al., 2008, 2012; Marquand et al., 2010). Recently, decoding models have also been integrated with biophysical models of brain function, such as dynamic causal models (Friston et al., 2003), to afford mechanistically interpretable classifications (Brodersen et al., 2011a,b).

Many applications of multivariate classification operate on data with a two-level hierarchical structure. Consider, for example, a study in which a classification algorithm is used to decode from fMRI data whether a subject chose option A or B on each of *n* experimental repetitions or trials. This analysis gives rise to *n* estimated labels (representing which choice the classifier predicted on each trial) and *n* true labels (indicating which option was truly chosen). Comparing predicted to true labels yields a sequence of classification *outcomes* (indicating for each trial whether the prediction was correct or incorrect). Repeating this analysis for each member of a group of *m* subjects yields the typical two-level structure (*m* subjects times *n* trials each) that is illustrated in Fig. 1; for a concrete example see Figs. 7a,e. A two-level structure underlies virtually all trial-by-trial decoding studies (see, among many others,

* Corresponding author at: Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Wilfriedstrasse 6, CH 8032 Zurich, Switzerland.
E-mail address: brodersen@biomed.ee.ethz.ch (K.H. Brodersen).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.03.008

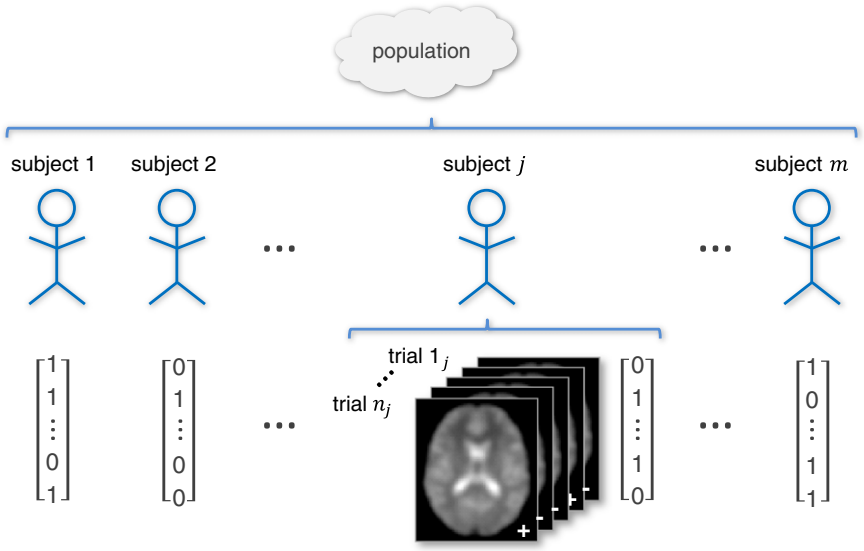

Fig. 1. Overview of the outcomes generated by a classification group study. In a trial-by-trial classification analysis, a classifier is trained and tested, separately for each subject, to predict a binary label (+ or −) from trial-wise correlates of brain activity. This constitutes a hierarchical design. The first level concerns trial-wise classification outcomes (where 1 and 0 represent correctly and incorrectly classified trials) that are drawn from latent subject-specific classification accuracies. The second level concerns subject-specific accuracies themselves, which are drawn from a population distribution. When evaluating the performance of a classification algorithm, we are interested in inference on subject-specific accuracies and on the population accuracy itself.

Brodersen et al., 2012b; Chadwick et al., 2010; Harrison and Tong, 2009; Johnson et al., 2009; Krajbich et al., 2009). The same two-level structure often applies to subject-by-subject classification studies (e.g., decoding a diagnostic state or predicting a clinical outcome), especially when subjects are partitioned into groups that are analyzed separately.

A hierarchical (or multilevel) design of this sort gives rise to the questions of what we can infer about the accuracy of the classifier in individual subjects, and what about the accuracy in the population from which the subjects were sampled. Any approach to answering these questions must provide a means of (i) *estimation* (e.g., of the accuracy itself as well as an appropriate interval that describes our uncertainty about the accuracy); and (ii) *testing* (e.g., whether the accuracy is significantly above chance). This paper is concerned with such subject-level and group-level inferences on classification accuracy for multilevel data.

The statistical evaluation of classification performance in non-hierarchical (e.g., single-subject) applications of classification has been discussed extensively in the literature (Brodersen et al., 2010a; Langford, 2005; Lemm et al., 2011; Pereira and Botvinick, 2011; Pereira et al., 2009). By contrast, relatively little attention has thus far been devoted to evaluating classification algorithms in hierarchical (i.e., group) settings (Goldstein, 2010; Olivetti et al., 2012). This is unfortunate since the field would benefit from a broadly accepted standard.

Such a standard approach to evaluating classification performance in a hierarchical setting should account for two independent sources of variability: *fixed-effects* (i.e., within-subjects) variance that results from uncertainty about the true classification accuracy in any given subject; and *random-effects* variance (i.e., between-subjects variability) that reflects the distribution of true accuracies in the population from which subjects were sampled. This distinction is crucial because classification outcomes obtained in different subjects cannot be treated as samples from the same distribution; in a hierarchical setting, each subject itself has been sampled from a population with an unknown intrinsic heterogeneity (Beckmann et al., 2003; Friston et al., 2005). Models that explicitly separate both sources of uncertainty are known as *mixed-effects* models. They are the objects of interest in this paper.

Contemporary approaches to performance evaluation in classification group studies fall into several groups.[1] One approach rests on the *pooled sample accuracy*, i.e., the number of correctly predicted trials, summed across all subjects, divided by the overall number of trials. The statistical significance of the pooled sample accuracy can be assessed using a simple classical binomial test (assuming the standard case of binary classification) that is based on the likelihood of obtaining the observed number of correct trials (or more) by chance (Langford, 2005). A less frequent variant of this analysis uses the *average sample accuracy* instead of the pooled sample accuracy (Clithero et al., 2011).

A second approach, more commonly used, is to consider *subject-specific sample accuracies* and estimate their distribution in the population. This method typically (explicitly or implicitly) uses a classical one-tailed *t*-test across subjects to assess whether the population mean accuracy is greater than what would be expected by chance (e.g., Harrison and Tong, 2009; Knops et al., 2009; Krajbich et al., 2009; Schurger et al., 2010).

In the case of single-subject studies, the first method (i.e., a binomial test on the pooled sample accuracy) is an appropriate approach. However, there are three reasons why neither method is optimal for group studies. Firstly, both of the above methods neglect the hierarchical nature of the experiment. The first method (based on the pooled sample accuracy) represents a fixed-effects approach and disregards variability across subjects. This leads to overly optimistic inferences and provides results that are only representative for the specific sample of subjects studied, not for the population they were drawn from. The second method (*t*-test on sample accuracies) does consider random effects; but it neither explicitly models the uncertainty associated with subject-specific accuracies, nor does it account for violations of homoscedasticity (i.e., the differences in variance of the data between subjects).

[1] This paper focuses on *parametric* models for performance evaluation. While *non-parametric* methods are available (e.g., based on permutation tests), these methods can be very time-consuming in hierarchical settings and are not considered in detail here (see e.g. Hassabis et al., 2009; Just et al., 2010; Pereira and Botvinick, 2011; Pereira et al., 2009; Stelzer et al., 2013).