NeuroImage: Clinical 4 (2014) 687–694

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions

Quentin Noirhomme[a,b,*], Damien Lesenfants[a,b], Francisco Gomez[c], Andrea Soddu[d], Jessica Schrouff[a,e], Gaëtan Garraux[a], André Luxen[a], Christophe Phillips[a,f,1], Steven Laureys[a,b,1]

[a] Cyclotron Research Centre, University of Liège, Liège, Belgium
[b] Coma Science Group, Neurology Department, University Hospital of Liège, Liège, Belgium
[c] Complexus Group, Computer Science Department, Universidad Central de Colombia, Bogotá, Colombia
[d] Department of Physics & Astronomy, Brain and Mind Institute, University of Western Ontario, London, ON, Canada
[e] Laboratory of Behavioral and Cognitive Neurology, Stanford University, Palo Alto, USA
[f] Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium

ARTICLE INFO

Article history:
Received 20 December 2013
Received in revised form 8 April 2014
Accepted 8 April 2014

Keywords:
classification
cross-validation
binomial
permutation test

ABSTRACT

Multivariate classification is used in neuroimaging studies to infer brain activation or in medical applications to infer diagnosis. Their results are often assessed through either a binomial or a permutation test. Here, we simulated classification results of generated random data to assess the influence of the cross-validation scheme on the significance of results. Distributions built from classification of random data with cross-validation did not follow the binomial distribution. The binomial test is therefore not adapted. On the contrary, the permutation test was unaffected by the cross-validation scheme. The influence of the cross-validation was further illustrated on real-data from a brain–computer interface experiment in patients with disorders of consciousness and from an fMRI study on patients with Parkinson disease. Three out of 16 patients with disorders of consciousness had significant accuracy on binomial testing, but only one showed significant accuracy using permutation testing. In the fMRI experiment, the mental imagery of gait could discriminate significantly between idiopathic Parkinson's disease patients and healthy subjects according to the permutation test but not according to the binomial test. Hence, binomial testing could lead to biased estimation of significance and false positive or negative results. In our view, permutation testing is thus recommended for clinical application of classification with cross-validation.

1. Introduction

In the last few years, there has been a growing interest in the statistical assessment of classification results in biomedical applications. Machine learning approaches are now increasingly used to study brain function (Etzel et al., 2009; Pereira et al., 2009; Lemm et al., 2011) and have been proposed as a diagnostic and prognostic tool for patients (e.g., in the field of severe brain injury see (Phillips et al., 2011; Galanaud et al., 2012; Luyt et al., 2012; Lule et al., 2013) or Parkinson disease (Focke et al., 2011; Orru et al., 2012; Schrouff et al., 2012; Garraux et al., 2013; Schrouff et al., 2013)). Such classification machines have also been designed for many other applications such as analyzing DNA microarray and predicting tumor subtype and clinical outcome (Golub et al., 1999; Simon et al., 2003). Limitations and controversies of these approaches have been recently highlighted in a study using brain–computer interfaces (BCIs) to unravel signs of consciousness in patients with disorders of consciousness (Cruse et al., 2011; Goldfine et al., 2013). A statistically significant classification accuracy is one where we can reject the null hypothesis that there is no information about task, patient diagnosis or outcome in the data from which it is being predicted. In a two-class problem with an equivalent number of elements in each class, e.g., disease vs. no-disease, the theoretical chance level, which is valid in the case of an infinite number of trials, is 50%. In practice, we only have a limited number of trials, which can be in the order of 10, due to patient fatigue. If a specific set of features can classify the data with for example 58% accuracy, the question is whether this accuracy is trustworthy. To tackle this issue, several approaches have been proposed in the literature.

A frequently used method is based on the binomial distribution (Müller-Putz et al., 2008; Pereira et al., 2009; Billinger et al., 2013). With a limited number of trials, the results of a classifier are seen as the results of tossing a coin, an unfair coin, which can be modeled as a Bernoulli trial with probability $p = 50\%$ of success. The probability of achieving k successes out of N independent trials is given by the

[1] Both authors contributed equally.
* Corresponding author.
E-mail address: quentin.noirhomme@ulg.ac.be (Q. Noirhomme).

2213-1582/$ - see front matter
http://dx.doi.org/10.1016/j.nicl.2014.04.004

binomial distribution. Knowing the distribution and a given *p*-value, we can compute a lower bound for any classification accuracy. If the lower bound is higher than the chance level, we can reject the hypothesis that the accuracy was obtained by chance. Here, we are only interested on the accuracies higher than the chance level. We are not interested in the chance of coincidental deviations below the expected 0.50 because we would not pretend our features contain information in that case. Another approach is based on the Pearson chi-square coefficient (Kubler and Birbaumer, 2008). However, for small number of trials, as it is often the case in the neuroimaging and electrophysiology literature, this approach is not reliable (Pereira et al., 2009) and matches the binomial test for higher number of trials (Howell, 2012).

Alternatively, permutation test based methods (Good, 2005) have been employed (Mukherjee et al., 2003; Etzel et al., 2009; Pereira et al., 2009; Schrouff et al., 2013b). A permutation test is a non-parametric test that has also been proposed as a substitute to the Student *t*-test in functional neuroimaging (Nichols and Holmes, 2002) and electrophysiology (Maris and Oostenveld, 2007) experiments. A permutation test estimates the distribution of the null hypothesis from the data. Assuming that there is no class information in the data, the labels are randomly permuted and the accuracy computed with the new labels. As the new labels are random, the new accuracy estimate is expected to reflect the chance distribution. The permutation is repeated hundreds to thousands of times. Then, the *p*-value is given by the fraction of the sample that is larger than or equal to the accuracy actually observed when using the correct labels.

To estimate classification accuracy, ideally, the original data are split into two independent, complementary subsets: a training set (which is used to train the classifier and to define all parameters) and a testing set (which is used to validate the results). In practice, with small datasets, a cross-validation (CV) scheme is often used. The process of splitting the data into two is repeated several times using different partitions. The results obtained from all partitions are then averaged (Lemm et al., 2011). The classification accuracy can then be tested. Following common practice (Pereira et al., 2009; Pereira et al., 2011), the accuracy estimate obtained through a CV could be treated as if it came from a single classifier. In that case, the binomial test sees all accuracies as independent.

In the following, we will show on simulated and real data that the CV scheme has an effect on the calculation of the chance level and that this influence is accounted for by the permutation test but not by the binomial test. We will first present results from simulated data illustrating the influence of the CV scheme. Next, we will exemplify how this may influence the "diagnosis" of patients with disorders of consciousness on real data from a previous EEG-based brain–computer interface (BCI) study (Lule et al., 2013). We will then further illustrate the influence with an fMRI study on activation patterns in Parkinson's disease (Cremers et al., 2012;Schrouff et al., 2012, 2013a). Finally, we will discuss some hypotheses underlying the observed differences between classification testing methods. Our simulations make a simplifying assumption, e.g. type of features, and our example from real data does not cover all possible data source and classification approaches, but the issues presented here are quite general and apply to studies employing a cross-validation scheme to estimate the accuracy of the data.

2. Material and methods

2.1. Simulated data

To test the validity of the binomial and permutation tests to assess classification accuracy, we generated random datasets for a two-class problem. We simulated three cases. First, we tested several scenarios with low number of features and trials. Second, we tested the influence of the number of repetitions of the CV scheme. Third, we tested scenarios with high number of features and low number of trials as often the case in the neuroimaging literature. The generation of the random data and the classifiers used built-in MATLAB (The MathWorks, Natick, MA, USA) functions (*rand, randperm, classify*)[1] and libsvm functions (Chang and Lin, 2011). Datasets were generated with 10,000 simulations. Each simulation included two sets with an equal number of trials. Trial number was 100, 50 or 30. Trials of the 100 trial set (respectively 50 and 30 trial sets) had 40 features (respectively 20 and 10). Features and labels were randomly assigned 0 and 1 (*rand* function thresholded at .5). We tested four CV schemes. In an ideal CV scheme, all possible partitions of the data should be tested. This is the case for the leave-one-out (LOO) CV but in practice for classical N-fold CV schemes it is computationally intractable. Nevertheless, repeating the N-fold CV several times with different partitions is recommended and can reduce the variance of the estimator (Efron and Tibshirani, 1997; Etzel et al., 2009; Lemm et al., 2011). The CV schemes were LOO, 10-fold, 5-fold and 2-fold CVs. The first three are the most used and recommended in the literature (e.g., Lemm et al., 2011). The 2-fold CV is an extreme case at the opposite of the LOO CV. A linear discriminant analysis and a support vector machine (Burges, 1998) with linear kernel classified the data.

To compute the binomial lower bound, the binomial distribution is often approximated by a normal distribution; for example to compute the Wald interval or adjusted Wald interval (Kohavi, 1995; Martin and Hirschberg, 1996; Berrar et al., 2006; Billinger et al., 2013). However, the approximation of the binomial distribution by the normal distribution is only valid whenever the number of trials N and the accuracy p satisfy the following equation: $N \times p \times (1-p) > 5$ (Berrar et al., 2006). In the absence of problem specific knowledge, the best choice for estimation of the bound is derived from Jeffreys' Beta distribution (Martin and Hirschberg, 1996; Berrar et al., 2006). This approximation is adequate for $10 \leq N$ (Martin and Hirschberg, 1996). The binomial lower bound (λ) was computed using Jeffreys' Beta distribution (Berrar et al., 2006) as follows:

$$\lambda \approx \left\{ a + \frac{2(N-2m)z\sqrt{0.5}}{2N(N+3)} \right\} - z\sqrt{\frac{a(1-a)}{N+2.5}}$$

where N is the number of trials, m is the number of successful trials, a is the estimated accuracy and z is the z-score (1.65 for one sided test with $p < .05$ (resp. 2.33 for $p < .01$)).

The permutation test (Good, 2005) was based on 999 permutations plus the original accuracy (Ojala and Garriga, 2010). Only accuracies higher than 0.5 were assessed using permutation testing. We did not compute permutation test for accuracies smaller or equal than 0.50 because we would not pretend that our classifications contain information in that case. The permutation test consisted of randomly exchanging the label and classifying the data with the CV scheme. The *p*-value was calculated as the sum of all values of the permutation distribution equal or higher than the results of the original data divided by the number of permutations.

In a first experiment, 12 datasets were built, three for each of the four CV schemes with 100, 50 or 30 trials, and with 10,000 simulations each. Every simulation involved two subsets with an equal number of trials and features. First, the classification accuracy of the trials from the first subset obtained with linear discriminant analysis was assessed with a chosen CV scheme (Fig. 1A). The distribution of accuracies obtained from all simulations was called: CV distribution. Second, to build an empirical binomial distribution, all trials from the first subset were used to train a classification algorithm which was applied to the second, independent, subset (Fig. 1B). A third distribution, the CV-independent distribution, was built by applying a mixed CV scheme where the N-1 training folds came from the first subset

[1] The MATLAB code can be found at https://github.com/CyclotronResearchCentre/BinomPermTest .