Classification-based Tests for Neuroimaging Data Analysis: Comparison of Best Practices

Muhaddisa Barat Ali and Emanuele Olivetti
NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation, Trento, Italy
Centro Interdipartimentale Mente e Cervello (CIMeC), University of Trento, Italy
muhaddisabarat.ali@unitn.it

***Abstract*—In neuroimaging data analysis, classification algorithms are frequently used to discriminate between two populations of interest, like patients and healthy controls, or between stimuli presented to the subject, like face and house. Usually, the ability of the classifier to discriminate populations is used within a statistical test, in order to evaluate scientific hypotheses. In the literature, different procedures are adopted to carry out such tests, like using permutations, assuming the binomial model or using confidence intervals. Moreover multiple choices are made by practitioners when implementing those tests, like the actual classification algorithm or the use of a resampling scheme. In this work we analyze those procedures and some of those choices with respect to their effect on the Type I (false discovery) and Type II (sensitivity) errors. With a simulation study, we compare the different procedures and show the impact in practice. The final aim is to characterize the best practices and give more insight for their use.**

***Index Terms*—classification algorithms ; testing hypotheses ; Type I and II error**

I. Introduction

Over the last decade, a growing number of studies has shown that classification algorithms can be used for pattern discrimination on neuroimaging data. For example, in clinical studies, classifiers are used to discriminate patients from healthy subjects on the basis of structural magnetic-resonance imaging (MRI) data [1]. When studying cognitive functions, e.g. by presenting multiple visual stimuli to the subject, classifiers are used to predict the category of the stimulus from the concurrent brain activity [2]. Even when creating brain maps, classifiers can be used to score each brain location, e.g. voxel [3], to show their relevance to the brain function under scrutiny.

The ability of the classifier to predict the correct category of the stimulus (or of the subject/patient), is considered positive evidence in support of the hypothesis that category-related information is present within the recorded brain data [4]. For this reason, in order to make inference, the classifier is used within a hypothesis testing framework, typically the frequentist one. Practitioners usually implement tests such as the binomial test, or the permutation test, or compute confidence intervals [4], to answer the question whether the classifier performed better than chance. The adopted test statistic is usually a measure of how correct the predictions of the classifiers are, like classification accuracy [4].

Conducting a statistical test with classifiers requires the practitioner to make a number of choices. Among them, the choice of the specific classification algorithm (linear, non-linear, etc.), the scoring function (e.g. accuracy), the estimation process of the score (e.g. k-fold cross-validation), the null distribution (e.g. binomial or based on resampling). Moreover the inferential process can be carried out with a significance test or, alternatively, with confidence intervals [4].

The exact choices made during the inferential process may have a substantial impact on the result and, more importantly, on the frequency of Type I error, i.e. the incorrect rejection of a true null hypothesis, and Type II error, i.e. the failure to reject a false null hypothesis (see [5]). It is our opinion that the literature on this topic is far from complete. Most of the related work is either on describing the available choices (see for example [4]) or on applying them. In a some cases, more prescriptive indications are provided for specific steps of the inferential process, like to improve the stability of cross-validation through multiple repetitions [6], or to avoid the bias of the binomial assumption for the null distribution through the use of permutation testing [7], [8] and indications of scheme to conduct it [9]. Nevertheless, a number of open questions remain unanswered.

In this work, we study some of the choices in the inferential process and their impact on Type I and Type II errors of the tests through simulation. In the proposed simulation, we adopt a simple generative model which shares some aspects of the actual neuroimaging data collected during real experiments, like the small sample size, the dimensionality and sparseness. We do not try to accurately mimic neuroimaging data, for two main reasons. First, we believe that the properties that we are investigating are general, and not necessarily specific of certain kinds of neuroimaging data, like those from MRI or from the magneto/electro-encephalographer (M/EEG). Second, to the best of our knowledge, accurate neurophysiological models of the signal recorded by neuroimaging devices are generally not available. When they are available, they are extremely complex to use and unfit for a setting like ours.

With the help of simulations, in this work we show that, in case of small sample size, the use of the binomial assumption increases the Type I error (see [8]) but also decreases the Type II error, i.e. has increased sensitivity. We observe this effect both in the high-dimensional setting, typical of neuroimaging data, and in the low dimensional setting, typical of certain brain mapping procedures like the searchlight algorithm [3]. Moreover, we observe that increasing the number of folds of cross-validation reduces the frequency of Type II error, even though at the cost of some increase of Type I error. Additionally, we observe that the additional flexibility of classifiers with more hyperparameters can be detrimental in case of low sample size.

In the remaining part of the paper we introduce and discuss all the previous points. Specifically, in Section II, we introduce the notation and the main theoretical ingredients used in Section III, where we report the results of the simulations designed to investigate our claims. To conclude, in Section IV, we discuss the claims in the light of the experimental results.

II. Methods

In this section we introduce the notation, hypothesis testing and some basic concepts about the use of classifiers in statistical tests.

A. Notation

Let $Y \in \mathcal{Y}$ be a binary random variable describing the category of data recorded with a neuroimaging device. In the case of a cognitive neuroscience investigation, Y can be the category of the stimulus presented to the subject, e.g. face vs. house, and in case of a clinical study Y can be the category of the individual, e.g. healthy subject vs. patient.

Let $X \in \mathbb{R}^d$ be the random vector of pre-processed neuroimaging data with category Y. The realization of X, i.e. x, can be, for example, the values of a beta-map of functional MRI (fMRI) data recorded during the presentation of one visual stimulus, i.e. a trial. Typically, d is in the order of $10^2 - 10^3$. We call the pair (x, y) as *example*.

As a result of the data collection phase of neuroimaging experiment, a dataset $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ of N examples is collected. N may be in the order of a few tens, equally distributed between the two categories.

B. Hypothesis Testing

In experimental science, hypotheses about the phenomenon under observation are formulated and then tested in the light of the data collected during experiments. The most common paradigm to test hypotheses is the frequentist one [10], which consists of the following steps:

1) Set up the null hypothesis H_0 to disprove.
2) Define an appropriate test statistic T, which is a function that, given the collected data, summarizes them into a real number.
3) Compute $p(T|H_0)$, i.e. the distribution of T when H_0 is true. Decide the rejection regions R, i.e. the values for the T such that H_0 has to be rejected (see later).
4) Run the experiment, collect the data and compute T^* as the value of the test statistic for the observed data.
5) Reject H_0 if $T^* \in R$.

A test is characterized by the frequency of Type I error:

$$P(\text{Type I}) = P(\text{rejecting } H_0 | H_0 \text{ is true}) \quad (1)$$

and by the frequency of Type II error:

$$P(\text{Type II}) = P(\text{not rejecting } H_0 | H_0 \text{ is false}) \quad (2)$$

There is a trade-off between the two quantities and, commonly, $P(\text{Type I})$ is set to a small value, e.g. 0.05 or 0.01, while $P(\text{Type II})$ is obtained from the analysis of the test, called *power analysis*, since $power = 1 - P(\text{Type II})$. The rejection regions R of the test are usually defined by setting a threshold on T that depends on the null-distribution and the pre-defined $P(\text{Type I})$, often called *significance level*.

C. Confidence Intervals

In the literature [4], it is reported that *confidence intervals* (CIs) can be used to quantify evidence about a hypothesis. The procedure is based on estimating the interval for the *true* value of the test statistic T, through the observed T^* plus model assumptions or resampling techniques. A confidence interval with a p *confidence level* means that, when repeating the experiment multiple times, in the long term, a fraction p of the times the confidence interval will include the true value of T. Typically, the confidence level is chosen as 0.95 or 0.99. It is common practice to use CIs to decide about the hypothesis, i.e. to see whether the CI excludes or not values of T expected from H_0.

D. Classification-Based Test

A classifier $f \in \mathcal{F}$ is a function $f : \mathcal{X} \mapsto \mathcal{Y}$ that returns the predicted class label of x. Classifiers are usually trained on a portion of the dataset D, called train set D_{train}. In order to quantify the ability of f to correctly predict, its performance is measured, the most common measure being the generalization error $\epsilon = E_{\mathcal{X}\times\mathcal{Y}}[I(Y, f(X))]$, where I is the indicator function. The standard unbiased estimator of ϵ is the error rate $\hat{\epsilon} = \frac{1}{|D_{test}|}\sum_{(x,y)\in D_{test}} I(y, f(x))$, where $D_{test} = D \setminus D_{train}$. A complementary and equivalent measure of performance is classification accuracy, $acc = 1 - \epsilon$, and its estimate $\hat{acc} = 1 - \hat{\epsilon}$. In certain cases, it may be convenient to resort to the discrete version $\hat{\epsilon}$, i.e. the number of incorrect predictions e. Other measures may be more effective in assessing performance in case of imbalanced data, like [11], [12], but are less popular.

The values $\hat{\epsilon}$, e, or $\hat{acc}$ may show high variability for small N. Moreover, the split of D in D_{train} and D_{test} is non-deterministic, adding more variability to the estimate. In order to reduce such variability, it is common to adopt a resampling technique, like k-folds cross-validation (k-CV). With k-CV, k estimates of the performance measure are produced and then averaged to improve stability.

Beyond the parameters that are fit during the training phase, classifiers usually have further parameters, called hyperparameters, that need to be set before training. For example, the regularization coefficient of regularized linear models, or the C and γ parameters of radial basis function support vector machines (RBF-SVMs). To this end, part of the data is used to estimate those parameters in advance. The standard process to estimate those parameters is based on a nested CV scheme [13]. It is important to note that parameter estimation, training and classifier evaluation compete in the use of the available data because N is fixed and they need non-overlapping sets of examples in order to avoid circularity.

Classifiers can be used within statistical tests to assess whether there is category-related information within the data. The measure of performance of the classifier is used as test statistic T within a test procedure (see Section II-B). The null-distribution of T under H_0, i.e. $P(T|H_0)$, depends on the actual choice of T. In case $T = e$, a typical choice is that $P(e|H_0) = Bin(e|N, p = \frac{1}{2})$ [4], which assumes examples to be independent and identically distributed (i.i.d.). Another popular choice, that is applicable for every choice of T, is based on the permutation test. In the permutation test, the vector of class-labels $[y_1, \ldots, y_N]$ is permuted in order to artificially break the (possible) systematic difference between the categories. Then, the test statistic is computed, now following $P(T|H_0)$ by construction. The two steps are repeated for all possible permutations (or a random subset, as approximation) in order to estimate the null distribution of T.

Classifiers can be used also together with confidence intervals (CIs). Typically, the confidence interval of the measure of performance is derived and tested whether it excludes or not the values associated with chance-level, i.e. H_0. The interval is usually defined through the binomial assumption, as described in detail in [14]. Resampling techniques, like k-CV or the bootstrap, can be adopted as well but are less common for the context of CIs and more complex to be properly implemented.

III. Experiments

We generated a large number of simulated datasets to study how different choices in the data analysis procedure affected the Type I error and Type II error. Each dataset consisted of examples/vectors from two classes. The probability distribution of each class was a multivariate Gaussian. One class had always the zero vector as mean ($\mu_A = \mathbf{0}$) and the identity matrix as covariance ($\Sigma_A = I$). By modulating the mean (μ_B) and covariance (Σ_B) of the second class, we created different scenarios:

- $\mu_A = \mu_B$, $\Sigma_A = \Sigma_B$. By keeping the same covariance and mean, we generated cases where there is no systematic difference between the classes, i.e. the H_0 is true by design. By counting the incorrect rejections of H_0 over many repetitions of the tests, we could estimate the Type I error.
- By gradually moving away μ_B from μ_A, we could gradually reduce the overlap between the classes, i.e. increase the effect size.
- With $\Sigma_A = \Sigma_B$ we could test the case in which the optimal Bayes classifier is linear, i.e. an hyperplane. By changing the value of the elements of Σ_B[1], we modulated the non-linearity of the optimal Bayes classifier.

Another main parameter of the simulation was the number of dimensions (d) of the feature space. Moreover, in order to obtain a setting more similar to typical neuroimaging data, we considered a further parameter, $d_{inf} \leq d$, representing the number of *informative* dimensions, i.e. the number of dimensions actually affected when shifting μ_B away from μ_A: $\mu_B = \mu_A + \Delta$, where $\Delta_i = \delta$ when $i \leq d_{inf}$, otherwise $\Delta_i = 0$. In this way, we introduced a *sparsity* parameter, i.e. way to simulate sparsity in the feature space, which is typical in neuroimaging data.

From the distributions described above, we drew 1000 datasets with the same number of examples per class and counted the number of times each test of hypothesis failed, incurring in a Type I or a Type II error. We tested different sample sizes ($N \in [10, 100]$), different dimensions ($d \in [5, 300]$, $d_{inf} = [5, min(50, d)]$), different effect sizes ($\delta \in [0, 1]$) and different degrees of non-linearity of the optimal boundary between the classes ($\sigma_B^i \in U[1-\lambda, 1+\lambda]$, for $\lambda \in [0, 2]$ and $i \leq d_{inf}$). We analyzed the datasets with multiple classifiers, i.e. Logistic Regression, Linear SVM, RBF SVM and different significance levels (0.05 and 0.01) and confidence levels (0.95, 0.99). When exploring the different configurations above, we compared different choices with statistical tests. For lack of space, here we show only part of the results.

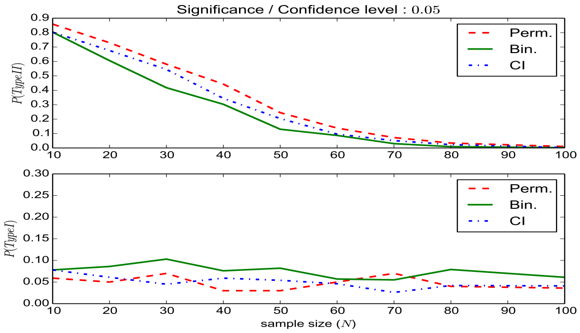

Fig. 1. Comparison of $P(TypeI)$ and $P(TypeII)$ as a function of the sample size for the binomial assumption (as null-distribution and for confidence interval) vs. the permutation-based null distribution. Classifier: linear SVM, $d = 200$, $d_{inf} = 10$, $\delta = 0.5$, $\lambda = 0$, $k = 5$.

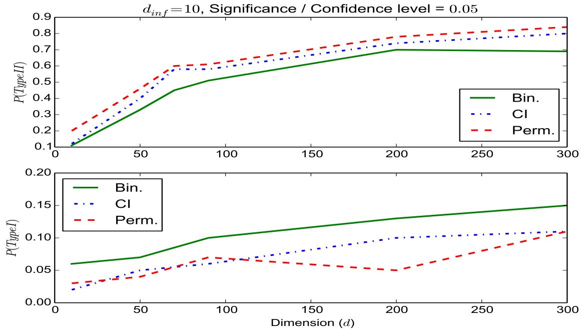

Fig. 2. Comparison of the three tests in terms of $P(TypeI)$ and $P(TypeII)$ with fixed $d_{inf} = 10$ but increasing dimension size (d). Classifier: linear SVM, $N = 40$, $\delta = 0.5$, $\lambda = 0$, $k = 5$.

Figure 1 shows the the binomial assumption vs. the permutation test in terms of $P(TypeI)$ and $P(TypeII)$ as a function of the sample size and Figure 2 as a function of d. Different choices for the data generation parameters provided analogous results.

We tested the effect of using classifiers with different number of hyperparameters, whose values were selected through nested k-CV. Both in case of linear and non-linear optimal Bayes classifier, we observed that less hyperparameters lead to lower $P(TypeII)$ and slightly increased $P(TypeI)$. We do not show quantitative results for lack of space. Anyway, we obtained qualitative analogous results for a wide range of values of the parameters of the simulation.

Figure 3 shows the $P(TypeI)$ and $P(TypeII)$ as a function of k, the number of folds of k-CV. We observed such qualitative behavior for a wide range of choices of parameters of the simulations.

[1] We restricted the changes to the diagonal elements of Σ_B, $\{\sigma_B^i\}_i$, to limit the the number of parameters to explore in experiments.

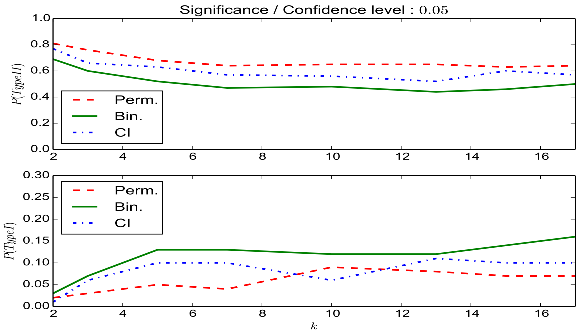

Fig. 3. Comparison of the three tests in terms of $P(TypeI)$ and $P(TypeII)$ as a function of the number of folds k of CV. Classifier: linear SVM, $N = 40$, $d = 200$, $d_{inf} = 10$, $\delta = 0.5$, $\lambda = 0$.

For full reproducibility of the results, all the code of this simulation study is publicly available with an OpenSource license at https://github.com/emanuele/prni2016_classification_test.

IV. Discussion and Conclusion

Our results on simulated data illustrate that the choices made during the inferential process have a clear influence on its efficacy, in terms of $P(TypeI)$ and $P(TypeII)$. Even though the simulation is not an accurate representation of neuroimaging data, we believe that the results can be interpreted in a qualitative way with respect to their impact on neuroimaging data analysis.

From a qualitative point of view, the results summarized in Figure 1, 2, 3 and in many other configurations of the simulations, show that the use of the binomial assumption lead to an increase to the $P(TypeI)$ with respect to the level defined at the beginning of the analysis, i.e. 0.05 for the hypothesis test (0.95 for the confidence level). As reported in [8], this is expected because the cross-validation scheme interferes with such assumption. Nevertheless, we observe a decrease of $P(TypeII)$, which was not reported before. This means that using the binomial assumption increases the sensitivity of the test procedure at the cost of false discovery. Moreover, using the binomial assumption within the approach based on confidence intervals result in similar $P(TypeI)$ and reduced $P(TypeII)$ with respect to the permutation-based approach.

The results of the comparison of classifiers with different number of hyperparameters match our expectations. A classifier with more hyperparameters exhibits an increase in $P(TypeII)$, that we motivate with the increased variance in the results due to need of fitting more hyperparameters with the same amount of data.

The results of Figure 3 shows that $P(TypeI)$ increases and $P(TypeII)$ decreases with the number k of folds of k-CV. A higher k means a larger train set in each fold, which leads to more stable classifiers, justifying the decrease of $P(TypeII)$.

As mentioned in Section III, we observed the qualitative findings above with many different settings of the parameters, not reported for lack of space. In particular we observed them with datasets of low dimension, typical of brain mapping procedures like searchlight [3]. So we are confident in extending the claims also to the low-dimension setting.

In conclusion, we believe that the results presented here fill some of the gaps in the literature of methods for neuroimaging data analysis. Moreover, the simulation-based approach proposed here can be used for further investigations on the efficacy of other data analysis strategies as well.

References

[1] N. K. Focke, G. Helms, S. Scheewe, P. M. Pantel, C. G. Bachmann, P. Dechent, J. Ebentheuer, A. Mohr, W. Paulus, and C. Trenkwalder, “Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic parkinson syndrome and healthy controls,” *Hum. Brain Mapp.*, vol. 32, no. 11, pp. 1905–1915, Nov. 2011. [Online]. Available: http://dx.doi.org/10.1002/hbm.21161

[2] J. V. Haxby, “Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex,” *Science*, vol. 293, no. 5539, pp. 2425–2430, Sep. 2001. [Online]. Available: http://dx.doi.org/10.1126/science.1063736

[3] N. Kriegeskorte, R. Goebel, and P. Bandettini, “Information-based functional brain mapping.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 10, pp. 3863–3868, Mar. 2006. [Online]. Available: http://dx.doi.org/10.1073/pnas.0600244103

[4] F. Pereira, T. Mitchell, and M. Botvinick, “Machine learning classifiers and fMRI: a tutorial overview.” *NeuroImage*, vol. 45, no. 1 Suppl, pp. 199–209, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2008.11.007

[5] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafo, “Power failure: why small sample size undermines the reliability of neuroscience,” *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 365–376, Apr. 2013. [Online]. Available: http://dx.doi.org/10.1038/nrn3475

[6] J. A. Etzel, V. Gazzola, and C. Keysers, “An introduction to anatomical ROI-based fMRI classification analysis,” *Brain Research*, vol. 1282, pp. 114–125, Jul. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.brainres.2009.05.090

[7] E. Maris and R. Oostenveld, “Nonparametric statistical testing of EEG- and MEG-data.” *Journal of neuroscience methods*, vol. 164, no. 1, pp. 177–190, Aug. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.jneumeth.2007.03.024

[8] Q. Noirhomme, D. Lesenfants, F. Gomez, A. Soddu, J. Schrouff, G. Garraux, A. Luxen, C. Phillips, and S. Laureys, “Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions,” *NeuroImage: Clinical*, vol. 4, pp. 687–694, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.nicl.2014.04.004

[9] J. A. Etzel and T. S. Braver, “MVPA Permutation Schemes: Permutation Testing in the Land of Cross-Validation,” in *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*. IEEE, Jun. 2013, pp. 140–143. [Online]. Available: http://dx.doi.org/10.1109/prni.2013.44

[10] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses (Springer Texts in Statistics)*. Springer, Nov. 2010. [Online]. Available: http://www.worldcat.org/isbn/1441931783

[11] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The Balanced Accuracy and Its Posterior Distribution,” in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, Aug. 2010, pp. 3121–3124. [Online]. Available: http://dx.doi.org/10.1109/icpr.2010.764

[12] E. Olivetti, S. Greiner, and P. Avesani, “Induction in Neuroscience with Classification: Issues and Solutions,” in *Machine Learning and Interpretation in Neuroimaging*, ser. Lecture Notes in Computer Science, G. Langs, I. Rish, M. Grosse-Wentrup, and B. Murphy, Eds. Springer Berlin Heidelberg, 2012, vol. 7263, pp. 42–50. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-34713-9_6

[13] E. Olivetti, A. Mognon, S. Greiner, and P. Avesani, “Brain decoding: Biases in error estimation,” in *Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), 2010 First Workshop on*, vol. 0. Los Alamitos, CA, USA: IEEE, Aug. 2010, pp. 40–43. [Online]. Available: http://dx.doi.org/10.1109/wbd.2010.9

[14] J. Langford, “Tutorial on Practical Prediction Theory for Classification,” 2005. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.4561