Bayesian Hypothesis Testing in Machine Learning

Giorgio Corani(✉), Alessio Benavoli,
Francesca Mangili, and Marco Zaffalon

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA),
USI - SUPSI, Manno, Switzerland
{giorgio,alessio,francesca,zaffalon}@idsia.ch

Abstract. Most hypothesis testing in machine learning is done using the frequentist null-hypothesis significance test, which has severe drawbacks. We review recent Bayesian tests which overcome the drawbacks of the frequentist ones.

Keywords: Bayesian hypothesis testing · Null hypothesis significance testing

1 Introduction

Hypothesis testing in machine learning (for instance to establish whether the performance of two algorithms is significantly different) is usually performed using frequentist tests.

The highly-cited tutorial by [1] makes some important points. It recommends non-parametric rather than parametric tests for comparing multiple classifiers on multiple data sets. The advantages of the non-parametric approach are that they do *not* average measures taken on different data sets; they do *not* assume normality; they are *robust* to outliers. In particular, [1] recommends the signed-rank test for the pairwise comparison of *two* classifier over multiple data sets and the Friedman test for the comparison of *multiple* classifiers over multiple data sets. Modern procedures for the multiple comparisons are discussed in [1,2]. They control the family-wise error rate (FWER) while providing more power than the traditional Bonferroni correction. Both [1,2] assume the post-hoc analysis of the Friedman test to be based on the mean-ranks test. When comparing two algorithms A and B, the statistic of the mean-ranks test is proportional to the difference between the average rank of A and B, $\overline{R}_A - \overline{R}_B$. In a recent note [3], we recommend instead to *avoid* the mean-ranks test, as both $\overline{R}_A$ and $\overline{R}_B$ depend on the performance of the other algorithms included in the original experiment. This can make the results non-repeatable. For instance the difference between A and B could turn out to be significant if the pool comprises algorithms C, D, E and not significant if the pool comprises algorithms F, G, H. We instead recommend using the sign-test or the Wilcoxon signed-rank test, whose outcome only depends on the performance of A and B.

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 199–202, 2015.
DOI: 10.1007/978-3-319-23461-8_13

However such tests are based on the frequentist framework of the null-hypothesis significance tests (NHST). The NHST controls the Type I error, namely the probability of rejecting the null hypothesis when it is true. When multiple comparisons are performed, the NHST approach prescribes to control the family-wise error rate, namely the probability of finding at least one Type I error among the null hypotheses which are rejected. Yet null hypothesis significance testing has severe drawbacks.

Consider analyzing a data set of n observations with a NHST test. The sampling distribution used to determine the critical value of the test assumes that your intention was to collect exactly n observations. If your intention was different (for instance in machine learning you typically compare two algorithms on all the data sets that are available) the sampling distribution should be changed to reflect your actual sampling intentions [4]. This is never done, given the difficulty of formalizing one's intention and of devising an appropriate sampling distribution. This problem is thus important but generally ignored.

NHST can reject the null hypothesis or fail to reject it, but it cannot verify the null hypothesis. In other words, it does not provide any measure of evidence *for* the null hypothesis. Within the NHST framework accepting the null hypothesis is a weak decision: it does not mean that the null hypothesis is true.

NHST decisions are taken on the basis of the chosen significance α, namely the probability of rejecting the null hypothesis when it is true. Usually one sets α=0.01 or 0.05, without having the possibility of a sound trade-off between Type I and Type II errors.

Bayesian hypothesis tests overcome these issues. The computation does not depend on the intention of the person who collected the data. The Bayesian test returns the posterior probability of the null and the alternative hypotheses. This allows to take decision which minimize the posterior expected value of the loss (posterior risk). For instance [5] reviews how to obtain Bayes-optimal decisions for a variety of different loss functions.

In [6] we proposed a Bayesian counterpart of the signed-rank test, which is the recommended test for comparing the score of two classifiers on multiple data sets. To devise this non-parametric test we adopted the Dirichlet process, which is often used in Bayesian non-parametrics. By means of simulations on artificial and real world data, we use our test to decide if a certain classifier is significantly better than another. The Bayesian and the frequentist signed-rank (α=0.05) take the same decisions only when we assume the Type I error to be 19 times more costly than the Type II error. In this case, the optimal decision is to declare that classifier Y is better than classifier X when the posterior probability of this hypothesis is greater than $1-\alpha = 0.95$. For any other different cost setting the frequentist test is tied to control the Type I error, fixing $\alpha = 0.05$. Instead the Bayesian decision rule allows to minimize the posterior risk. The rule for optimal decisions (accepting or rejecting the null hypothesis) is equivalent to that of cost-sensitive classification [7]. For any other setting of the costs, the Bayesian test incurs *lower* costs than the frequentist test.

Assume now that the two classifiers have been assessed via cross-validation on a collection of data sets $\boldsymbol{D} = \{D_1, D_2, \ldots, D_q\}$. One has to decide if the difference of accuracy between the two classifiers on the multiple data sets of $\boldsymbol{D}$ is significant. The signed-rank test both in its frequentist and Bayesian variant considers only the mean difference of accuracy measured on each data set, ignoring the associated uncertainty of the cross-validation estimates obtained on each data set.

In [8] we propose a test which performs inference on multiple data sets accounting for the correlation and the uncertainty of the estimates yielded by cross-validation on each data set. Our solution is based on two steps. First we develop a Bayesian counterpart of the correlated frequentist t-test [9], which is the standard test for analyzing cross-validation results. Under a specific matching prior the inferences of the Bayesian correlated t-test and of the frequentist correlated t-test are numerically equivalent. The meaning of the inferences is however different. The inference of the frequentist test is a p-value; the inference of the Bayesian test is a posterior probability. The posterior probabilities computed on the individual data sets can be combined to make further Bayesian inference on multiple data sets.

After having computed the posterior probabilities on each individual data set through the correlated Bayesian t-test, we merge them to make inference on $\boldsymbol{D}$. We model each data set as a Bernoulli trial (borrowing the intuition of [10]), whose possible outcomes are the win of the first or the second classifier. The probability of success of the Bernoulli trial corresponds to the posterior probability computed by the Bayesian correlated t-test on that data set. The number of data sets on which the first classifier is more accurate than the second is a random variable which follows a Poisson-binomial distribution. We use this distribution to make inference about the difference of accuracy of the two classifiers on $\boldsymbol{D}$. We are unaware of other approaches able to compare cross-validated classifiers on multiple data sets, accounting for the correlation and the uncertainty of the cross-validation estimates.

When comparing multiple classifiers, the recommended frequentist approach is the Friedman test. If it rejects the null hypothesis, one runs a procedure for multiple comparisons. A problem also of the modern procedures for multiple comparisons [1,2] is that they simplistically treat the multiple comparisons as independent from each other. But when comparing algorithms $\{a, b, c\}$, the outcome of the comparisons (a,b), (a,c), (b,c) are *not* independent.

In [11] we devised a Bayesian non-parametric procedure for comparing multiple classifiers. Adopting again the Dirichlet process (DP) [12] as a model for the prior, we first devised a Bayesian Friedman test. Then we designed a *joint* procedure for the analysis of the multiple comparisons which accounts for their dependencies. We analyze the posterior probability computed through the Dirichlet process, identifying statements of *joint* comparisons which have high posterior probability. The proposed procedure is a compromise between controlling the FWER and performing no correction of the significance level for the multiple comparisons. Our Bayesian procedure produces more Type I errors but fewer

Type II errors than procedures which control the family-wise error. In fact, it does not aim at controlling the family-wise error. We show the effectiveness of this approach in a simulation of sequential model selection among a large number of candidates (*racing*). Our procedure yields superior results compared to the traditional frequentist procedure thanks to both ability to manage dependencies among the multiple comparisons and to recognize equivalent models, narrowing down the pool of competing models. To recognize that the models have equivalent performance corresponds to verify the null hypothesis, which is impossible within NHST.

2 Software

The software for all our methods is available from http://ipg.idsia.ch/software/.

References

1. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**, 1–30 (2006)
2. Garcıa, S., Herrera, F.: An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. Journal of Machine Learning Research **9**, 2677–2694 (2008)
3. Benavoli, A., Corani, G., Mangili, F.: Should we really use post-hoc tests based on mean-ranks? Journal of Machine Learning Research (2015) (in press)
4. Kruschke, J.: Doing Bayesian data analysis: A tutorial introduction with R. Academic Press (2010)
5. Müller, P., Parmigiani, G., Robert, C., Rousseau, J.: Optimal sample size for multiple testing: the case of gene expression microarrays. Journal of the American Statistical Association **99**(468), 990–1001 (2004)
6. Benavoli, A., Mangili, F., Corani, G., Zaffalon, M., Ruggeri, F.: A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pp. 1026–1034 (2014)
7. Elkan, C.: The foundations of cost-sensitive learning. In: International Joint Conference on Artificial Intelligence, vol. 17, pp. 973–978 (2001)
8. Corani, G., Benavoli, A.: A Bayesian approach for comparing cross-validated algorithms on multiple data sets. Machine Learning (2015) (in press)
9. Nadeau, C., Bengio, Y.: Inference for the generalization error. Machine Learning **52**(3), 239–281 (2003)
10. Lacoste, A., Laviolette, F., Marchand, M.: Bayesian comparison of machine learning algorithms on single and multiple datasets. In: Proc. of the Fifteenth Int. Conf. on Artificial Intelligence and Statistics (AISTATS 2012), pp. 665–675 (2012)
11. Benavoli, A., Mangili, F., Corani, G., Zaffalon, M.: A Bayesian nonparametric procedure for comparing algorithms. In: Proceedings of the 32nd International Conference on Machine Learning (ICML 2015) (2015) (in press)
12. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. The Annals of Statistics **1**(2), 209–230 (1973)