The Empirical Variance Estimator for Computer Aided Diagnosis: Lessons for Algorithm Validation

Alex F. Mendelson[1], Maria A. Zuluaga[1], Lennart Thurfjell[2], Brian F. Hutton[3,4], and Sébastien Ourselin[1,5]

[1] Translational Imaging Group, Centre for Medical Image Computing, University College London, London, UK
[2] GE Healthcare, Uppsala, Sweden
[3] Institute of Nuclear Medicine, University College London, London, UK
[4] Centre for Medical Radiation Physics, University of Wollongong, NSW, Australia
[5] Dementia Research Centre, University College London, UK

Abstract. Computer aided diagnosis is an established field in medical image analysis; a great deal of effort goes into the development and refinement of pipelines to achieve greater performance. This improvement is dependent on reliable comparison, which is intimately related to variance estimation. For supervised methods, this can be confounded by statistical issues at the comparatively small sample sizes typical of the field. Given the importance of reliable comparison to pipeline development, this issue has received relatively little attention. As a solution, we advocate an empirical variance estimator based on validation within disjoint subsets of the available data. Using Alzheimer's disease classification in the ADNI dataset as an examplar, we investigate the behaviour of different variance estimators in a series of resampling experiments. We show that the proposed estimator is unbiased, and that it exceeds the estimates of naive approaches, which are biased down. Because the estimator avoids independence assumptions, it is able to accommodate arbitrary validation strategies and performance metrics. As it is unbiased, it is able to provide statistically convincing comparison and confidence intervals for algorithm performance. Finally, we show how the estimator can be used to compare different validation strategies, and make some recommendations about which should be used.

1 Introduction

There has been great hope for supervised methods in the characterization of neurological disorders, where they may be able to pick out subtle or distributed changes in images that may not be apparent to the radiologist. One area that has received particular attention is computer aided diagnosis (CAD) in Alzheimer's disease (AD) and its prodrome, mild cognitive impairment. Anticipated disease modifying interventions will need to occur at an early phase in the disease course, where gross cognitive and neurological changes may not be apparent.

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 236–243, 2014.

The defining feature of supervised learning is the use of a training set of labeled examples to build an explicit or an implicit model which is then applied to quantify or classify previously unseen examples. In order to avoid the upward bias of over-fitting, classification algorithms must be evaluated using separate test sets that were not used in classifier construction. While we speak here primarily of algorithms, our discussion is just as relevant to the comparison of whole CAD pipelines; the methods that compare the performance of algorithms can also be used for modality or feature choices.

Algorithm validation can be performed using simple hold out strategies or cross validation (CV). In the former, only a single classifier is trained and evaluated. The real quantity of interest is always the expected performance of the algorithm in this scenario, marginalized over all test subjects and training sets of a particular size. CV combines multiple hold out tests to reduce the variance of the performance estimate while remaining unbiased. The most common for of this, K-fold CV (KCV), divides the data into K disjoint subsets of equal size. Each of these is left out to become the test set while the remainder are used for training. In this way, each example is used exactly once for testing, and so all have equal weight in the final performance estimate.

One reason for the popularity of simple hold out tests is the presumed independence of performance results on each subject in the test data, allowing for binomial model confidence intervals and table based tests [1]. Using these tests for algorithms is fine provided the performance of their derived classifier models is constant enough across training sets. Much of machine learning deals with large datasets that make this assumption reasonable, but in medical image classification, where complex classifier models are built on necessarily limited data this seems less likely. Here, validation based on a single training set may be incorrect in generalising inference conclusions based on classifier models to their producing algorithms [2]. An inference test for algorithms based on simple hold out may then have a type I error rate above the nominal value (e.g. 0.05), which then loses its meaning.

While some of the variation due to training set will be apparent in the results of CV, the correlations between the evaluations on different folds makes unbiased variance estimation impossible with a function of the observed performances alone [3]. As the choice of partition is another source of variation, it is possible to provide more accurate estimates through the combination of M KCV runs performed on the same data (MKCV), though this further violates the independence assumptions that would be used to estimate variance. The effect of these correlations in testing against the null hypothesis that an algorithm performs no better than chance was demonstrated in a recent paper [4], highlighting the need for permutation tests. These problems are already acknowledged by much of the research community, and classification papers often omit the statistical analysis common in group difference studies.

As there is no appropriate permutation test for the null hypothesis that two algorithms have the same performance, comparative inference must rely on distribution assumptions and variance estimation. Though naive variance models

assuming independence may have bias either way, the results are most harmful when variance is underestimated in comparison studies. When the probability of incorrectly rejecting the null hypothesis may be above the nominal level, statistical tests are undermined. Reliable or conservative estimates for the variance are just as crucial in algorithm comparison as estimates for the expectation.

In response to this problem, we advocate an empirical solution based on the work of [5], which involves performing validation in two disjoint subsets of the available data. We conduct a series of resampling experiments to examine the behaviour of different variance estimators in a typical classification context. We are able to demonstrate the failure of naive variance estimators built on independence assumptions, and to show that the expectation of the proposed measure is unaffected by finite sample sizes. Given its necessary convergence, this strongly indicates that it is an unbiased estimator for the true variance. With this knowledge, we show how it can be used to investigate the stability of different validation strategies, and give some advice as to which should be used.

2 Variance Estimation Using Disjoint Subset Pairs

Though there exist model based estimators that rest on much weaker assumptions than independence, these are not straightforward. In particular, we consider the estimator developed in [6] for KCV. This has an adjustable parameter to trade between power and error rate. While it would be possible to choose a parameter which made tests conservative beyond all reasonable doubt, this would come with severely diminished power. We prefer empirical estimation because it is intuitive and general. It can be used with arbitrary validation strategies and performance metrics (e.g. MKCV and AUC), and has no parameter selection requiring prior knowledge of a problem. Though it comes at an increased computational cost, it is not an unreasonable one. The bulk of computation load in most medical image classification studies lies in the image processing and registration steps rather than the production and testing of classifiers.

In the absence of a model, we cannot estimate the variance of a validation strategy's performance estimate from a single observation. By conducting a validation strategy on two randomly sampled disjoint subsets of the same size one is able to produce two independent observations, $x^{(1)}$ and $x^{(2)}$. It is then possible to get an unbiased variance estimate $\mathcal{V}$ using the standard variance estimator for a sample of two. To reduce variability due to choice of subsets, one can repeat the selection R times and combine results in an average.

$$\mathcal{V} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_r^{(i)} - \bar{x}_r)^2\right), \text{ where } n = 2 \tag{1}$$

Where there are N examples available, this variance estimate can then be used in one of two comparison schemes:

1. Use $\mathcal{V}$ as a conservative estimate for the variance of CV performance on the full N. This requires only the modest assumption that it becomes more stable at greater sample sizes [5].

2. Use $\mathcal{V}$ as an unbiased estimate for CV performance on $\frac{1}{2}N$. $\frac{1}{2}\mathcal{V}$ is then an unbiased variance estimator for the mean of two of these estimates.

Though the main purpose of our estimator is to provide a conservative estimator, there are other good reasons for its use. One of these is the accommodation of arbitrary performance metrics. Another is the ability to account for variance reduction due to repeated partitioning. There are also situations where anti-correlation between performance in the folds of CV leads naive estimators to underestimate the variance [6], when this occurs, the proposed estimator may actually produce narrower confidence intervals than naive models. Generally, we suspect that inference with the second method will be more powerful, though researchers must be mindful that the reduced training set size involved may alter algorithms comparative performance.

3 Resampling Experiments

Our aim here is to reproduce a typical CAD task in neuroimaging. We present a commonly used algorithm, dataset, and feature set. MRI and assessment data were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database. In all experiments, we classified AD patients and healthy control (HC) subjects using T1 weighted images acquired at baseline on a mixture of 1.5T and 3.0T scanners. All images were automatically corrected for spatial distortion and subjected to quality control. Any subjects with suspected alternative etiologies, including fronto-temporal dementia, were excluded. This left us with a total of 370 HC subjects, and 252 AD subjects.

In order to generate features for classification we used label propagation to parcellate our images with an 83 region atlas[1] [7,8]. Non-rigid registration was performed using the niftyreg package[2] [9]. We performed a six-class tissue segmentation using the new segment module of SPM12[3] with the maximum cleanup option, and used the resulting tissue compartments to segment the intra-cranial space. Our feature set was then the sum of the grey matter compartment in each of the cerebral regions, normalized by the intra-cranial volume.

In all experiments presented here, we used balanced CV implemented exactly as in the commonly used LIBSVM package [10]. We used a linear support vector machine (SVM) with the C parameter selected so that the SVM behaviour was in the 'hard margin' limit, and increasing it further brought about no change in the solutions obtained. To avoid issues relating to class balance, we restrict our analysis to subsets with equal numbers of subjects from each class. In all our experiments we take a number of disjoint pairs of validation experiments, and used them to produce the following variance estimators:

– The proposed empirical variance estimate (EVE) using disjoint pairs (2. from section 2).

[1] `http://www.brain-development.org`
[2] `http://sourceforge.net/projects/niftyreg`
[3] `http://www.fil.ion.ucl.ac.uk/spm/software/`

- A naive empirical variance estimate (NEVE) that treats the performance on each randomly selected subset as independent. Subsets are randomly selected from all available data, and allowed to overlap.
- A naive fold-wise model (NFW) which treats the performance in each fold as an independent observation. We take the mean variance estimate across all subsets used.

3.1 Drawstring Experiments

By their definition, both empirical estimators for the variance of a validation experiment of a fixed size must converge to the true variance as the size of the available dataset increases. If the expectation of our estimator was affected by finite sample size, then it would not be an unbiased estimator of the true variance. To determine whether EVE was affected by a finite dataset, and compare its behaviour to NEVE, we performed a drawstring experiment as follows: we drew 500 "parent" subsets with a varying P subjects of each class, and within each of these produced both empirical estimates using 500 subset pairs. These smaller subsets had a fixed 25 subjects of each class, and on each of these we performed 10-fold CV. P was varied from 25 to 155 in increments of two.

3.2 Variance Characterization

In this experiment we sought to compare the behaviours of the empirical estimators to the naive model-based ones at a variety of sample sizes. We produced the three variance estimates using 10000 paired subsets for 10-fold CV. We varied the size of the subset from 10 to 125 items of each class. In addition to the expectation of our variance estimators, we include a variance estimate from a naive binomial model which assumes results on each subject are independent and binomial distributed, i.e., where the mean performance is p on n items, the variance estimate is $\frac{p(1-p)}{n}$.

3.3 Validation Strategy Comparison

Different CV strategies will have different variances depending on the character of the classification problem, which is not known in advance. Having an empirical estimator allows one to compare these variances for real problems using finite datasets. Having fewer folds decreases training set size, but allows split variability to be reduced through repeated partitioning. The choice of the best strategy to estimate performance may then be a trade-off between bias and variance. As a demonstration, we used the EVE to compare 5 strategies as estimators for the expected leave one out performance. These were 10 and 3 fold CV with either 1 or 10 repeats, and leave one out itself. To this end, we produced 2500 disjoint subset pairs, and measured the bias, variance, and the mean squared error of each strategy based on assumptions of normality.

4 Results and Discussion

4.1 Drawstring Experiments

In Fig. 1, we can see that the expectation of NEVE is clearly affected by finite sample size, while the proposed estimator stays constant. As both necessarily converge to the true variance as parent sample size increases, this provides convincing evidence that the proposed method is an unbiased estimator even at finite sample size. While we report accuracy here as it is the most commonly used performance metric, the same qualitative behaviour was observed for sensitivity, specificity, and area under curve. The EVE value is omitted where there are too few subjects in the parent sample for two disjoint pairs.

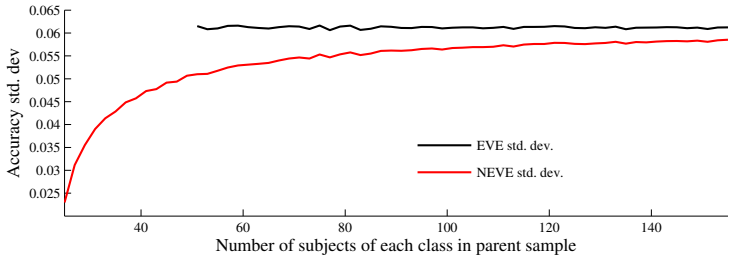

Fig. 1. Expectation of empirical variance estimators with changing sample size

4.2 Characterization Experiments

On the left side of Fig. 2, we can see the expected accuracy increase with sample size, while variance decreases. The naive empirical estimator acquires an increasing negative bias as the subsets used in validation become an increasing fraction of our full dataset. On the right, using a naive binomial variance estimate based on the mean performance to normalize, we can see relative sizes of the variance estimators. Both the naive binomial and fold-wise estimators are consistently below the proposed estimator. This is consistent with positive correlations between performance evaluations in our validation experiments. Here, we have demonstrated that models based on independence can underestimate variance in real CAD problems. We caution against their general use in hypothesis tests where exactness is required.

4.3 Comparison of Validation Strategies

Fig. 3 shows the relative variance of the different validation strategies as compared to leave one out, and their bias and mean squared error as an estimator for its expected performance. We can see that the 3-fold validations have a negative bias due to their smaller training sets, and that averaging over 10 partitions is

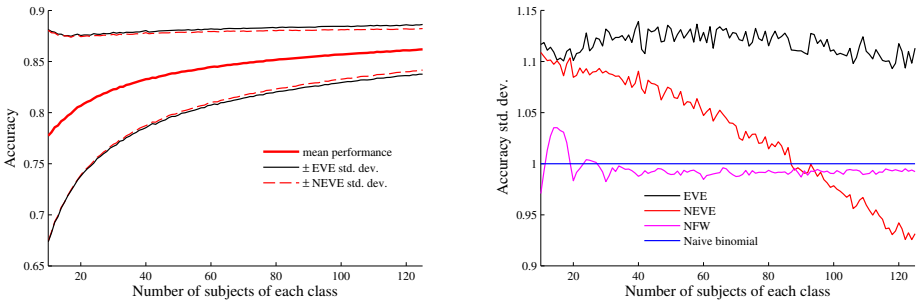

Fig. 2. Expectation of standard deviation estimates with changing validation set size. On the right, variance estimates are normalised by that of a naive binomial model.

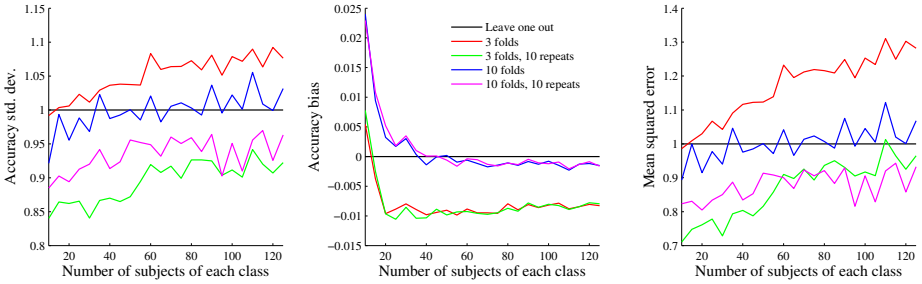

Fig. 3. Standard deviation, bias and mean squared error of validation strategies as estimator for the expectation of a leave one out experiment. Estimates are normalised by those of the leave one out validation itself.

able to reduce the variance. Notably, we can see that the repeated strategies have the lowest mean squared errors, particularly at smaller sample sizes. If one is willing to trade increased bias for decreased variance, leave one out is not the best estimator for its own expected performance.

5 Conclusion

This paper seeks to draw attention to the problem of performance variance estimation in CAD research. Having demonstrated that some naive estimators currently used can have negative bias, we strongly recommend the use of alternative methods in algorithm comparison. We demonstrate an empirical estimator as a possible solution to this problem, and show experimentally that it is unbiased. We then use it to compare different validation strategies as estimators, and find that a bias variance trade-off is possible using few folds with repeated partitioning. Because of its high variance, we recommend against the use of leave one out validation with all but exceptionally small sample sizes.

Acknowledgements. This work is funded by UCL (code ELCX), a CASE studentship with the EPSRC and GE healthcare, EPSRC grants (EP/H046410/1, EP/ H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH-DARE@IT (FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative). Imaging data were provided by the Alzheimer's disease neuroimaging initiative (http://adni.loni.ucla.edu/).

References

1. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O.: Automatic classification of patients with alzheimer's disease from structural MRI: a comparison of ten methods using the adni database. Neuroimage 56(2), 766–781 (2011)
2. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10(7), 1895–1923 (1998)
3. Bengio, Y., Grandvalet, Y.: No unbiased estimator of the variance of k-fold cross-validation. J. Mach. Learn. Res. 5, 1089–1105 (2004)
4. Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S.: Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. NeuroImage: Clinical 4, 687–694 (2014)
5. Nadeau, C., Bengio, Y.: Inference for the generalization error. Machine Learning 52(3), 239–281 (2003)
6. Grandvalet, Y., Bengio, Y.: Hypothesis testing for cross-validation. Montreal Universite de Montreal, Operationnelle DdIeR (2006)
7. Gousias, I.S., Rueckert, D., Heckemann, R.A., Dyet, L.E., Boardman, J.P., Edwards, A.D., Hammers, A.: Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. NeuroImage 40(2), 672–684 (2008)
8. Cardoso, M.J., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S.: STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Medical Image Analysis 17(6), 671–684 (2013)
9. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine 98(3), 278–284 (2010)
10. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`