Pattern Recognition Letters 29 (2008) 1960–1965

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Cross-validation and bootstrapping are unreliable in small sample classification

A. Isaksson [a,*], M. Wallman [a,c], H. Göransson [a], M.G. Gustafsson [a,b,*]

[a] Department of Medical Sciences, Uppsala University, Academic Hospital, SE-751 85 Uppasala, Sweden
[b] Department of Engineering Sciences, Uppsala University, P.O. Box 534, SE-751 21 Uppsala, Sweden
[c] Fraunhofer Chalmers Research Centre for Industrial Mathematics, Gothenburg, Sweden

ARTICLE INFO

Article history:
Received 22 October 2007
Received in revised form 11 June 2008
Available online 15 July 2008

Communicated by R.P.W. Duin

Keywords:
Supervised classification
Performance estimation
Confidence interval

ABSTRACT

The interest in statistical classification for critical applications such as diagnoses of patient samples based on supervised learning is rapidly growing. To gain acceptance in applications where the subsequent decisions have serious consequences, e.g. choice of cancer therapy, any such decision support system must come with a reliable performance estimate. Tailored for small sample problems, cross-validation (CV) and bootstrapping (BTS) have been the most commonly used methods to determine such estimates in virtually all branches of science for the last 20 years. Here, we address the often overlooked fact that the uncertainty in a point estimate obtained with CV and BTS is unknown and quite large for small sample classification problems encountered in biomedical applications and elsewhere. To avoid this fundamental problem of employing CV and BTS, until improved alternatives have been established, we suggest that the final classification performance always should be reported in the form of a Bayesian confidence interval obtained from a simple holdout test or using some other method that yields conservative measures of the uncertainty.

1. Introduction

Currently there is a rapidly growing interest to use supervised statistical learning techniques (Michell, 1997; Hastie et al., 2001; Webb, 2002) to design classifiers for different forms of decision support in performance sensitive applications found e.g. in biomedicine. Important examples are predictions of tumour subtype and clinical outcome based on mRNA levels in tumour samples measured using modern large-scale microarray technologies (Rosenwald et al., 2002; van't Veer et al., 2002; Yeoh et al., 2002). In these examples the classification of a new tumour sample is intended to guide the choice of treatment. Misclassification would lead to sub-standard treatment that may have serious consequences for the patient. Therefore, reliable information about the classifier performance is critical for the acceptance of classification guided therapy.

Since the 1980s, cross-validation (CV) (Lachenbrusch and Mickey, 1968; Stone, 1974) and bootstrapping (BTS) (Efron, 1979, 1983; Efron and Tibshirani, 1993) have been the dominating methods for estimation of the unknown performance of a classifier designed for discrimination. CV and BTS are acknowledged for making efficient use of the samples available, a feature which makes them especially suitable for use in situations where only a limited number of examples are available. To the present date, CV and BTS still are the most commonly used methods, even though many theoretical contributions and practical alternatives have been reported (Hand, 1986, 1997; Vapnik, 1995; Schiavo and Hand, 2000). The importance of reliable performance estimation when using small data sets must not be underestimated. No matter how sophisticated and powerful algorithms for classification are developed and applied, if no reliable performance estimates are obtained, no reliable decisions can be made based on classification results.

There are several ways to interpret the quantity estimated by CV and BTS. In this work, we choose the most common approach which is to consider CV and BTS as estimators of the true error rate e_{true}, i.e. the error rate of a designed classifier when applied to a very large independent test set. In other words, the true error rate is the probability of misclassification when applying the classifier to an unknown future test example, which is the quantity of interest for a decision maker. Here, we ignore the additional complexity that the size of the training sets in CV are slightly smaller than the total number of training examples available, a fact that introduces a bias in the performance estimates. CV and BTS are methods designed to estimate the classifier performance using the samples available via partitioning (CV) or resampling with replacement (BTS). An almost trivial, but often overlooked, fact is that a single point estimate, e_{est}, of a true performance, e_{true}, is not useful unless it is known to be close to e_{true}. For simulated data sets describing the same problem, a good estimator is reflected by a narrow joint

* Corresponding authors. Address: Department of Medical Sciences, Uppsala University, Academic Hospital, SE-751 85 Uppasala, Sweden. Tel.: +46 18 611 97 82; fax: +46 18 611 37 03.
E-mail addresses: Anders.Isaksson@medsci.uu.se (A. Isaksson), Mats.Gustafsson@angstrom.uu.se (M.G. Gustafsson).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.06.018

distribution (probability density function, pdf) $p(e_{true}, e_{est})$ of the estimate and the corresponding true value. For real problems there is only a single data set available, and the true performance is not known. Thus, $p(e_{true}, e_{est})$ cannot be estimated in a real application but it is important to understand its properties based on simulated data.

A given combination of classification problem, classification and learning algorithm, performance estimator, and data set size, defines an unknown joint distribution of estimated and true performance, $p(e_{true}, e_{est})$. In Fig. 1a one example of such a joint distribution is shown. For a given e_{est} the uncertainty about e_{true} is described by the conditional distribution $p(e_{true}|e_{est})$ which is illustrated in Fig. 1b. Notably, in a real application we would only have access to a single error rate point estimate e_{est} and we would not know the associated conditional distribution $p(e_{true}|e_{est})$ of interest. In this work, we address this fact by first illustrating the uncertainty in the CV and BTS estimates for realistic situations and then by suggesting the use of Bayesian confidence intervals or other conservative alternatives to assess and report classification performance.

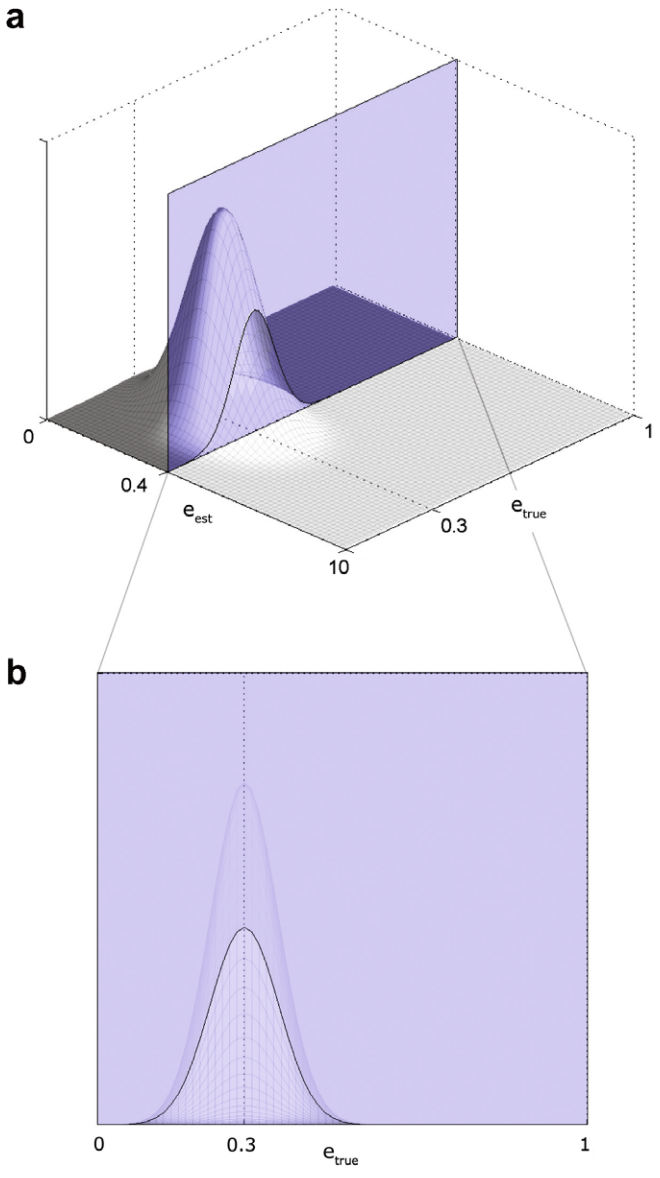

Fig. 1. (a) Schematic figure illustrating the joint distribution of true and estimated error rates $p(e_{true}, e_{est})$ for a particular combination of classification problem, classification/learning algorithm, and data set size obtained from simulated data sets by using a very large test set and one estimation method like CV or BTS. (b) Given an estimated error rate e_{est} of 0.4 the uncertainty about e_{true} is described by the conditional pdf $p(e_{true}|e_{est} = 0.4_t)$. This pdf is illustrated as the intersection $p(e_{true}, e_{est} = 0.4)$, a slice of the joint pdf which is identical to $p(e_{true}|e_{est} = 0.4)$ except for a normalization factor. Since the intersection $p(e_{true}, e_{est} = 0.4)$, and consequently the corresponding conditional pdf $p(e_{true}|e_{est} = 0.4_t)$, are centred at $e_{true} = 0.3$ this reflects a situation where, on average, the true performance is 0.3 when the estimated performance is 0.4. Notably, in a real application only the point estimate e_{est} would be available. Thus, there would be no information on the uncertainty of the underlying true error rate for the classifier.

2. Uncertainty of CV and BTS estimates

To illustrate the performance of CV and BTS estimates compared to true error rates when the methods are applied to problems of realistic size, we performed Monte Carlo simulation two-class classification experiments in Matlab™ (Mathworks Inc., USA), described in detail elsewhere (Wickenberg Bolin et al., 2006), using a pair of two-dimensional overlapping Gaussian distributions with means, variances and co-variances estimated from real mRNA expression data From the simulation results we determined $p(e_{true}, e_{est})$ for data sets of different sizes. All simulations were based on our own code together with routines available in version 4.0 of PRTools (Duin et al., 2004), a toolbox (plugin) for Matlab™ specialized on statistical pattern recognition. In particular, a PRTools function called parzendc was employed for classifier design. This design procedure is based on a Parzen kernel density estimate (Parzen, 1962; Webb, 2002) with its smoothing kernel parameter obtained as a maximum likelihood estimate. However, this particular choice is not important for the main conclusions made here.

The different algorithms for performance estimation applied were the 0.632 bootstrap estimator and a hybrid algorithm called repeated 10-fold CV (rCV10) which performs conventional 10-fold CV repeatedly using different partitionings of the data set (Hastie et al., 2001). One thousand different data sets were generated and for each classifier designed, the true performance was obtained by testing it using a large test set consisting of 5000 samples per class. The results were similar for both performance estimation algorithms. In Fig. 2, the simulation results from evaluating the 0.632 bootstrap estimator is presented for three different data set sizes ($N = 20$, 100, and 1000). Similar results are shown in Fig. 3 for a 30 times repeated 10-fold CV. For any chosen estimated error rate the broad conditional distributions show that the BTS and CV estimates are poor approximations of the true performance for both $N = 20$ and $N = 100$ samples. Only for $N = 1000$, the estimates deviate less than 10% from the true performance.

3. Performance evaluation and data set size

For a given classification problem, like the one studied in our simulations, an ideal situation would be if the true performance e_{true} was equal to the performance estimate e_{est} for all randomly selected data sets. In this situation, the joint distribution $p(e_{true}, e_{est})$ illustrated in Fig. 1a would be completely restricted to the straight line $e_{true} = e_{est}$ and there would be no uncertainty left about the true error rate. In both Figs. 2 and 3 the presented distributions of results from simulated data suggest that the underlying joint pdfs are broad and not at all restricted to the line $e_{true} = e_{est}$. Since a broad joint distribution results in a broad conditional distribution (as in Fig. 1a and b), the results from the simulations shown in Figs. 2a and b, 3a and b indicate that for small sample sizes the conditional distributions are so broad that observed performance estimates e_{est} become practically useless for prediction of the true performance e_{true}.

In Figs. 2a and 3a it can be noted that there is a larger variance in the e_{est}-axis direction compared to that along the e_{true}-axis. This

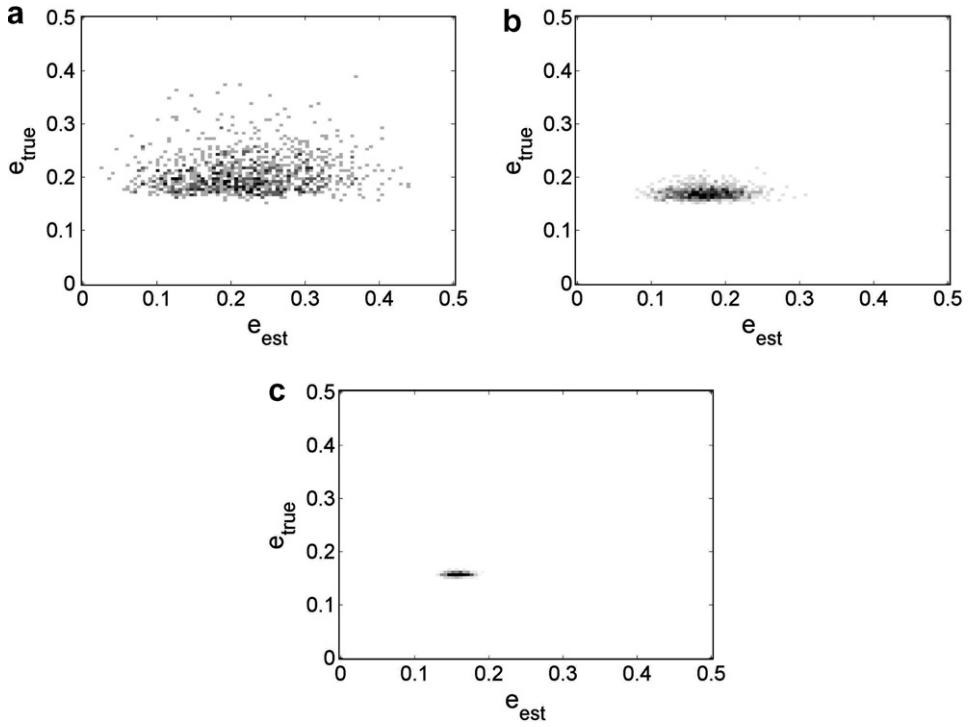

Fig. 2. Results from 0.632-bootstrap performance estimation of a Parzen kernel density based classifier applied to a standard two-class problem. 1000 data sets were simulated for each of three different sample sizes N. (a) $N = 20$. (b) $N = 100$. (c) $N = 1000$. Note the wide joint distributions that become progressively wider with smaller data set sizes.

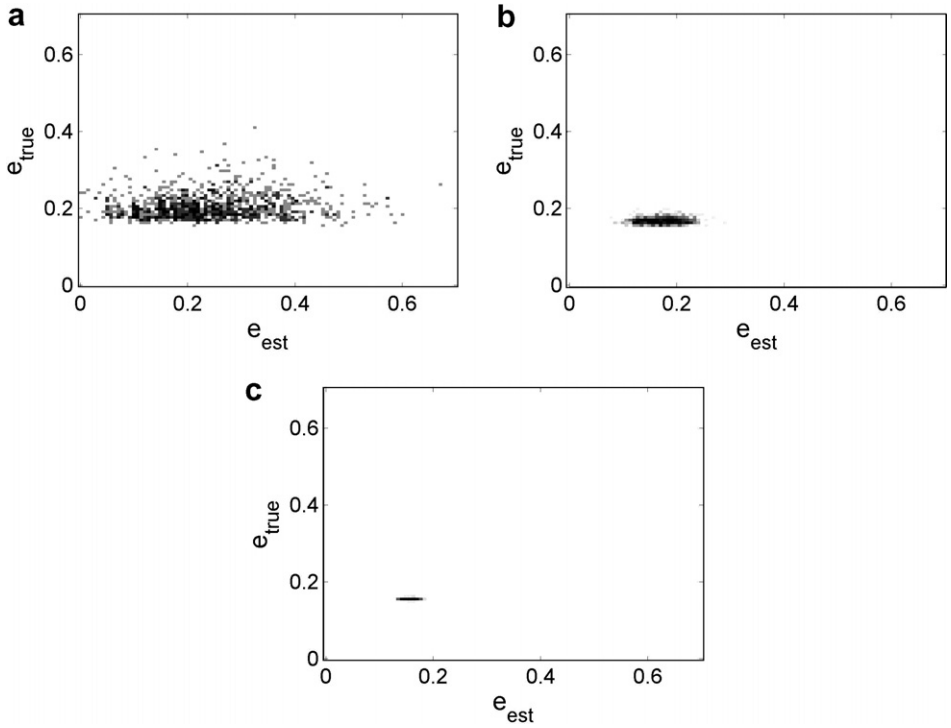

Fig. 3. Results from 30 times repeated 10-fold cross-validation estimation of a Parzen kernel density based classifier with 1000 simulated data sets for each of 3 different sample sizes N. (a) $N = 20$. (b) $N = 100$. (c) $N = 1000$. Note the wide joint distributions that become progressively wider with smaller data set sizes.

difference may be partly explained as a bias caused by the small test set size used to obtain the individual performance estimates along the e_{est}-axis. It will disappear as the number of examples used increases, see Figs. 2b and c, 3b and c. There is at present no detailed quantitative description of this phenomenon. However, issues contributing to this effect are qualitatively explained in the next section. For the less complicated holdout estimate ($e_{est} = k_t/N_t$ where k_t is the number of errors and N_t the number of inde-

pendent test samples) Wickenberg Bolin et al. has quantitatively described how its variance depends on the size of the test sets (Wickenberg Bolin et al., 2006).

For larger data sets sizes, here illustrated by $N = 1000$ (Figs. 2c and 3c), the joint distributions and consequently the conditional distributions become more narrow. For an unbiased performance estimator the joint pdf becomes restricted to a small region close to the line $e_{true} = e_{est}$. Consequently, as should be expected, unbiased versions of CV and BTS estimates become reasonable approximations of the true performance when the data set size is sufficiently large. The results from the simulations presented here are consistent with other recent and more comprehensive simulation results (Braga-Neto and Dougherty, 2004; Xu et al., 2006).

4. Estimating the uncertainty of CV and BTS estimates

Intuitively, the uncertainty of a particular CV estimate could be estimated as the variation between individual holdout estimates used to compute the CV estimate or as the variation between individual CV estimates obtained for different splits of the data set. If the variation between the individual estimates is small, this would indicate that the CV estimate is close to the true performance of a final design based on all examples available. Several methods based on this idea have been reported recently (Mukherjee et al., 2003; McLachlan et al., 2004; Michiels et al., 2005). Unfortunately, there are many pitfalls associated with this idea that must not be ignored.

For example, Michiels et al. demonstrated that CV estimates based on different partitions of the data may vary substantially in real medical applications (Michiels et al., 2005). Their results show that the CV estimates are unreliable but this does not at all prove that there is large variation in the true performances of the classifiers designed and tested. Due to the small test sets used in conventional CVs, the observed variance between the CV estimates might be completely dominated by the variance contribution from the small test set size used. This dominance can be understood e.g. from a recently derived analytical expression for the observed variance for an idealized CV procedure where all design and test sets are independent, here called repeated holdout testing (RHT) (Wickenberg Bolin et al., 2006). Therefore, one must not misinterpret large variations between CV estimates as in the paper by Michiels et al. to mean large variation also between the underlying true performances. Of course, a real large variation between underlying true performances will result in large variation in the corresponding CV estimates even if the test set size used is large. However, in all applications where the total number of samples is on the order of hundreds, theory and simulations suggest that the contribution to the observed variance from the small test set sizes still is substantial (Wickenberg Bolin et al., 2006). In conclusion, it might be quite misleading to interpret the variation between CV estimates to be valid also for the variation between true performances.

In addition to the variance caused by small test sets, there are at least two additional effects on the variance of CV estimates that must not be neglected. One is caused by the fact that in a CV procedure, all the classifiers are dependent due to the more or less overlapping design sets used. For example, in the extreme case of N-fold CV (also known as leave-one-out CV) where each test set consist of a single example, $N-2$ of the $N-1$ design examples used for any pair of the N classifiers designed and tested are identical. This large overlap makes the variation in the true performance e_{true} of the designed classifiers smaller than one should expect from a corresponding RHT. Another undesirable effect is caused by the fact that the CV estimates are contaminated by inter-dependencies between the design and test sets used. Since the test set in each of the CV iterations is directly given from the selection of the corresponding design set, the resulting holdout estimate is not completely independent of the design set as in an ideal RHT procedure. This deviation from the RHT yields a complicated variance contribution that is difficult to analyse in detail mathematically.

It should be pointed out that the variation in the e_{est}-axis direction in Fig. 2 is based on truly independent data. Therefore, it is fundamentally different from the variation in CV estimates obtained from different partitions of a single data set as discussed in this section. Thus the variation in Fig. 2 along the e_{est}-axis direction is the true variation in the CV estimates (that never is available in a real application) whereas in this section we consider estimates of this variation. However, the larger variance in the e_{est}-axis direction compared to that along the e_{true}-axis in Fig. 2 can be qualitatively explained by the fact that the CV estimates presented are based on small test sets.

As an attempt to compensate for the over-pessimistic effect caused by small test set sizes (the first effect) and to eliminate the problem of dependent design and test sets (the third effect), a method called Repeated Independent Design and Test (RIDT) has recently been developed that can yield almost unbiased estimates of the variation between the true performances of the classifiers designed and tested (Wickenberg Bolin et al., 2006). In other words, the RIDT procedure can yield an estimate of the variation between the true performances that equals the true variation on average. This is significantly different from the earlier approaches like the one proposed by Michiels et al. which yield an estimate of the variation between CV estimates, a quantity that is heavily contaminated by the small test set size effect. Thus, RIDT is designed to determine the expected variation between the true performances, not the variation between the CV estimates. Although the RIDT procedure is a significant step forward towards bounds on the true performance of a final classifier design, it is important to note that an estimate is of limited practical value unless it comes with information about its own uncertainty. However, since the uncertainties in the mean and variance estimates obtained with RIDT are unknown, RIDT cannot deliver the desired confidence interval. In conclusion, even with procedures like RIDT that can eliminate most of the undesirable small sample effects, it is not possible to determine reliable bounds (confidence intervals) on the true performance of interest.

To determine the uncertainty in a BTS estimate, an intuitive idea is to obtain a distribution of BTS estimates via a second external bootstrap that generates different data sets that subsequently are used to obtain individual BTS estimates. In analogy with the discussion above for the variation between CV estimates for different splits of the data set, the observed variation between the BTS estimates could then be used as a measure of uncertainty for the BTS estimates. Notably, such an approach would be completely different from the double bootstrap proposed early by Efron to reduce the bias of an individual BTS estimate (Efron, 1983). Unfortunately, this alternative double bootstrap suggested here would suffer from the same kind of small sample effects as discussed for the CV estimates above. In conclusion, neither CV/RIDT nor double bootstrap variance estimates can deliver reliable performance bounds on the true performance as long as the data set sizes are small (on the order of hundreds).

Attempts to provide information about the uncertainty of CV or BTS point estimates have also been made by performing permutation tests, for recent examples in medicine see (Hedenfalk et al., 2001; Radmacher et al., 2002). In each iteration, the class labels of the training data are permuted randomly and the resulting CV or BTS point estimate is obtained. Then the probability to observe a CV or BTS point performance estimate that is equal or better than the one obtained with the correct class labels is calculated. A low probability suggests that the classifier detects a real difference

between the classes. However, it should be noted that a low probability (p-value) does not provide a meaningful estimate of the uncertainty associated with the CV or BTS estimates. In fact, it has been shown in simulations that the p-values essentially are deterministic non-invertible functions of the corresponding CV estimates (Hsing et al., 2003). This means that the p-values are less informative than the CV estimates themselves in quantifying the associated uncertainty. Thus, permutation tests cannot be used to put trust in CV or BTS point estimates as being useful estimates of the true error rate.

In conclusion, CV and BTS can be used to estimate classification performance when the data set is large enough. However, there is presently no way of ensuring the uncertainty of the estimates for many real applications with sample sizes in the hundreds. This underscores the need for alternatives to CV and BTS estimates that can offer reliable confidence intervals and/or bounds on the true performance of interest.

5. Alternatives to CV and BTS

There are a number of recently proposed approaches to obtain confidence intervals for the performance of supervised classifiers based on a wide range of different theoretical foundations. For an excellent introduction see the tutorial by Langford (2005). There are few studies reported that compare the performance of different confidence intervals but recently the empirical applicability of five different upper bounds was compared on 29 real world data sets (Kaariainen and Langford, 2005). One of the most successful bounds was the binomial test bound recently proposed by Langford which is based on the binomial distribution. Another comparison of different methods to obtain confidence intervals (CIs) for the performance is the study by Brown et al. (2001).

A general conclusion is that many of the methods are not conservative in the sense that the resulting confidence intervals do not cover the true performance in the desired percentage of independent experiments, for example 95%. One explanation is that many methods rely on particular assumptions or approximations that are not valid for small sample sizes. Moreover, many bounds reported are not tight enough for practical use when the sample sizes are small. Therefore, the classical Bayesian CI (Jaynes, 1976, 2003; Webb, 2002) is of particular interest here as it yields conservative intervals and has already been employed in small sample medical applications such as tumour classification (Simon et al., 2003). The Bayesian posterior $p(e_{\text{true}}|k_t,N_t)$, which defines our uncertainty about the true error rate after observing k_t errors in N_t tests, is expressed by means of Bayes' rule as

$$p(e_{\text{true}}|k_tN_t) = \frac{P(k_t|e_{\text{true}}N_t)p(e_{\text{true}|N_t})}{P(k_t|N_t)} = \frac{P(k_t|e_{\text{true}}N_t)p(e_{\text{true}})}{P(k_t|N_t)} \quad (1)$$

where the last equality follows from the fact $p(e_{\text{true}}|N_t) = p(e_{\text{true}})$, i.e. that the integer N_t does not come with any information with respect to our prior knowledge about the true error rate. The factor $P(k_t|e_{\text{true}},N_t)$ defines the binomial distribution, i.e. the probability of observing k_t errors in N_t independent trials when the probability of making an error in each trial is e_{true}. $P(k|N_t)$ is independent of e_{true} and can therefore be viewed as a normalization constant. Thus, Bayesian inference makes it possible to calculate a posterior distribution $p(e_{\text{true}}|k_t,N_t)$ that describes our posterior uncertainty about the true error rate after making k_t errors on a data set with N_t test samples. Note that the posterior uncertainty is not conditioned on a particular data set but only on the fact that there were k_t errors made on N_t test examples. However, the particular classifier is defined by the samples used for training. A Bayesian CI at the 95%-level covers 95% of the area under the posterior $p(e_{\text{true}}|k_t,N(_t)$. For the small sample problems of interest here, unless the prior $p(e_{\text{true}})$ is narrow, the resulting posterior and the associated confidence intervals become broad.

Particularly attractive features of the Bayesian interval are that it is easily derived (see above) and that it relies on explicit use of a prior which quantifies the initial knowledge about the true performance. By contrast, alternative intervals like the Clopper–Pearson interval discussed by Brown et al. (2001) may also be employed but they are known to be more conservative than necessary (on average), they are not as easy to derive, and they do not allow simple incorporation of prior knowledge. An interesting connection between Bayesian inference and frequency based classical statistical inference in this context should be mentioned. The most informative prior possible is the true frequency distribution of true error rates that one would obtain when designing repeatedly with N_d samples. With access to this prior any α-level Bayesian CI of the posterior, defined by covering α% of the distribution, would be equivalent to the corresponding α-level confidence interval used in classical statistics. In practice the most informative prior is usually unknown and therefore replaced by a prior that reflects the

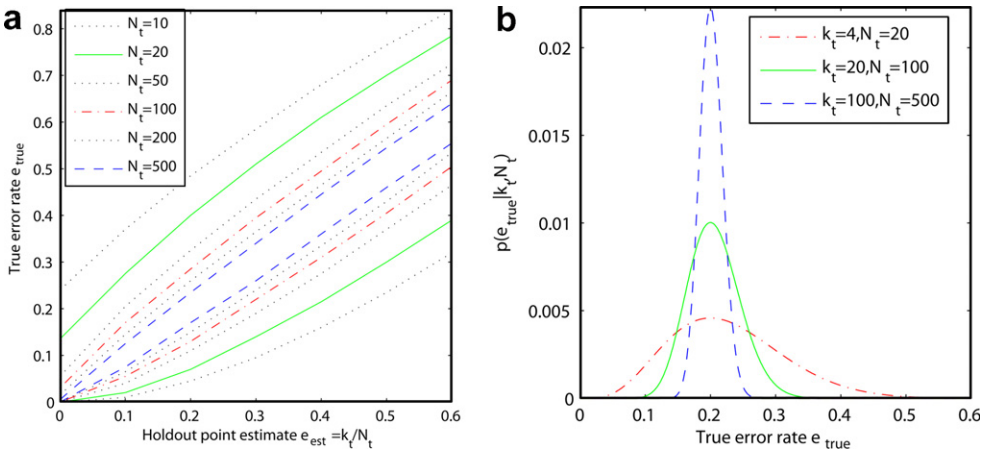

Fig. 4. (a) Three different Bayesian posterior distributions (pdfs) $P(e_{\text{true}}|k_t,N_t)$ for a uniform prior $p(e_{\text{true}}) = 1$ as calculated using Eq. (2). All distributions correspond to the same holdout point estimate $e_{\text{est}} = k_t/N_t = 20\%$ but the posteriors become progressively wider with smaller test sets. (b) Bayesian confidence intervals for different error rate estimates e_{est} and test set sizes N_t in a conventional holdout test. The intervals are the highest 95% probability density confidence regions, i.e. the shortest intervals that can be constructed from the HO results. As an example making $k_t = 4$ errors on $N_t = 20$ test examples yields a confidence interval for the true error rate as the interval between the green lines for $e_{\text{est}} = k_t/N_t = 20\%$ which is [7%,40%]. Similar graphs suitable for obtaining Bayesian confidence intervals can be found for example on page 254 in the text book by Webb (2002).

present state of knowledge. When the probability of misclassification is completely unknown, except that it is restricted to the unit interval [0,1], the uniform distribution, $p(e_{\text{true}}) = 1$ is used. In this case, the posterior may be written out explicitly as

$$p(e_{\text{true}}|k_t, N_t) = \beta e_{\text{true}}^{k_t}(1 - e_{\text{true}})^{N_t - k_t} \quad (2)$$

where is β normalization constant. Examples of this posterior for different values of k_t and N_t are presented in Fig. 4a. The Bayesian CI corresponding to Eq. (2) quantifies our uncertainty about the true performance. In terms of classical statistics, this Bayesian CI is exact in a situation where the distribution of true performances for the classifiers designed is uniform on the interval [0,1].

In Fig. 4b a graph is presented based on the Bayesian posterior distribution in Eq. (2) which makes it possible to determine the shortest Bayesian confidence interval for a particular holdout (HO) test result using a uniform prior. Thus, Fig. 4b provides a compact description of posterior pdfs in the form of intervals. For example, the solid green curve in Fig. 4a presents the whole posterior after making $k_t = 4$ errors using $N_t = 20$ test examples whereas the solid green curves in Fig. 4b show that the associated 95% confidence interval is [7%,40%]. A major conclusion is that for such small sample sizes, the Bayesian CIs provide rigorous intervals, which are necessary for using the classifier in a real life situation with any degree of certainty of its performance. If the interval is too wide for the particular application, the performance estimate is simply not good enough to permit the use of the classifier in practice.

One should remember that CV and BTS estimates become unnecessary in applications where the number of examples available is large. In such cases, a simple HO test is sufficient as the resulting Bayesian confidence interval will become very short. One should also note that a classifier may also be equipped with an option which will leave the sample of interest unclassified (Webb, 2002; Simon et al., 2003). To reject samples difficult to classify may be an attractive way to obtain reliable decisions, but this is not trivial as it requires reliable estimation also of the probability of rejection using the examples available.

6. Conclusion

In conclusion, for classification problems we would like to point out that, (1) For a single data set, there is no way of knowing how reliable a particular CV or BTS performance estimate is. (2) Simulations show that CV and BTS estimates are unreliable for sample sizes commonly encountered in real world applications. (3) Calculating a Bayesian confidence interval (or some alternative conservative confidence interval) based on a holdout test still seem to be the only rigorous yet practically useful alternative for assessing and reporting classifier performance.

Acknowledgements

This work was supported by the Wallenberg Consortium North, Cancerfonden, The Swedish Research Council (VR), The Swedish Society for Medical Research (SSMF), the Beijer foundation, the Göran Gustafsson foundation, Carl Tryggers stiftelse, the Magnus Bergvall foundation, the Marcus Borgström foundation, and the Faculty of Science and Technology (Uppsala University).

References

Braga-Neto, U.M., Dougherty, E.R., 2004. Is cross-validation valid for small-sample microarray classification? Bioinformatics 20 (3), 374–380.
Brown, L.D. et al., 2001. Interval estimation for a binomial proportion. Stat. Sci. 16 (2), 101–133.
Duin, R.P.W. et al., 2004. PRTools4. A Matlab Toolbox for Pattern Recognition. Delft University of Technology, Delft.
Efron, B., 1979. Bootstrap methods: Another look at the jackknife. Ann. Statist. 7, 1–26.
Efron, B., 1983. Estimating the error rate of a prediction rule: Improvements on cross-validation. J. Amer. Statist. Assoc. 78, 316–331.
Efron, B., Tibshirani, R., 1993. Introduction to the Bootstrap. Chapman and Hall, London.
Hand, D.J., 1986. Recent advances in error rate estimation. Pattern Recognition Lett. 4, 335–346.
Hand, D.J., 1997. Construction and Assessment of Classification Rules. John Wiley & Sons, Chichester, UK.
Hastie, T. et al., 2001. The Elements of Statistical Learning. Springer, New York.
Hedenfalk, I. et al., 2001. Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med. 344 (8), 539–548.
Hsing, T. et al., 2003. Relation between permutation-test P values and classifier error estimates. Mach. Learn. 52, 11–30.
Jaynes, E.T., 1976. Confidence Intervals vs Bayesian Intervals. In: Harper, W.L., Hooker, C.A. (Eds.), Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, vol. II. D. Reidel, Dordrecht, pp. 175–257.
Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK.
Kaariainen, M., Langford, J., 2005. A Comparison of Tight Generalization Error Bounds. Proceedings of the 22nd International Conference on Machine Learning. ACM, Bonn, Germany. pp. 409–416.
Lachenbrusch, P., Mickey, M., 1968. Estimation of error rates in discriminant analysis. Technometrics (10), 1–11.
Langford, J., 2005. Tutorial on practical prediction theory for classification. J. Mach. Learn. Res. 6, 273–306.
McLachlan, G.J. et al., 2004. Analyzing Microarray Gene Expression Data. Wiley, Hoboken, New Jersey.
Michell, T.M., 1997. Machine Learning. McGraw-Hill, New York.
Michiels, S. et al., 2005. Prediction of cancer outcome with microarrays: A multiple random validation strategy. Lancet 365 (9458), 488–492.
Mukherjee, S. et al., 2003. Estimating dataset size requirements for classifying DNA microarray data. J. Comput. Biol. 10 (2), 119–142.
Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Statist. 33 (3), 1065–1076.
Radmacher, M.D. et al., 2002. A paradigm for class prediction using gene expression profiles. J. Comput. Biol. 9 (3), 5005–5011.
Rosenwald, A. et al., 2002. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N. Engl. J. Med. 346 (25), 1937–1947.
Schiavo, R.A., Hand, D.J., 2000. Ten more years of error rate research. Int. Stat. Rev. 68 (3), 295–310.
Simon, R. et al., 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J. Natl. Cancer Inst. 95, 14–18.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. Ser. B, 36: Ser. B 36, 111–148.
van't Veer, L.J. et al., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415 (6871), 530–536.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, Berlin.
Webb, A.R., 2002. Statistical Pattern Recognition. Wiley, Chichester.
Wickenberg Bolin, U. et al., 2006. Improved variance estimation of classification performance via reduction of bias caused by small sample size. BMC Bioinformatics 7 (Mar 13), 127.
Xu, Q. et al., 2006. Confidence intervals for the true classification error conditioned on the estimated error. Technol. Cancer Res. Treat 5 (6), 579–590.
Yeoh, E.J. et al., 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1 (2), 133–143.