BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 10 2006, pages 1245–1250
doi:10.1093/bioinformatics/btl066

Gene expression

Avoiding model selection bias in small-sample genomic datasets

Daniel Berrar*, Ian Bradbury and Werner Dubitzky
School of Biomedical Sciences, University of Ulster at Coleraine, Northern Ireland

Received on October 19, 2005; revised on January 29, 2006; accepted on February 20, 2006
Advance Access publication February 24, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Genomic datasets generated by high-throughput technologies are typically characterized by a moderate number of samples and a large number of measurements per sample. As a consequence, classification models are commonly compared based on resampling techniques. This investigation discusses the conceptual difficulties involved in comparative classification studies. Conclusions derived from such studies are often optimistically biased, because the apparent differences in performance are usually not controlled in a statistically stringent framework taking into account the adopted sampling strategy. We investigate this problem by means of a comparison of various classifiers in the context of multiclass microarray data.

Results: Commonly used accuracy-based performance values, with or without confidence intervals, are inadequate for comparing classifiers for small-sample data. We present a statistical methodology that avoids bias in cross-validated model selection in the context of small-sample scenarios. This methodology is valid for both *k*-fold cross-validation and repeated random sampling.

Contact: dp.berrar@ulster.ac.uk

1 INTRODUCTION

Classification is arguably one of the most important and most widely studied analytical tasks for analyzing datasets generated by high-throughput technologies in biology and biotechnology. Numerous novel classifiers have been developed to address the specific challenges posed by such datasets. Owing to the high-dimensionality and dataset sparsity, these models are typically benchmarked against existing models based on their performance on resampled data subsets. In one form or another, such comparative studies have the same question in mind: Does the novel model provide a significant improvement over existing models? It is a standard practice to address this question on the basis of a cost function for classification, e.g. accuracy-based performance measures such as the correct classification rate, or based on more intricate cost functions for false positive and false negative classifications (Brown *et al*., 2000).

In this study, we demonstrate that for small-sample datasets using cost functions alone can be highly problematic in answering the aforementioned question, because they often involve a model selection bias. We illustrate this problem by analyzing high-dimensional gene expression data generated by microarray experiments investigating cancer. To address the potentially serious implications in biomedical settings, we propose a stringent statistical methodology for assessing classifiers.

This study is concerned with the classification of different types of cancer based on expression profiles. In contrast to frequently encountered binary classification tasks, the classification tasks chosen for this study are multiclass problems, which are considered more challenging. Comparative studies addressing multiclass microarray data include, for example, Ross *et al*. (2000), Yeoh *et al*. (2002), Dudoit *et al*. (2002) and Li *et al*. (2004). Normally, these studies compare the performance of various classifiers based on the (observed) correct classification rate, or alternatively, the (observed) error rate, implying a 0–1 loss function. The literature provides numerous examples of novel classifiers claimed to be superior to competitors on the basis of benchmark tests involving observed accuracy measures, e.g. Brown *et al*. (2000), Wang *et al*. (2003), Berrar *et al*. (2003). However, conclusions claiming superiority of one method over the other often lack sufficient evidence to support such claims. Three of the main reasons for this problem to occur include the (1) 'orphaned' values problem (absence of suitable confidence intervals), (2) difference in data processing and sampling strategy and (3) inadequate testing strategies. These are now discussed in turn.

First, the observed error rate is only an estimate for the model's true error rate on the population of interest, i.e. the set of samples that are described by an expression profile similar to the investigated dataset (e.g. a 'population' of similar microarraystudies). Reporting the observed error rate without estimated confidence intervals (or, as the Bayesian analogue, credibility intervals) of the true error rate is of limited value. Indeed, the interpretation of such 'orphaned' values is non-trivial and has contributed to heated debates (Liotta *et al*., 2004, http://erc.endocrinology-journals.org/cgi/content/full/11/4/585). Given the study of a particular population, the true error rate constitutes an inherent property of the model. The observed error rate is only an estimate and depends largely on the adopted sampling strategy. Single performance scores represent estimates of a true value and, in particular in small-sample settings, are difficult to interpret without confidence intervals for the true statistic. For example, it does clearly make a difference whether a classifier's correct classification rate of, say, 80% is based on 100 or on 10 000 test cases. Confidence intervals for the statistic of interest are of particular importance in scenarios comprising small datasets such as

*To whom correspondence should be addressed.

microarray and other high-throughput data in biology. Nevertheless, confidence intervals are rarely reported in the microarray literature.

Second, even when observed error rates and their confidence intervals for the true errors are available in small-sample scenarios, it is logically inappropriate to directly use these values for assessing whether the classifiers' performance is significantly different. Even if exactly the same datasets are used and confidence intervals are reported, it is problematic to compare the results from different studies. These studies usually differ with respect to the adopted data pre-processing techniques (e.g. different data pre-processing steps), sampling strategies (e.g. different cross-validation procedures) and model learning techniques (e.g. parameter settings). For example, it is of rather ambiguous benefit to know that a classifier *A* achieved an accuracy rate of *x*% on a particular dataset *D*, whereas a classifier *B* achieved *y*% on the same dataset. The observed difference, $|x\% - y\%|$, in accuracy could, for instance, have been caused by different experimental settings, rather than by systemic differences in the analytical classification methods.

Third, even if a comparative study adopts the same sampling strategy for identical datasets, there is a critical flaw in the logic underlying such simplistic comparisons. From a statistical perspective it is conceptually unsatisfactory to compare multiple classifiers based only on their achieved accuracy scores without considering statistical significance tests that (1) take into account the specific data sampling strategy and (2) correct for multiple testing.

In the context of microarray data, Somorjai *et al*. (2003) asked the question 'What do we want from a classifier?' and identified the model's robustness as a critical feature. Somorjai *et al*. (2003) criticized the common practice in classifying microarray data where the most intricate models are chosen, without assessing whether simpler models might be adequate. In the present study we address the question 'What do we want from a comparative study?' Bluntly put, we want a ranking from the 'best' to the 'worst' classifier for the specific task at hand. In some form or another all comparative studies are concerned with the following question: Do the observed differences in performance provide sufficient evidence to conclude that the models perform significantly differently, or can we not exclude the possibility (with reasonably confidence) that this difference may be due to chance alone or to the random variation introduced by the sampling strategy? This question, however, cannot be answered on the basis of cost functions alone, but requires an adequate statistical significance test for the difference in performance.

This study demonstrates (1) that accuracy-based measures—even with appropriate intervals—are insufficient to fairly compare classifiers on small-sample datasets and (2) that claims of the form Model *A* outperforms model B because of the observed accuracy rate difference of *x*% are to be taken with extreme caution. The aim of this comparative study is not to identify the best classifier for microarray data. Instead, we seek to pinpoint common pitfalls and caveats in microarray data classification and present a statistical methodology for detecting significantly different performance of cross-validated classifiers in small-sample scenarios.

2 MATERIALS AND METHODS

We demonstrate the methodology on the basis of three well-studied, publicly available datasets, one based on cDNA chips and two on Affymetrix oligonucleotide arrays. These publicly available datasets are chosen because they have been widely used as benchmarking datasets.

The NCI60 dataset comprises gene expression profiles of 60 human cancer cell lines of various origins (Ross *et al*., 2000). We pre-processed, normalized and cleansed this dataset following the protocol of Scherf *et al*. (2000). The thus prepared cDNA data comprises 60 cases from 9 cancer classes.

The ALL dataset represents the expression profiles of 327 pediatric acute lymphoblastic leukemia samples (Yeoh *et al*., 2002). This data comprises 10 classes and the expression profiles of a total of 12 600 annotated genes and multiple ESTs. The dataset is pre-processed according to the protocol by Yeoh *et al*. (2002).

The GCM dataset contains the expression profiles of 198 specimens of predominantly solid tumors of 14 cancer types (Ramaswamy *et al*., 2001). The normalized dataset comprises the expression profiles of 16 063 genes (Ramaswamy *et al*., 2001).

Various families of classifiers have been applied to the analysis of microarray data. For the comparative study, we decided to include at least one classifier per family, and to focus on those models that have been widely used in the context of microarray data. As a representative of large-margin classifiers, we chose support vector machines (SVMs), which have been used in, for example, Brown *et al*. (2000); Ramaswamy *et al*. (2001), Yeoh *et al*. (2002) and Li *et al*. (2004). For the large family of neural networks, we chose three representatives, (1) multilayer perceptrons (MLPs), which have been used in, for example, Khan *et al*. (2001), Yeoh *et al*. (2002) and Li *et al*. (2004); (2) radial basis function networks (RBFs) (Broomhead and Lowe, 1988) and (3) probabilistic neural networks (PNN) as parallel implementation of Parzen windows, used in, for example, Berrar *et al*. (2003). Decision trees have been used in, e.g. (Zhang *et al*., 2001; Dudoit *et al*., 2002; Li *et al*., 2004). We chose two widely used models, the decision tree *C*5.0 (Quinlan, 1993) and classification and regression trees (CART) (Breiman *et al*., 1984). As a method for generating ensemble classifiers, we chose boosting because this approach showed superior performance to other aggregation methods (Dudoit *et al*., 2002). As a representative of instance-based classifiers, we chose a *k*-nearest neighbor (*k*-NN) model with a distance-weighted voting scheme, because it has shown excellent performance in previous studies (Dudoit *et al*., 2002; Radmacher *et al*., 2002).

For assessing the classification accuracy, various data re-sampling strategies have been suggested, including leave-one-out cross-validation (LOOCV), *k*-fold cross-validation, repeated random subsampling (a.k.a. repeated hold-out method), and bootstrapping. The difference between repeated random sampling and cross-validation is that in the latter, the test sets do not overlap. Notice that *k*-fold cross-validation may not be practical for small-sample datasets, because the test sets would be too small (Dudoit *et al*., 2002; Li *et al*., 2004). We adopt the widely used repeated random subsampling approach (Scherf *et al*., 2000; Dudoit *et al*., 2002), noting that the methodology also applies to *k*-fold cross-validation.

The NCI60 dataset is pre-processed using principal component analysis based on singular value decomposition, and the first 23 'eigengenes' (explaining >75% of the total variance), are selected. The 10 dataset pairs (L_i, T_i), $i = 1, \ldots, 10$, are generated by randomly sampling (without replacement) 45 cases for the learning set L_i and 15 cases for the test set T_i.

The ALL and GCM datasets are analyzed in a procedure involving a repeated random sampling of learning and test cases to generate 10 pairs of (pairwise disjoint) learning (L_i) and test sets (T_i). Based on the learning set L_i only, we determined the signal-to-noise weight in a one-versus-all approach (Slonim *et al*., 2000) for each gene with respect to each class. Then, we randomly permuted the class labels and performed a random permutation test involving 1000 iterations to assess the significance of the signal-to-noise weights (Radmacher *et al*., 2002). We ranked the genes according to their weight and the associated *P*-value to construct the predictor sets containing only class-discriminatory genes. The learning and test cases are identical for all models. The models are constructed on the basis of the learning set L_i in LOOCV and tested on the test set T_i with the corresponding genes. Those parameters that led to the smallest cumulative error in L_i are then used to classify the cases in T_i. It is critical that the sets L_i and T_i are disjoint, but any given two learning sets do overlap. The test sets are never used for model selection or feature selection (external cross-validation) (Ambroise and MacLachlan, 2002). Each learning phase encompasses a complete re-calibration of the models' parameters, including various distance metrics for the k-NN; the optimal number of nearest neighbors for the k-NN, and so on.

3 BIAS IN CROSS-VALIDATED MODEL SELECTION

3.1 Caveat 1: observed versus true error rates

In general, single performance estimates (e.g. the observed correct classification rate or the observed error rate) convey little information without confidence intervals for the true statistic. Confidence intervals for the statistic of interest are of particular importance in scenarios that comprise small datasets such as microarray data. Unfortunately, such confidence intervals are rarely reported, rendering the interpretation of single accuracy measures problematic. Moreover, these studies often do not distinguish between the observed statistic and the true statistic. For example, if a classifier achieves a misclassification rate of 20%, then the observed error rate is 20%, while the true error rate of this classifier is (probably) ∼20%. This true error rate is an inherent property of the classifier for the population under investigation and can be estimated using the observed statistic.

There exist different approaches for deriving statistical confidence intervals. Depending on the size of the test set and the number of observed misclassifications/correct classifications, some caveats in deriving such intervals should be noted. In the following, we focus on the true error rate, but the calculation for the true accuracy is analogous.

Let M denote the number of test cases and let m denote the number of incorrectly classified test cases. The estimated (observed) error rate is $\varepsilon = m/M$. Let τ denote the true error rate of the classifier for the population under investigation. Statistical textbooks (e.g. Anderson and Sclove, 1986) give the following equation for the deriving a confidence interval:

$$\tau \approx \varepsilon \pm (0.5/M + zs), \tag{1}$$

with $s = \sqrt{\varepsilon(1-\varepsilon)/M}$ and $z = \Phi^{-1}(1 - \frac{1}{2}\alpha)$, e.g. $z = 1.96$ for 95% confidence, with $\Phi(\bullet)$ being the standard normal cumulative distribution function. The term $0.5/M$ is for continuity correction and should not be neglected for small M.

Kohavi (1995) proposes the following normal approximation to the binomial: $|\varepsilon - \tau| / \sqrt{\varepsilon(1-\varepsilon)/M} < \Phi^{-1}(1 - \frac{1}{2}\alpha)$. This translates into a confidence interval for τ as follows:

$$\tau \approx \left\{ \varepsilon + \frac{(1-2\varepsilon)z^2}{2(M+z^2)} \right\} \pm zs,$$

with

$$s = \sqrt{\varepsilon(1-\varepsilon)/M + \frac{z^2}{[2(M+z^2)]^2}[1 - 4\varepsilon(1-\varepsilon)(2 + z^2/M)]} \tag{2}$$

Both intervals in Equations (1) and (2) are based on the assumption that the number of observed errors, m, follows a binomial distribution. The approximation of the binomial by the normal distribution with mean $M\tau$ and variance $M\tau(1-\tau)$ is valid whenever $M\varepsilon(1-\varepsilon) \geq 5$ (Rosner, 2000). However, for relatively small test sets or small observed error rates, e.g. $M = 80$ and $\varepsilon = 1/16$, this approximation may not be appropriate. Both equations quantify the probability to observe m errors given the test set size M and the true error τ, i.e. $P\{m\,|\,M, \tau\}$. Note that this approach derives a frequentist confidence intervals for τ. Arguably more interesting is the likelihood that τ is smaller than a certain value τ_a, given the test set size M and the observed errors, m. This means that we are interested in the posterior distribution of τ, hence $P\{\tau \leq \tau_a\,|\,M, m\}$. It has been shown that for deriving confidence intervals—or, more precisely, Bayesian credibility intervals—for the true error rate and in the absence of problem-specific knowledge and an assumed binomial distribution of the errors, the best choice for the prior is given by Jeffreys' Beta distribution (Bernardo, 1979; Martin and Hirschberg, 1996). Based on Jeffreys' Beta distribution, an approximate $(1-\alpha)$%-credibility interval for the true error rate can be derived as follows:

$$\tau \approx \left\{ \varepsilon + \frac{2(M-2m)z\sqrt{0.5}}{2M(M+3)} \right\} \pm z \cdot \sqrt{\frac{\varepsilon(1-\varepsilon)}{M+2.5}} \tag{3}$$

Martin and Hirschberg (1996) compared this approximation to the precise numeric solutions for various values of m and M. The approximation is adequate for $10 \leq M \leq 200$ and $0 \leq m \leq \frac{1}{2}M$.

Statistics textbooks' methods implicitly assume a uniform prior distribution for τ, $f\{\tau\} = 1$. As the test set size increases, the intervals obtained by these methods become narrower and the assumptions made about the prior become less important. However, if the test set size is small, then the influence of the assumed prior should not be neglected. A uniform prior $f\{\tau\} = 1$ implies complete ignorance of the classifier's true error rate, and is equivalent to the a priori assumption that the classifier is equally likely to exhibit a true error rate of 0 or 100% on the population of interest. This assumption, however, does not take into account that a classifier usually classifies at least some of the cases of the learning set correctly. Since the samples in the learning set belong to the underlying population, the true error rate cannot possibly be 100%, but must be smaller. Real-world datasets in biology and biotechnology, particularly microarray datasets, are characterized by a large degree of noise and missing values, so that it is unlikely that a classifier will be able to correctly classify all cases of the population of interest, thus, it is also unlikely that the true error rate is 0%.

Figure 1 shows τ as a function of the test set size, M. Merely for the sake of this argument, we make the assumption that the observed error rate is constant with $\varepsilon = 0.15$, i.e. for all sizes of test sets, 15% of the cases are misclassified (here, $\tau = \varepsilon$). Figure 1 shows that the differences between the intervals—both with respect to the midpoints and widths—are more extreme for smaller test sets and converge to ε for $M \rightarrow \propto$.

For a test set size of $M = 300$, the midpoints of the three intervals differ by <3%, but both Kohavi's interval and the interval based on Jeffrey's prior are considerably wider than the narrow interval given by Equation (1). Both Kohavi's interval and the interval based on Jeffrey's prior are ∼96 times wider than the interval in Equation (1). This difference becomes even more apparent for smaller test sets,

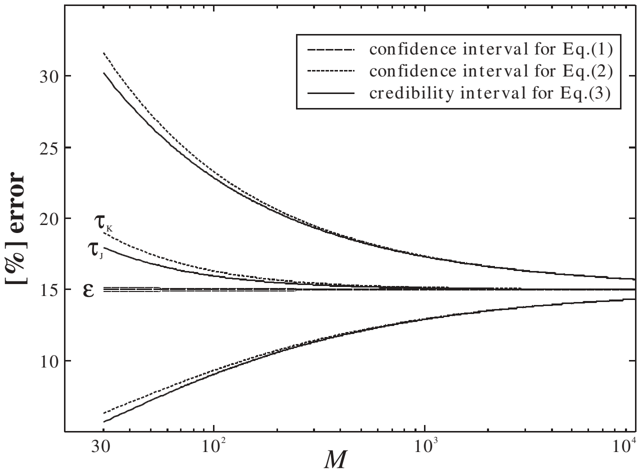

Fig. 1. The 95%-confidence/credibility intervals for τ based on Equations (1)–(3), with constant observed error rate $\varepsilon = 0.15$ for $M = 30$ to $M = 10\,000$. The midpoints of the intervals are ε for Equation (1), τ_K for Equation (2) and τ_J for Equation (3).

Table 1. The 95%-CI for the true error rate τ(in %) based on the Beta distribution using Jeffreys' prior

Model	Dataset NCI60	ALL	GCM
k-NN	27.74 ± 7.07	**22.12 ± 2.43**	25.56 ± 3.88
SVMs	**21.20 ± 6.43**	22.39 ± 2.44	24.11 ± 3.81
C5.0	36.91 ± 7.65	31.12 ± 2.71	35.09 ± 4.26
2-fold boosted	35.60 ± 7.59	31.39 ± 2.71	35.50 ± 4.27
3-fold boosted	41.49 ± 7.82	28.98 ± 2.65	36.95 ± 4.31
4-fold boosted	37.56 ± 7.68	27.38 ± 2.61	33.43 ± 4.21
5-fold boosted	37.56 ± 7.68	27.47 ± 2.61	31.98 ± 4.16
CART	41.49 ± 7.82	36.02 ± 2.81	65.54 ± 4.24
PNN	23.16 ± 6.65	22.57 ± 2.44	**21.00 ± 3.62**
RBF	44.76 ± 7.89	31.12 ± 2.71	24.94 ± 3.85
MLP	38.22 ± 7.70	29.78 ± 2.67	44.82 ± 4.44

Smallest error rates are in bold face.

e.g. $M = 100$. Here, the widths of the intervals from Equations (2) and (3) differ by ~1% and are both much wider than the interval given by Equation (1), $\tau = 15.00 \pm 0.08$. The interval from Equation (2) is ~93 times wider than the interval from Equation (1), while the interval from Equation (3) is ~92 times wider.

Table 1 shows the intervals for the true error rates of the models in the present study.

3.2 Caveat 2: error rates versus significance of differences

The sampling procedure introduces a random variation in the sampled datasets, which must be taken into account by the statistical test. In the context of small-sample microarray data, three sources of variation owing to the random sampling procedure can be identified (Dietterich, 1998):

1. *Random variation in the make-up of the test sets*. It is possible that a classifier *A* outperforms classifier *B* on a particular test set T_i, although on the whole population the two classifiers might perform equally.
2. *Random variation in the predictor sets*. It is possible that *A* achieves a higher classification accuracy than *B* when trained on a particular predictor set L_i, although both classifiers may perform equally on the entire population.
3. *Random classification errors*. If the dataset contains a fraction of randomly mislabeled cases, then no classifier can achieve an error rate less than this fraction.

Conceptually, a 95% confidence (or credibility) interval for an estimate (e.g. the true error rate) is completely different from a 95% confidence level for the difference of two estimates (e.g. the difference between the error rate of model *A* and *B*). Therefore it should be noted explicitly that it is logically inappropriate to use the derived confidence intervals for assessing whether there is a significant difference in performance of the classifiers!

Let p_{Ai} be the observed proportion of test cases misclassified by *A* and let p_{Bi} be the observed proportion of misclassified test cases by *B* during the *i*-th fold. If we assume that the differences $p_i = p_A - p_{Bi}$ were drawn independently from a normal distribution, then we could apply Student's *t*-test. However, the assumptions underlying this test are violated! This is because in cross-validation and repeated random subsampling, the learning sets necessarily overlap; in repeated random subsampling, the test sets may overlap as well. Hence, the individual differences p_i are not independent from each other. The high Type I error of Student's *t*-test is due to an underestimation of the variance because the samples are not independent. Assume that in each fold *N* cases are used for learning and *M* cases are used for testing. Let the number of folds be *k*, and let the difference of proportion of misclassified cases be $p_i = p_{Ai} - p_{Bi}$, with $i = 1, \ldots, k$ and $p_{Ai} = m_{Ai}/M$, with m_{Ai} the number of errors on the *i*-th test set comprising *M* cases (p_{Bi} and m_{Bi} analogous). Further, let the average of p_i over the *k* folds be $\bar{p} = k^{-1} \sum_{i=1}^{k} p_i$. The estimated variance of the *k* differences is $s^2 = (k-1)^{-1} \sum_{i=1}^{k} (p_i - \bar{p})^2$. The statistic for the variance-corrected resampled paired *t*-test is then given by Equation (4).

$$T_c = \frac{\bar{p}}{\sqrt{(k^{-1} + c)s^2}} \sim t_{k-1} \qquad (4)$$

Nadeau and Bengio (2003) set the constant *c* to *M*/*N*, assuming that *N* is about five times larger than *M*. Empirical results show that this corrected statistic drastically improves the standard resampled *t*-test with respect to the Type I error (Nadeau and Bengio, 2003; Bouckaert and Frank, 2004). However, this improvement might come at the cost of an increased Type II error, particularly when the training set is not ~5 times larger than the test set.

3.3 Caveat 3: correction for multiple testing

In the context of multiple comparisons, another important caveat must be respected. When the study comprises *n* classifiers, a total of $\kappa = \frac{1}{2}n(n-1)$ pairwise comparisons are possible. The α of each individual test is the comparison-wise error rate, while the family-wise error rate, α_κ, is made up of the κ individual comparisons. It is essential to adjust the comparison-wise error for multiple testing. Bonferroni's correction for multiple testing, $\alpha = \alpha_\kappa/\kappa$, is known to be a rather conservative approach, which implies that true null hypotheses will not be rejected too often; however, false ones will frequently fail to be rejected.

The Holm test (Holm, 1979) is a compromise between the too conservative Bonferroni correction and Fisher's more liberal least significant difference (LSD) test. The Holm test rejects more false null hypotheses than the Bonferroni method, hence has greater power, and in contrast to Fisher's LSD test, it controls the family-wise α_κ at the desired level.

For the NCI60 dataset, the smallest P-values (based on the variance-corrected paired t-test) are $P_1 = 0.04$ for the comparison SVMs versus 3-fold boosted decision tree and $P_2 = 0.05$ for the comparison SVMs versus RBF. In the third sampling fold, however, RBF classified more cases correctly than SVMs, hence explaining the larger P-value for the comparison SVMs versus RBF in contrast to SVMs versus 3-fold boosted decision trees. With $n = 11$ models being compared, however, we obtain $\kappa = \frac{1}{2}n(n-1) = 55$ and $\alpha = 0.05/\kappa = 9.1 \times 10^{-4}$, hence the Holm test fails to reject the null hypothesis of equal performance between SVMs and 3-fold boosted decision trees. Even when a standard paired t-test (i.e. without correction term) is used, the difference is not significant ($P = 0.001 > 9.1 \times 10^{-4}$). The apparent best performers are the SVMs with a test error rate of 21.20 ± 6.43%, while the apparent worst performer is RBF with a test error rate of 44.76 ± 7.89%. Despite the fact that the intervals do not overlap, the experiment does not allow to reject the null hypothesis at the chosen confidence level, i.e. we cannot exclude that the observed difference in performance is because of chance alone. Surprisingly, no model significantly outperformed any other model on the NCI60 dataset. For example, the SVMs classified more cases correctly than the RBF on average, but the SVMs did not consistently (i.e. over all 10-folds) classify more cases correctly. The apparent difference in performance between SVMs and RBF (correct test classification rate of 79.3% versus 55.3%) does not imply a statistical difference at the chosen confidence level (see Caveat 2 in Section 3.2). This is also true when the term for variance-correction is omitted ($P = 0.001 > 9.1 \times 10^{-4}$).

For the ALL dataset, the smallest P-value is $P_1 = 1.67 \times 10^{-6}$ for the comparison between SVMs and C5.0. Comparing this value with $\alpha = 0.05/\kappa = 9.1 \times 10^{-4}$ allows to reject the null hypothesis of equal performance. The second smallest P-value is $P_2 = 6.24 \times 10^{-6}$ for the comparison between SVMs and 2-fold boosted decision trees; this value is compared with $\alpha = 0.05/(\kappa - 1) = 9.3 \times 10^{-4}$ and leads to the rejection of the null hypothesis. Analogously, the following models perform significantly differently: PNN versus 2-fold boosted decision trees ($P_3 = 4.71 \times 10^{-5}$) and PNN versus C5.0 ($P_4 = 1.22 \times 10^{-4}$). The null hypothesis cannot be rejected for the next comparison, PNN versus RBF with $P_5 = 2.38 \times 10^{-3}$, because $P_5 > 0.05/(\kappa - 4) = 9.8 \times 10^{-4}$.

For the GCM dataset, the difference in performance of the k-NN and the MLP is significant, whereas the difference between SVM and MLP is not. This might be surprising, given that the interval for τ of the SVM is more to the left (lower error rate) than the respective interval of the k-NN as can be seen in Figure 2.

Considering the intervals depicted in Figure 2, it seems counter-intuitive that the null hypothesis of equal performance of the SVMs and MLP cannot be rejected. However, as explained above, intervals for an estimate (here, an estimate of the true error rate) are logically inappropriate for testing the difference between two estimates. It might be possible that for the GCM dataset, SVMs are to be preferred over MLPs; it might be possible that by including more folds in the sampling procedure, the difference in performance could become significant. However, the experiments as carried out in the present study (as well as numerous similar experiments reported in the literature) do not provide sufficient evidence for reporting that, in this case, SVMs performed significantly better than MLPs. This is an important result of this discourse and should be noted.

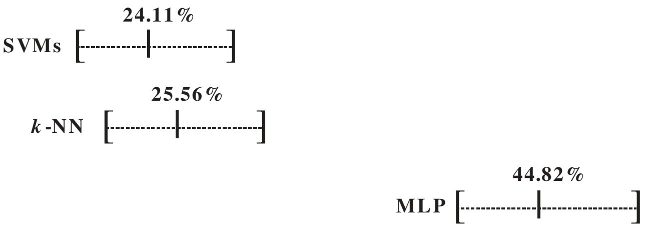

Fig. 2. The 95%-credibility interval for the true error rate τ on the GCM test set for the SVMs, k-NN and MLP.

Being primarily interested in how classifiers fare with respect to the individual datasets, this study did not correct the comparison-wise error rate for the multiple datasets. However, if a comparative study tries to identify an overall 'winner' over multiple datasets, then the comparison-wise error rates need to be adjusted accordingly.

4 DISCUSSION AND CONCLUSIONS

The observed levels of accuracy or error can vary substantially as a function of the adopted sampling strategy. Clearly, confidence or credibility intervals for the true error rates represent a more informative criterion than monolithic accuracy or error values. However, in classification scenarios involving small sample sizes and relatively low error rates, the derived intervals can be too narrow, particularly if they are based on textbook formulae without a continuity correction. This can lead to misleading conclusions that claim one model outperforms the other.

Therefore, single accuracy measures and even confidence intervals for the predictive accuracy should not be used for answering the question of interest. Particularly in tasks involving relatively small test sets (in the order of 10^2) and classifiers with small error rates, different classification accuracies seem to suggest a difference in performance, whereas rigorous statistical testing cannot reject the null hypothesis of equal performance. Random permutation tests could represent an alternative to parametric tests in this context. Recently, Statnikov *et al.* (2005) have conducted a comparative study and assessed the differences in performance using a random permutation test. However, their study did not correct for multiple testing.

Since many recent studies focus on multi-class classification, we discussed the caveats and pitfalls of datasets involving multiple classes. Some of the discussed caveats may manifest themselves less severely in two-class scenarios, particularly when the class distribution is well balanced. When multiple classes are involved (particularly with unbalanced distribution) comparisons of accuracy measures can be highly misleading (Provost and Fawcett, 1998).

Which classifiers should be included in a comparative study? This question deserves careful consideration. If a novel method is to be benchmarked against established techniques, then it is essential that the most similar methods are chosen as the competitors. Including too many (and possibly largely different in terms of their properties) classifiers in the study, however, may result in an adjusted comparison-wise error rate that will not reject a true null hypothesis and can therefore unnecessarily increase the Type II error rate. Essentially, the nature of the dataset at hand should guide the choice of the benchmark classifiers. For example, if the dataset involves a binary classification problem and the novel model is an optimal separating hyperplane tailored to this type of data, then the comparative study should include binary SVMs that have demonstrated to perform well for these tasks. Different chores require different tools—comparing a screwdriver with a hammer is questionable.

For comparing the performance of classifiers on a particular dataset, it is essential that (1) the training and test sets are identical for the classifiers, (2) the sampling strategies are the same, (3) the learning phases include a complete parameter re-calibration and external cross-validation, (4) the difference in performance is assessed by means of a suitable statistical test, which is appropriate for the adopted sampling strategy and accounts for both comparison- and family-wise error rates by adjusting for multiple testing and (5) the competing classifiers are carefully chosen.

The proposed methodology for variance correction is applicable in the context of cross-validation and repeated random subsampling. However, we note that while the correction leads to a decrease of the Type I error, the Type II error necessarily increases. Future work will need to focus on a weighting approach for the number of folds, k, and the correction term, M/N, in Equation (4).

The correction for multiple testing is less conservative than Bonferroni's approach, and with respect to Type I and II error rates, it is comparable with more intricate adjustments (Manly *et al*., 2004). Most importantly, the methodology presented in this work can be used to address the question of interest that all comparative studies ask, be it implicitly or explicitly. The methodology is limited in that it is based on a parametric test, which can be criticized from a purely Bayesian's perspective.

Minimizing the misclassification error rate is only one aspect that needs to be taken into account in constructing classifiers. Another critical feature of a classifier is its complexity. From an Occam's razor perspective, the simpler model is generally to be preferred over a more sophisticated model.

Conflict of Interest: none declared.

REFERENCES

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on th basis of microarray gene expression data. *Proc. Natl Acad. Sci. USA*, **98**, 6562–6566.

Anderson,T.W. and Sclove,S.L. (1986) *The Statistical Analysis of Data*, 2nd edn. Scientific Press, Palo Alto.

Bernado,J.M. (1979) Reference posterior distributions for Bayesian inference. *J. R. Statist. Soc*., **B41**, 113–147.

Berrar,D.P., Downes,C.S. and Dubitzky,W. (2003) Multiclass cancer classification using gene expression profiling and probabilistic neural networks. In *Proceedings of the 8th Pacific Symposium on Biocomputing,* Lihue, Hawaii, USA, World Scientific, NJ, pp. 5–16.

Bouckaert,R.R. and Frank,E. (2004) Evaluating the replicability of significance tests for comparing learning algorithms. In *Proceedings of the 8th Pacific-Asian Conference on Knowledge Discovery and Data Mining*, Australia, pp. 3–12.

Breiman,L., Friedman,J., Olshen,R. and Stone,C. (1984) *Classification and Regression Trees*. Chapman & Hall, New York.

Broomhead,D.S. and Lowe,D. (1988) Multivariate functional interpolation and adaptive networks. *Complex Systems*, **2**, 321–355.

Brown,M.P.S. *et al*. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 263–267.

Dietterich,T. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*., **10**, 1895–1923.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat*., **6**, 65–70.

Khan,J. *et al*. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med*., **7**, 673–679.

Kohavi,R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence,* Montréal, Québec, Canada, pp. 223–228.

Li,T. *et al*. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**, 2429–2437.

Liotta,L.A. *et al*. High-resolution serum proteomic patterns for ovarian cancer detection. Letter to the editor. http://erc.endocrinology-journals.org/cgi/content/full/11/4/585.

Manly,K.F. *et al*. (2004) Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res*., **14**, 997–1001.

Martin,J.K. and Hirschberg,D.S. (1996) Small sample statistics for classification error rates II: confidence intervals and significance tests. *Technical Report #96-22*, University of California, Irvine, CA.

Nadeau,C. and Bengio,Y. (2003) Inference for generalization error. *Mach. Learn*., **52**, 239–281.

Provost,F.J. and Fawcett,T. (1998) Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, New York, USA, pp. 43–48.

Quinlan,J.R. (1993) *C4.5:Programs for Machine Learning*. Morgan Kaufmann, San Francisco.

Radmacher,M.D. *et al*. (2002) A paradigm for class prediction using gene expression profiles. *J. Comput. Biol*., **9**, 505–511.

Ramaswamy,S. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Rosner,B. (2000) *Fundamentals of Biostatistics*, 5th edn. Duxbury Press, USA.

Ross,D.T. *et al*. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Gen*., **24**, 227–235.

Scherf,U. *et al*. (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Gen*., **24**, 236–244.

Slonim,D.K., Tamayo,P., Mesirov,J.P., Golub,T.R. and Lander,E.S. (2000) Class prediction and discovery using gene expression data. In *Proceedings of the fourth Annual International Conference on Computational Molecular Biology*, Universal Academy Press, Tokyo, Japan, pp. 263–272.

Somorjai,R.L. *et al*. (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, **19**, 1484–1491.

Statnikov,A., Aliferis,C.F., Tsamardinos,I., Hardin,D. and Levy,S. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**, 631–643.

Wang,J. *et al*. (2003) Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC Bioinformatics*, **4**, 60.

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**, 133–143.

Zhang,H. *et al*. (2001) Recursive partitioning for tumor classification with gene expression microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 6730–6735.