Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Chemometrics and Intelligent Laboratory Systems 79 (2005) 91 – 100

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Variance reduction in estimating classification error using sparse datasets

Claudia Beleites[b], Richard Baumgartner[a], Christopher Bowman[a], Ray Somorjai[a], Gerald Steiner[b], Reiner Salzer[b], Michael G. Sowa[a,*]

[a]National Research Council Canada, Institute for Biodiagnostics, 435 Ellice Avenue, Winnipeg, Manitoba, Canada MB R3B 1Y6
[b]Technische Universität, D-01062 Dresden, Germany

Received 17 May 2004; received in revised form 27 January 2005; accepted 14 April 2005
Available online 31 May 2005

Abstract

In biomedical applications, frequently only a limited number of samples are available for the development and testing of classification rules. Understanding the behavior of the error estimators in this setting is therefore highly desirable. In an extensive study using simulated as well as real-life data we investigated the properties of commonly used error estimators in terms of their bias and variance, and have found that in these small-sample size situations, the influence of variance on the error estimates can be significant, and can dominate the bias. Consequently, our results strongly suggest that bootstrap resampling and/or *k*-fold crossvalidation-based estimators, especially when computed over multiple data splits, should be preferred in these small-sample size scenarios, because of their reduced variance compared to the more routinely used crossvalidation approaches. While linear partial least squares was used as the classifier/regressor, the general conclusions arising from this study are not qualitatively affected for other classifiers, linear or nonlinear.

Keywords: Error rate estimation; Crossvalidation; Bootstrap resampling; Small sample size

1. Introduction

Classification deals with the problem of grouping observations into a number of known classes. Developing reliable classifiers for the class prediction of future observations requires some means to assess the quality of the developed classification rules. A common measure of model quality is the probability of misclassifying new observations that are not included in the development of the classification model. One calculates this generalization error as the proportion of misclassifications in a *test set*. One refers to the samples used to develop the classification model as the *training set*. In biomedical applications, the number of available samples that are well characterized and therefore suitable for use in classifier training and testing is often limited. When developing a classifier based on a sparse dataset, a number of caveats apply. The crux of the problem is that there are too few samples to adequately train and test the classifier. Too few training samples compromise the quality of the classification model and cast doubt on the reliability of the classification rules. Too few test samples provide only a crude measure of model quality, i.e. gross approximations to the generalization error of the classification model. Since one draws inference from the estimated error, overlooking the properties of the error estimator can lead to false conclusions regarding the prediction capability of the classification model.

This manuscript reviews common strategies for estimating the classification error when only a small number of samples are available. The properties of these estimators, in particular their bias and variance, are investigated. Bootstrap re-sampling and/or *k*-fold crossvalidation-based estimators are recommended because of their reduced variance compared to the more routinely used crossvalidation approaches.

1.1. Error estimation methods

The characteristics of different error estimation methods have been investigated for different sample sizes and

* Corresponding author. Tel.: +1 204 984 5193; fax: +1 204 984 5472.
E-mail address: mike.sowa@nrc-cnrc.gc.ca (M.G. Sowa).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.04.008

classifiers [1–12]. Estimators of the generalization error should be both precise and accurate. Inaccurate error estimators have a systematic deviation from the true error; they are *biased* estimators. This bias can only be determined by comparison with a reference method that provides the true generalization error. Imprecise estimators exhibit high variance. The variance of an estimator can be approximately determined by repeatedly assessing the estimator on models trained and tested on different partitions of the data. Thus, estimating a model's generalization error rate, both bias and variance contribute to the total uncertainty. However, usually only a single error estimate is employed as the quality parameter of the classification model. This practice provides no information on the variance of the error estimation, thus impairing the inferences available from such a measure of model quality. The problem is particularly acute when sample sizes are small.

1.2. Independent test set or hold-out method

When sample sizes are large, one can divide the data into independent training and test sets. The classification error predicted for the test set, using the independent training set to build the classification model, provides an unbiased estimator of the generalization error of the model. However, the variance of the estimator can be quite large with small test set sample sizes. Perhaps more importantly, the hold-out method requires that precious samples be further partitioned between training and testing the classifier model. Depriving the training and validating sequences of classifier development from samples can seriously compromise both the quality of the model and the reliability of the error estimation. The inefficient use of data inherent in the hold-out method is a serious liability for small sample sizes.

1.3. Resubstitution

The most obvious way to use all data for model training and additionally estimating the error is to predict the class memberships for the same data that was used for model building. This resubstitution estimate of the model's error actually measures the classifier's ability to adapt to the training data. Using this estimate leads to an optimistic bias. In many cases, the resubstitution error is extremely optimistically biased and is of limited value in judging the quality of the classification model, i.e. its prediction capability.

1.4. Resampling methods

To avoid optimistic bias, *independent* test and training sets should be used. Thus, in the small-sample case, one faces the conflicting goals of using as many samples as possible to build the classification model (training) and yet retaining as many samples as possible to estimate the model's error (testing). Resampling methods attempt to resolve this conflict by using all the data for both training and testing; this is done iteratively, ensuring that at each iteration the training and testing data are independent. Resampling partitions the original data into subsets, each of these being used to build a *submodel*. Samples excluded from the subset used to build the submodel are available as an independent test set, to be used to estimate the generalization error of the submodel. Averaging over the errors for the submodels provides an estimate of the generalization error for the model built using the full dataset.

In the following one tacitly assumes that the samples used are random, unbiased samples from the parent population. In general, this cannot be guaranteed for real-life data. Of course, no error estimator can compensate for biased sampling. This caveat should always be kept in mind. There are two assumptions underpinning resampling-based error estimation: (1) The submodel obtained from resampling is similar to the model that would be created if the entire dataset were used, and (2) The different submodels obtained through resampling are similar. However, these assumptions begin to break down as the sample size decreases. Since the average error for models based on smaller sample sizes is generally higher, there is an inherent source of pessimistic bias in this approach. The variability between the training samples used to build the various submodels leads to errors conditional on the particular training set used to build the submodel. For small sample sizes, this can be an important contributor to the variance in the estimated error. The various resampling methods suffer from different degrees of bias and variance in the estimation of the error.

1.4.1. k-fold crossvalidation (k-fold CV)

This crossvalidation (CV) strategy ensures that every sample is used for testing and these have the same weight in the error estimation. Data are randomly split into k sets of nearly equal size and k different submodels are built by iteratively using $k-1$ of subsets in each submodel. For each submodel the subset of the data excluded from building the model is used as the test set for that submodel. In this way all samples are used in both model training and testing over the sequence of k submodels and the error estimated over the k submodels (k-fold CV error estimate) provides an estimate of the generalization error of the model built on the entire dataset. The test and training sample sizes depend on k; common choices are $k=5$ or 10. In practice, using a single random split of the data is common; however, multiple splits can be done to help control the variance of the estimator.

1.4.2. Leave-one-out crossvalidation (LOO CV)

Leave-one-out (LOO) is the special case of the k-fold CV with k equaling the sample size N. Each sample is left out one at a time; N submodels are calculated and tested on the excluded samples. LOO error estimation is known to have a small pessimistic bias but may suffer from high variance. However, for small sample sizes, the bias can also be very large.

1.4.3. Out-of-bootstrap error estimation [4]

Crossvalidation methods form their training data by drawing samples from the dataset without replacement, i.e. once a sample is selected from the dataset, it is no longer available for future selections. Bootstrap methods use random selection with replacement; hence, samples can occur more than once in a bootstrap training set. The replacement provides the desired training sample size. Samples not selected or "out-of-bootstrap" are independent of that particular bootstrap set and can serve as test data for a submodel built on the bootstrap (training) set.

By drawing (with replacement) bootstrap sets of the same size as the source dataset, on average approximately 37% of the samples are left out of the bootstrap set and can be used to estimate the error. A common implementation averages the errors estimated for 100–200 submodels based on different bootstrap sets of the same size as the original dataset. Hence, small variance is expected due to averaging over the large number of submodels, but the estimate should be biased as the submodels on average are based on only 63.2% of the samples in the original dataset.

One approach to correct for this bias is to use 0.632 bootstrap estimation [4,5], which is a weighted average of the out-of-bootstrap (63.2%) and of the resubstitution (36.8%) estimates.

2. Experimental

2.1. Classification

Partial least squares discriminant analysis (PLS-DA) [13,14] was used to carry out the classifications. PLS-DA is a partial least squares regression of a set Y of binary response variables describing the categories or class membership on a set X of predictor variables. PLS models attempt to find orthogonal linear combinations of predictor variables (latent variables or factors) that account for the variability in the predictor space, while simultaneously being highly correlated with the response variable. This process consists of two parts: estimating a smaller set of orthogonal variables that are able to describe the regression problem, and building a regression model using these new variables. Retaining only a few latent variables effectively reduces the dimensionality of the data used to build the regression/classification model. This technique is particularly suited to deal with situations for which the number of predictor variables exceeds the number of samples and where collinearity exists among the predictor variables. These two situations arise when high-dimensional spectroscopic data are used to predict or classify a small number of samples or observations.

PLS models were built using the SIMPLS algorithm [15] (PLS Toolbox© 1997–1998 Eigenvector Research, Inc., Manson, WA, USA) using the Matlab scripting language (Matlab Version 6, Mathworks, Natick, MA, USA). To perform discrimination, the two classes were assigned dummy regression labels of $-N/n_1$ and N/n_2 and a threshold of zero was used ($N=n_1+n_2$ is the total number of samples, n_1=number of samples in class 1, n_2=number of samples in class 2). Hence, samples with regression values falling below the threshold are assigned to be in class 1, whereas those samples with regression values above the threshold belong to class 2.

2.2. Datasets

Large sample sizes are required to provide a reliable estimate of the true generalization error of the classification model. The ratio of the number of samples or observations to the number of features or predictor variables (sample to feature ratio, SFR) dictates the effective sample size. Common recommendations are that the SFR should be between 5 and 10 (e.g. [16–18]).

2.2.1. UCI Pima Diabetes dataset

This dataset is available from the UCI repository [19]. The samples have eight features or predictor variables and the dataset consists of 268 diabetes and 500 control cases. Classification models based on one to eight latent variables with randomly drawn training and test sets containing 5, 10, 25 and 50 samples per class were investigated. With respect to dimensionality, these datasets span the range from small to medium sample sizes. From the remaining data, independent sets of samples were drawn randomly, and used to provide large-sample estimates of the generalization error of the classification model. Stratified random draws of the data were used to ensure that each class had a balanced representation in the small sample training and test sets as well as the large sample reserved to provide a reliable estimate of the generalization error.

2.2.2. Simulated data

Inherent characteristics of spectroscopic data are the high dimensionality of the feature space and the presence of correlation between adjacent features. Under these conditions, the number of samples required to satisfy the recommendations regarding sample size becomes enormous and is rarely attainable in practice. Using two infrared spectra in the range of 1000 to 1800 cm^{-1} to represent true class means and a common covariance matrix proportional to the identity matrix, normally distributed data with an intrinsic class overlap or Bayes error rate of 10.5% were generated through simulation. Each base spectrum consists of 208 data points, resulting in a problem with high dimensionality. The diagonal covariance matrix provides a simple covariance structure consistent with other simulation studies appearing in the pattern recognition literature

[20,21]. Datasets of size 5, 10, 25, 50 and 100 samples per class were generated, resulting in small sample sizes compared to the dimensionality. Large-sample estimators of the generalization error were based on an additional independent test set of 100,000 samples (50,000 samples per class); these large test sets provided the reference error for a given classification model. Models from one to six latent variables were examined.

2.3. Error estimations

Assuming that the large-sample estimate of the generalized error represents the true error of the classification model, the quality (bias and variance) of the various small-sample error estimators can be judged by comparing them to the large-sample error estimate. Small sample subsets were drawn randomly from the original data in a manner that ensured equal class representation in each of the subsets. Data not used in the subset to build the classification model provided an independent large-sample estimator of the generalization error of the classification model. A measure of the mean and variance of the small-sample error estimators was obtained by forming 1000 small-sample partitions of the data and calculating the mean and variance of the small-sample error rate over the partitions. Fig. 1 illustrates the error estimation strategy.

For each small-sample partition, the resubstitution, LOO, 5-fold CV, and out-of-bootstrap errors were calculated. Bootstrap-based error estimations used 200 stratified bootstrap draws. In this implementation, the stratified bootstrap draws an equal number of samples from each class to ensure that balanced datasets were used to develop each bootstrap model. The 0.632 bootstrap estimation was computed as the sum of 63.2% of the out-of-bootstrap error and 36.8% of the resubstitution error of the bootstrap models. The 5-fold CV implementation used stratified random subsets, without repetition to ensure that an equal number of samples from each of the two classes were used to develop the classification models. The performance of the hold-out method was assessed somewhat differently: each small-sample partition was split further into stratified parts comprising 60% and 40% of the samples for building the model and estimating the hold-out error, respectively.

The difference between the large-sample reference error estimator and the small-sample error estimators is reported as a measure of the quality of the latter. Both the bias and the variance are investigated in this study. The bias estimate is the mean difference between the small- and the large-sample error estimators over the 1000 partitions. The variance is approximated by the variance of the difference between the large- and small-sample error estimators.

3. Results and discussion

Fig. 2 displays the mean and standard deviation (S.D.) of the reference error as a function of training set size for both the simulated and Diabetes data. A large independent subset

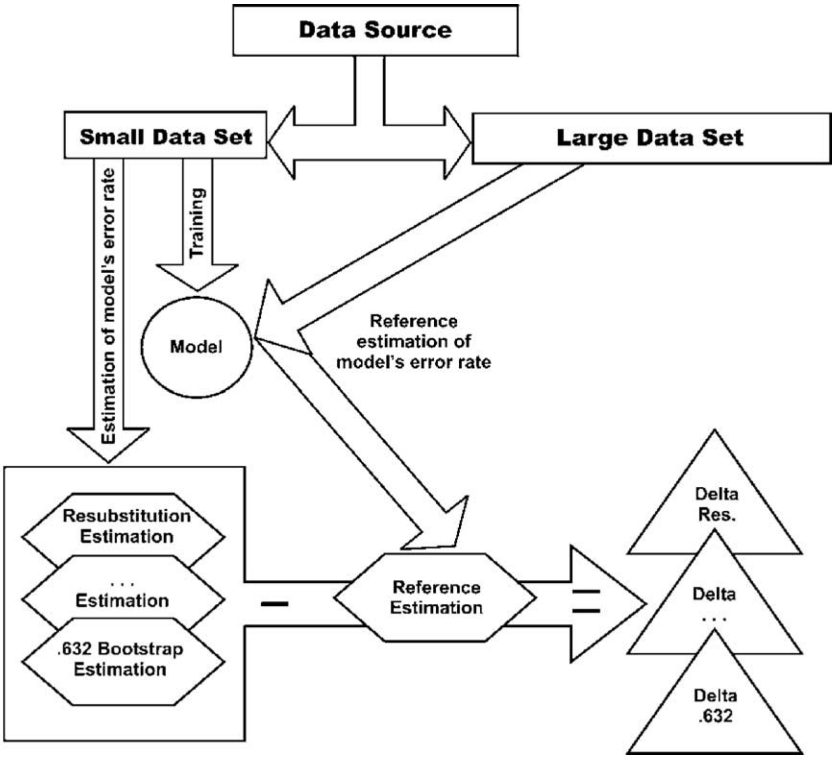
Data Source
Small Data Set
Large Data Set
Estimation of model's error rate
Training
Model
Reference estimation of model's error rate
Resubstitution Estimation
. . . Estimation
.632 Bootstrap Estimation
–
Reference Estimation
=
Delta Res.
Delta . . .
Delta .632

Fig. 1. Diagram of the error estimation strategy used to determine the bias and variance of the small-sample estimators relative to the large-sample error rate estimation. The large-sample error rate is assumed to be a good approximation to the true generalization error of the classification model.

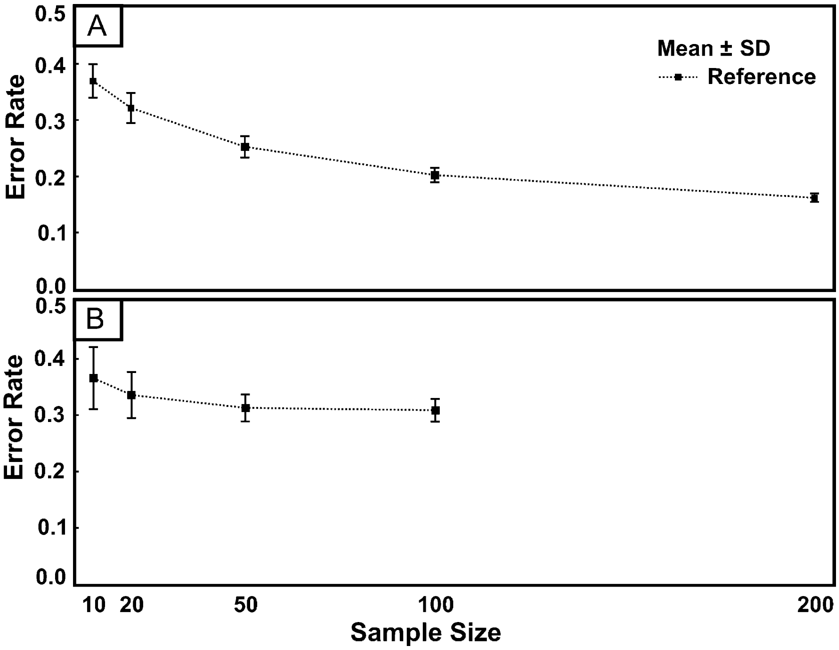

Fig. 2. (A) Reference error rate estimation for the simulated data based on using a large test set, 50,000 per class, to provide an estimate of the generalization error of the classification models built with varying training set sizes using a two-latent variable PLS-DA model. The mean and standard deviation of the reference error rate are reported over the 1000 different estimations performed on different subsets of the simulated data. (B) Reference error rate estimation for the Diabetes data based on using a large test set (200 per class) to provide an estimate of the generalization error of the classification models built with varying training set sizes using a two–latent variable PLS-DA model. The mean and standard deviation of the reference error rate are reported over the 1000 different estimations performed on different subsets of the Diabetes data.

of the data is reserved as a test set to provide a reference error estimation of the classification models. As expected, the reference error decreases as the size of the training set used to build the models increases. Similarly, the variance in the reference error estimation also decreases, mainly because of model improvement when the training set increases. The variability observed in the reference error derives from the finite sample size of both the test and training sets. However, given the large size of the subset of data used in the reference error estimation, we assume that variance arising from the finite test set is negligible in the reference error estimation compared to the variance arising from the finite training set. Although this should be a reasonable assumption for the simulated data, for which the test set sizes are extremely large (50,000 per class), the higher variability observed in the reference error for the Diabetes data suggests the presence of some residual variability due to the finite test set size (200 samples per class used in the large-sample reference error estimate). Since the variance introduced by the finite test set cannot exceed the variance observed over the 1000 reference error estimations, we can set an upper limit on the magnitude of the variance introduced by the finite test set size. For the Diabetes data, this effect was always less than 1/5 of the standard deviation in the difference between the reference error and the small-sample error estimators.

Examining the difference between the small-sample error estimators and the reference or large test sample error estimator provides some insight into the sources of bias and variance of the various small-sample error estimators. Taking the difference between small-sample error and reference error estimations effectively cancels out the variability arising from the small-sample training set used to build the models. We can attribute the difference between the small-sample and reference error estimations largely to the small-sample error estimator. Fig. 3A and B display the mean and S.D. of the difference between the reference error and small-sample error estimations as a function of training set size. The mean value of the difference gives the bias of the small-sample estimator, whereas the standard deviation approximates its variance.

3.1. Resubstitution error estimation

The resubstitution error estimates are quite distinct from the other estimators and show an extremely optimistic bias, especially for the simulated spectral data. For the small-sample cases, the resubstitution error is zero for the simulated data (see Fig. 4). The low variance of the resubstitution estimator for the simulated data can also be explained by the low, but extremely optimistic error estimation. For the Diabetes data, bias is still optimistic but less pronounced. The variance of the resubstitution estimator for the Diabetes data is comparable to that of other estimators. The resubstitution estimator underestimates the generalization error to such an extent that disparate models can often have the same resubstitution error. These results suggest that in the small-sample setting the resubstitution

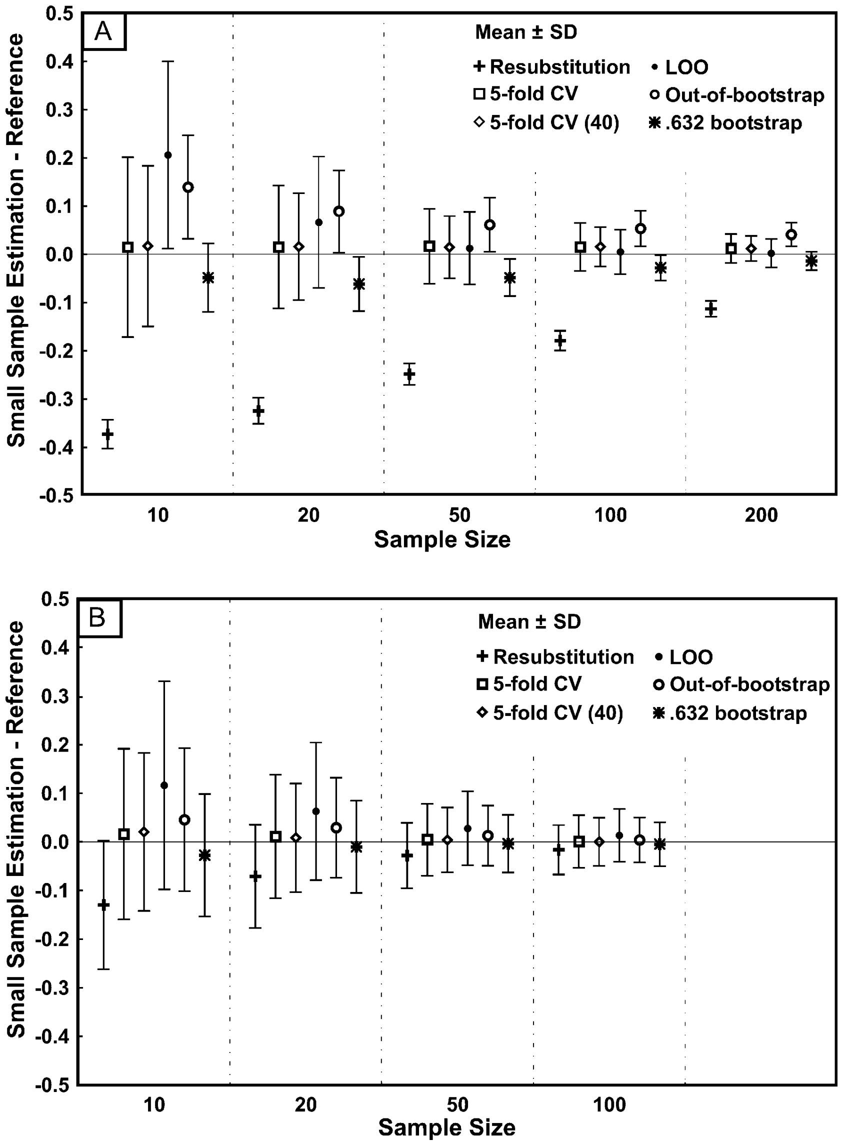

Fig. 3. (A) Performance of small-sample error estimators: simulated data. Small-sample error rate estimation minus the large-sample (reference) error rate estimation as a function of training set size. Differences in the error rates were calculated over 1000 subsets of the data (submodels) and the mean and standard deviation of the differences are reported. The mean and standard deviation respectively provide a measure of the bias and variance of the small-sample estimator. (B) Performance of small-sample error estimators: Diabetes data. Small-sample error rate estimation minus the large-sample (reference) error rate estimation as a function of training set size. Differences in the error rates were calculated over 1000 subsets of the data (submodels) and the mean and standard deviation of the differences are reported. The mean and standard deviation respectively provide a measure of the bias and variance of the small-sample estimator.

error rate estimator should not be used for choosing between different classification models (model selection or determining model complexity) or as a reliable indicator of the model's generalization error.

3.2. LOO CV error estimation

Figs. 3 and 5 show the difference between the reference error and the LOO error, and the LOO error as a function of sample size, respectively. For the smallest sample sizes investigated using the simulated dataset, the mean LOO error estimate was 58%, which is larger than the error expected when simply guessing class membership. Fig. 3 indicates that for small sample sizes, LOO CV suffers from both large variance and a large pessimistic bias. While the large variance of LOO estimation is known, LOO is usually reported as being a reasonably unbiased estimator. A possible reason for this unexpected high bias is the fact that the submodels are built with unequal sample sizes for the two classes and the test sample always belongs to the underrepresented class. This inherent lack of stratification in the LOO strategy leads to the high bias encountered for small sample sizes. Overall, the combined high variance and large pessimistic

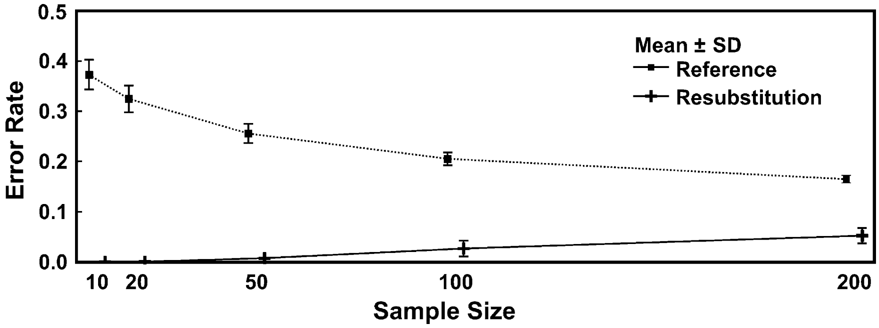

Fig. 4. Resubstitution and reference error rate estimations for the simulated data. The mean and standard deviation of the resubstitution and reference error rates are reported over the 1000 different estimations performed on different subsets of the simulated data. Note that the resubstitution error is essentially zero for small sample sizes.

bias makes LOO error estimators of little value for small-sample size problems. In addition, one cannot reduce the variance of this estimation method.

3.3. k-fold CV error estimation

The *k*-fold estimators are expected to make up for some of the deficiencies of the LOO estimator. While the 5-fold CV estimator was relatively unbiased, the variance turned out to be comparable to that of the LOO estimator. However, unlike with the LOO estimator, further resampling can be carried out to help reduce the variance of the *k*-fold estimator. Using more than one random split of the data and averaging the *k*-fold error estimations over these different random splits, the variance of the estimator should be reduced. Using 40 random splits of the data (a total of 200 submodels) resulted in a somewhat reduced variance for both the simulated and Diabetes data. Such an iterated *k*-fold CV procedure is recommended for datasets with fewer than 100 samples, instead of the conventional *k*-fold CV that uses a single random split. While the iterated *k*-fold CV procedure helped reduce the variance of the error estimation, its observed variance was higher than that of the out-of-bootstrap estimator. The 5-fold CV error estimate showed remarkably low bias, particularly for the smaller training sets (for which the LOO method failed). For large sample sizes with the Diabetes data, only the 0.632 estimator showed smaller bias, and with the synthetic data LOO performed slightly better. Generally the 5-fold CV error estimator was the least biased estimation method.

3.4. Bootstrap-based error estimators

The out-of-bootstrap estimate generally showed more pessimistic bias than the crossvalidation estimates, except for the described *failures* of the LOO CV. The bias of the out-of-bootstrap estimator is partly related to the slight imbalance between submodels, suggesting that the bias could be reduced if both the training and test sets were stratified. Compared to the crossvalidation methods, the variance was lower, particularly with small sample sizes. This estimator showed the lowest variance among the pessimistically biased methods. Only the 0.632 bootstrap estimator showed lower

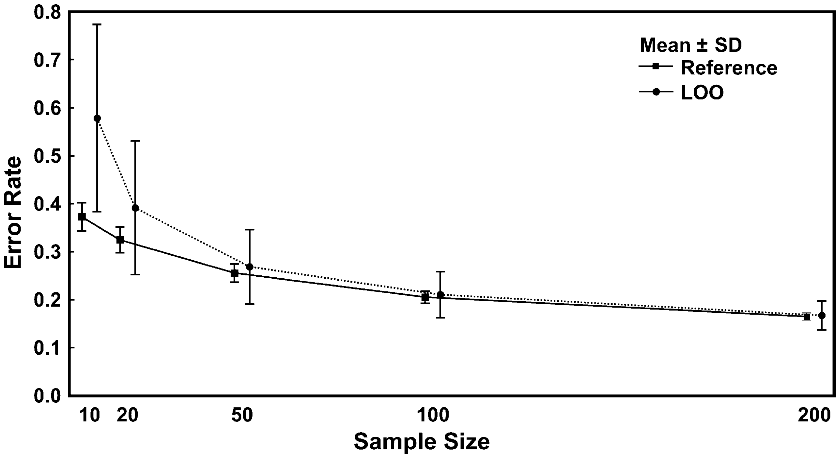

Fig. 5. Leave-one-out and reference error rate estimations for the simulated data. The mean and standard deviation of the LOO and reference error rates are reported over the 1000 different estimations performed on different subsets of the simulated data. Note the poor performance of the leave-one-out estimate (large variance as well as bias) for small sample sizes.

variance, however as noted below, the 0.632 bootstrap estimator should be used with caution. The low variance properties of the out-of-bootstrap error estimator make it attractive for small-sample error estimation.

The 0.632 bootstrap estimator showed lower bias and lower variance than the out-of-bootstrap estimator. The bias was optimistic for both datasets. The resubstitution estimates used to correct for the pessimistic bias of the out-of-bootstrap estimator generally provided little or no information due to their high optimistic bias. In these situations the 0.632 estimate is de facto reduced to a fixed proportion of the out-of-bootstrap estimate.

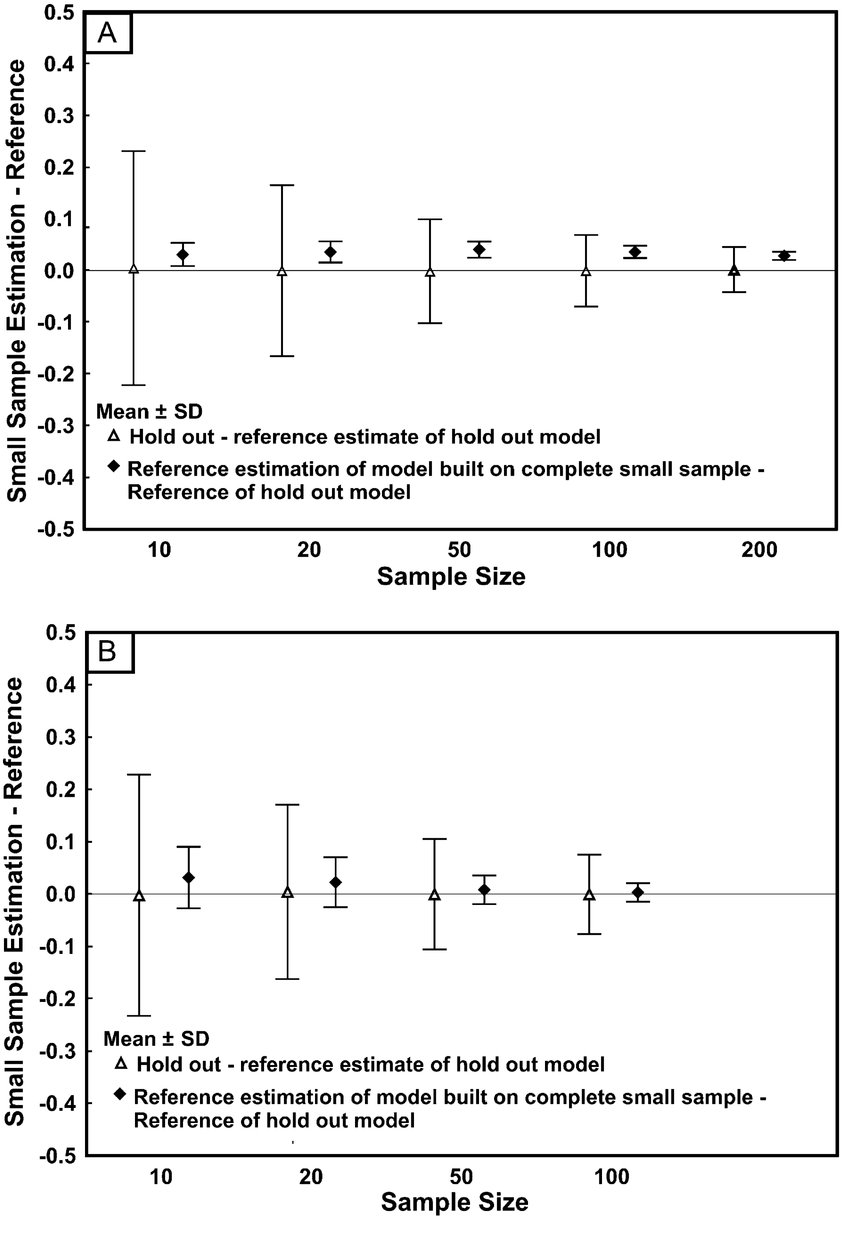

Fig. 6. (A) Performance of small-sample hold-out error estimator: simulated data. Hold-out error rate estimation based on 60:40 training and test set splits of the small sample. Open triangles plot the hold-out error minus the large-sample (reference) error rate estimation as a function of training set size. Differences in the error rates were calculated over 1000 subsets of the data (submodels) and the mean and standard deviation of the differences are reported. The mean and standard deviation respectively provide a measure of the bias and variance of the small-sample estimator. The solid diamond plots compare the difference in error estimates using the large-sample test set with the error calculated for both the hold-out models (built on 60% of the small dataset) and models built on the entire small dataset. These results show that on average the hold-out models are worse than models built on the entire small dataset. (B) Performance of small-sample hold-out error estimator: Diabetes data. Hold-out error rate estimation based on 60:40 training and test set splits of the small sample. Open triangles plot the hold-out error minus the large-sample (reference) error rate estimation as a function of training set size. Differences in the error rates were calculated over 1000 subsets of the data (submodels) and the mean and standard deviation of the differences are reported. The mean and standard deviation respectively provide a measure of the bias and variance of the small-sample estimator. The solid diamond plots compare the difference in error estimates using the large-sample test set with the error calculated for both the hold-out models (built on 60% of the small dataset) and models built on the entire small dataset. These results show that on average the hold-out models are worse than models built on the entire small dataset.

Table 1
Summary of the performance of some common error rate estimation methods for small sample size

	Bias	Variance	Repetitions	Comment
Hold-out	Unbiased	High	N/A	Deprives classifier training and testing from samples, compromising model quality and error estimation reliability.
Resubstitution	Large, optimistic	Low	N/A	Low variance, result of large optimistic bias.
Leave-one-out	Pessimistic, usually low—but fails in certain situations	High	N/A	Bias and variance can be large for small sample sizes, LOO fails.
k-fold	Low, pessimistic	High	Yes, but not commonly implemented	Recommends to use *k*-fold with repetition to further reduce variance.
Out-of-bootstrap	Pessimistic	Low	Yes	Uses repetitions to reduce variance.
0.632 Bootstrap	Low, pessimistic or optimistic	Low	Yes	Overtrained situations: optimistic, reduction in bias and variance based on no information.

3.5. Hold-out estimate

The hold-out estimator differs from the other methods already discussed. For this method, the small dataset is divided into two disjoint parts. One part is used to develop the model and the second part to test the model. Hold-out models were built using 60% of the small dataset and tested using the remaining 40% of the small dataset. These error estimates are referred to as the hold-out error estimator. The reference error used the reserved large independent test sets to calculate the error for the models built on the 60% splits of the small dataset. The open triangles plotted in Fig. 6 show the performance of the hold-out error estimates compared to the reference error estimates where the hold-out error estimates calculated using the 60:40 training and test splits of the small dataset are subtracted from the reference error. As an estimator of the error, the hold-out error estimate is unbiased. This is apparent in Fig. 6 (see open triangles in Fig. 6) where the mean difference between the hold-out error estimates and the large-sample reference error estimates of the hold-out model is essentially zero. However, the variance of the hold-out error estimator was the largest among the discussed error estimators.

The solid diamond plots of Fig. 6 compare error estimates using the large sample test set with the error calculated for both the hold-out models (built on 60% of the small dataset) and models built on the entire small dataset. These results show that on average the hold-out models are worse than models built on the entire small dataset. The data splitting scheme of the hold-out method generally results in inferior models compared to models that use the entire available data. In addition the high variance of the hold-out error estimator can be attributed to the small test sample size. The high variance of the hold-out error estimator makes it particularly undesirable when small sample sizes are employed. Table 1 summarizes the findings from this study.

4. Conclusion

In practice, the feasibility of a classification approach is often assessed with a limited number of well-characterized samples. Usually an estimate of the generalization error of the classifier model is used to make this assessment. Both bias and variance contribute to the uncertainty of the error estimation, yet these characteristics are not apparent in the usual procedure of estimating the generalization error of the classification model. In this manuscript the uncertainty introduced by different estimators of a model's generalization error was examined in terms of bias and variance for small sample sizes.

Inaccurate or biased error estimators exhibit a systematic deviation from the true error while imprecise error estimators exhibit high variance. Except for the resubstitution error estimates, which generally suffered from large, optimistic bias, variance dominated the uncertainty for all of the other estimators examined. Partitioning the small dataset into an independent training and test set, the hold-out method led to unbiased error estimates. However, sacrificing samples with which to build the classification models resulted in poor models. Also, the limited test set yielded an imprecise estimation of the error leading to high variance of the error estimation. The benefit of an unbiased error estimation is grossly outweighed by the high variance of the error estimation when small sample sizes are involved. The commonly used crossvalidation methods, *k*-fold and LOO CV, also exhibited high variance. In addition, LOO estimates showed high bias for extremely small sample sizes. Unlike for *k*-fold CV, the variance of LOO estimators cannot be reduced by further resampling. Stratified 5-fold CV had remarkably low bias, and the variance could be somewhat reduced by repeating the 5-fold error estimation over more than one random split of the data. Repeated *k*-fold CV is a simple extension of the common *k*-fold CV procedure that should be considered in order to reduce the variance of the estimator. The out-of-bootstrap estimates were pessimistically biased, but showed the least variance among the pessimistically biased estimators. The low variance properties of this estimator are particularly attractive for situations where one has a small dataset for which to develop and test a classifier. The 0.632-corrected bootstrap estimator had lower variance than the out-of-bootstrap method. The resubstitution error estimates are near zero, effectively reducing the weighted average used as the 0.632 bootstrap error estimate to the fixed proportion (0.632) of the bootstrap error estimate.

However, the resubstitution estimate was extremely optimistically biased in a small-sample size situation, and therefore caution is advised when using it to correct the out-of-bootstrap error with the 0.632 estimate [11].

Uncertainty in the error estimation can compromise model selection, and can lead to false conclusions about the integrity of the classification approach. A small sample size exacerbates these problems, making the choice of error estimation method that much more critical.

Acknowledgements

The authors would like to thank Elicia Kohlenberg and Nichole Tao for their help in preparing this manuscript.

References

[1] M.R. Chernick, V.K. Murthy, C.D. Nealy, Pattern Recogn. Lett. 3 (1985) 167–178.

[2] M.R. Chernick, V.K. Murthy, C.D. Nealy, Pattern Recogn. Lett. 4 (1986) 133–142.

[3] M.R. Chernick, V.K. Murthy, C.D. Nealy, Comput. Math. Appl. 15 (1988) 29–37.

[4] B. Efron, J. Am. Stat. Assoc. 78 (1983) 316–331.

[5] B. Efron, R. Tibshirani, J. Am. Stat. Assoc. 92 (1997) 548–560.

[6] A.K. Jain, R.C. Dubes, C.-C. Chen, IEEE Trans. Pattern Anal. (1987) 628–633.

[7] R. Kohavi, in: C.S. Mellish (Ed.), Artificial Intelligence. Proceedings. 14th International Joint Conference, 20–25 August 1995, Montréal, Québec, Canada, Morgan Kaufmann, USA, 1995, pp. 1137–1145.

[8] S. Chatterjee, S. Chatterjee, Commun. Stat., Simul. Comput. 12 (1983) 645–656.

[9] H. Schulerud, in: A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquézar, A.K. Jain, J. Kittler (Eds.), Pattern Recognition, Proceedings. 15th International Conference on, Vol. II, 3–7 September 2000, Barcelona, Spain, IEEE Computer Society, USA, 2000, pp. 372–377.

[10] N. Ueda, R. Nakano, Neural Networks, Proceedings of the International Conference, vol. I, 27 Nov.–1 Dec. 1995, Perth, WA, Australia, IEEE, USA, 1995, pp. 101–104.

[11] U.M. Braga-Neto, E.R. Dougherty, Bioinformatics 20 (2004) 374–380.

[12] R.L. Somorjai, A.E. Nikulin, The curse of small sample sizes in medical diagnosis via MR Spectroscopy, Proceedings of Society of Magnetic Resonance in Medicine, Twelfth Annual Scientific Meeting, New York, 14–20 August, 1993, p. 685.

[13] S. Wold, M. Sjöström, L. Eriksson, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130.

[14] M. Baker, W. Rayens, J. Chemom. 17 (2003) 166–173.

[15] S. de Jong, Chemometr. Intell. Lab. Syst. 18 (1993) 251–263.

[16] B. Kowalski, S. Wold, in: P.R. Krishnajah, L.N. Kanal (Eds.), Handbook of Statistics: Classification, Pattern Recognition and Reduction of Dimensionality, vol. II, North-Holland, Amsterdam, 1982, pp. 673–697.

[17] S.J. Raudys, A.K. Jain, Pattern Recognition. Proceedings. 10th International Conference on, vol. I, Atlantic City, New Jersey, USA, 16–21 June, IEEE, USA, 1990, pp. 417–423.

[18] S.J. Raudys, A.K. Jain, IEEE Trans. Pattern Anal. 13 (1991) 252–264.

[19] C.L. Blake, C.J. Merz. UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA, University of California, Department of Information and Computer Science, 1998.

[20] M. Skurichina, R.P.W. Duin, Pattern Recogn. 31 (7) (1998) 909–930.

[21] M. Skurichina, R.P.W. Duin, Pattern Anal. Appl. 5 (2002) 121–135.