Approximation of Unbiased Convex Classification Error Rate Estimator

Mindaugas Gvardinskas

Department of System Analysis, Vytautas Magnus University
Vileikos St. 8, LT-44404 Kaunas, Lithuania
e-mail: m.gvardinskas@if.vdu.lt.

Minija Tamosiunaite

Department for Computational Neuroscience, III Physikalisches Institut, Georg-August-Universitaet
Goettingen, Goettingen D-37077, Germany
e-mail:minija.tamosiunaite@phys.uni-goettingen.de

crossref http://dx.doi.org/10.5755/j01.itc.45.2.12052

Abstract. Convex classification error rate estimator is described as weighted combination of the low-biased estimator and the high-biased estimator. If the underlying data model is known, the coefficients (weights) can be optimized so that the bias and root-mean-square error of the estimator is minimized. However, in most situations, data model is unknown. In this paper we propose a new error estimation method, based on approximation of unbiased convex error rate estimator. Experiments with real world and synthetic data sets show that common error estimation methods, such as resubstitution, repeated 10-fold cross-validation, leave-one-out and random subsampling are outperformed (in terms of root-mean-square error) by the proposed method.

Keywords: Error estimation; Classification; Resubstitution; Cross-validation; Bootstrap.

1. Introduction

Classification error rate estimation is critical for hyperparameter selection [3, 5], feature selection [10] and combining classifiers [20]. This problem has been studied by many researchers and a number of error estimation methods have been proposed [6, 7, 8, 11, 12, 18]. A large part of these techniques are so called error-counting methods. According to this approach, classification error is estimated as the ratio between misclassified data vectors and total number of data vectors. One way to make an error-counting estimator is to use a convex combination of the low-biased estimator and the high-biased estimator. First estimator of this type was proposed by Toussaint and Sharpe [19]. They suggested to choose the coefficients a and b so that the estimator

$$\hat{\varepsilon}_N = a\hat{\varepsilon}_N^{(1)} + b\hat{\varepsilon}_N^{(2)} \qquad (1)$$

is unbiased estimator of true conditional probability of misclassification (conditional PMC). Here $\hat{\varepsilon}_N^{(1)}$ and $\hat{\varepsilon}_N^{(2)}$ are error estimates, the coefficients a and b are nonnegative, $a+b=1$ and N is the training set size. In the absence of either theoretical work or empirical results to guide in the selection of a Toussaint took $a=0.5$. The problem of coefficient selection was further investigated by McLachlan [14]. For the rule based on the Fisher linear discriminant function with zero cutoff point, McLachlan proposed to use theoretically derived coefficients. However, proposed coefficients were functions of the true model parameters. Different approach to coefficient selection was taken by Efron [7]. The coefficients $a=0.632$ and $b=0.368$ were suggested by an argument based on the fact that in 0.632 bootstrap, expected number of distinct points from the original data set appearing in the training set is approximately 0.632*N*. However, it was demonstrated that 0.632 bootstrap fails to give good estimates when Bayes error is very high [17].

In this paper, we propose a new estimator of the error rate of the Euclidean distance classifier, based on approximation of unbiased convex error rate estimator. In order to demonstrate the effectiveness of the new method we compare it with other common error counting methods, such as 0.632 bootstrap, 10-fold cross-validation, resubstitution and random subsampling.

This paper is organized as follows. Section 2 describes common error counting estimators. Section 3 introduces the new error estimation method. Section 4 is devoted to experimental analysis. Section 5 contains concluding remarks.

2. Error estimation

2.1. Basic definitions

Consider two category classification problem where class label $\omega \in \{0,\ 1\}$, feature vector $\mathbf{x} \in R^n$ and a classifier is a function f: $R^n \to \{0,\ 1\}$. An induction algorithm builds a classifier from a set of N independent observations $D_N = \{(\mathbf{x}_1, \omega_1), \ldots, (\mathbf{x}_N, \omega_N)\}$ drawn from some distribution T. Formally, it is a mapping g: $\{R^n \times \{0,\ 1\}\}^N \times R^n \to \{0,\ 1\}$. The performance of a classifier is measured by conditional probability of misclassification:

$$\varepsilon_N = P(g(D_N, \mathbf{x}) \neq \omega)\ . \qquad (2)$$

This error is conditioned on one particular training set D_N and induction algorithm g. If the underlying distribution is known, one can calculate conditional PMC exactly. However, in practice, this distribution is unknown and an error estimator $\hat{\varepsilon}_N$ is needed.

2.2. Resubstitution

Resubstitution error can be used as an estimate of conditional PMC. The resubstitution estimated error is defined as:

$$\hat{\varepsilon}_N^{(R)} = \frac{1}{N}\sum_{i=1}^{N} |\, g(D_N, \mathbf{x}_i) - \omega_i \,| \qquad (3)$$

This estimator is optimistic (i.e. low-biased), especially for small data sets.

2.3. Cross-validation

In k-fold cross-validation, the data set is randomly partitioned into k subsets of approximately equal size. Each subset is used as a test set and the remaining k-1 subsets are used as the training set. The cross-validation error estimate is defined as:

$$\hat{\varepsilon}_N^{(CV)} = \frac{1}{N}\sum_{i=1}^{k} \sum_{j=1,(\mathbf{x}_j,\omega_j)\in D_i}^{N} |\, g(D_N \setminus D_i, \mathbf{x}_j) - \omega_j \,| \qquad (4)$$

where D_i is the i-th fold of the data set D_N, k is the number of folds and N is the size of D_N. Leave-one-out estimator is a special case of k-fold cross-validation with k equal to the size of the original data set. In stratified cross-validation, each of the k subsets contains approximately the same proportion of class labels as the original data. Kohavi recommends stratified 10-fold cross-validation as the best error estimation method [9]. However, cross-validation has large variance, especially for small sample data sets [2]. One way to reduce variance is to repeat cross-validation procedure several times [2].

2.4. Holdout

In holdout method, the data set is randomly split into two parts: one is used as the training set and the other as the test set. It is common to allocate two thirds of the data as the training set and the remaining one third as the test set [9]. The holdout estimate is pessimistic (i.e. high-biased) since only a portion of the initial data is used to train the classifier. The holdout estimate of the conditional PMC of a classifier is defined as:

$$\hat{\varepsilon}_N^{(H)} = \frac{1}{h}\sum_{i=1,(\mathbf{x}_i,\omega_i)\in D_h}^{N} |\, g(D_N \setminus D_h, \mathbf{x}_i) - \omega_i \,| \qquad (5)$$

where $D_h \subset D_N$ is the holdout set (the test set) of size h. In random subsampling, the holdout method is repeated r times and the estimated error is derived by averaging the runs [9].

2.5. Bootstrap

Bootstrap sampling concept was introduced by Efron [6]. A bootstrap sample is formed by sampling N data points uniformly and with replacement from the original data set. On average it contains 0.632N of the original data. In 0.632 bootstrap method [7], an induction algorithm is trained on the bootstrap sample and the resubstitution error estimate $\hat{\varepsilon}_N^{(R)}$ is found. The rest of the data are used for classifier testing. The 0.632 bootstrap estimated error is defined as

$$\hat{\varepsilon}_N^{(B)} = \frac{1}{r}\sum_{i=1}^{r} (0.632 \cdot \hat{\varepsilon}_N + 0.368 \cdot \hat{\varepsilon}_N^{(R)}) \qquad (6)$$

where r is the number of bootstrap samples, $\hat{\varepsilon}_N^{(R)}$ is resubstitution error on the bootstrap data and $\hat{\varepsilon}_N$ is error estimate for bootstrap sample i.

2.6. Performance of error estimators

Commonly used performance measures of an error estimator $\hat{\varepsilon}_N$ are the bias, deviation variance and root-mean-square error (RMS) [2, 4, 17]:

$$Bias[\hat{\varepsilon}_N] = E[\hat{\varepsilon}_N] - E[\varepsilon_N] \qquad (7)$$

$$\begin{aligned} Var_{dev}[\hat{\varepsilon}_N] &= Var(\hat{\varepsilon}_N - \varepsilon_N) = Var(\hat{\varepsilon}_N) + Var(\varepsilon_N) \\ &- 2Cov(\hat{\varepsilon}_N, \varepsilon_N) \end{aligned} \qquad (8)$$

$$\begin{aligned} RMS[\hat{\varepsilon}_N] &= \sqrt{E[(\varepsilon_N - \hat{\varepsilon}_N)^2]} \\ &= \sqrt{E[\varepsilon_N{}^2] + E[\hat{\varepsilon}_N{}^2] - 2E[\varepsilon_N \hat{\varepsilon}_N]} \end{aligned}\ . \qquad (9)$$

The most important performance measure is RMS, because it combines bias and the deviation variance into a single metric.

3. Proposed method

3.1. Basic expressions

Expected error of the Euclidean distance classifier is expressed as [15, 16]

$$E[\varepsilon_N] \approx \Phi\left\{-\frac{\delta}{2}\frac{1}{\sqrt{T_M}}\right\} \tag{10}$$

and expected resubstitution error is given by

$$E[\varepsilon_N^R] \approx \Phi\left\{-\frac{\delta}{2}\sqrt{T_M}\right\} \tag{11}$$

where Φ is a standard Gaussian cumulative distribution function, $\delta = \sqrt{(\mathbf{M}_1 - \mathbf{M}_2)^T \mathbf{\Sigma}^{-1}(\mathbf{M}_1 - \mathbf{M}_2)}$ is a Mahalanobis distance between two pattern classes, $\mathbf{M}_1$ and $\mathbf{M}_2$ are class mean vectors, $\mathbf{\Sigma}^{-1}$ is the inverse of the common covariance matrix, $T_M = 1 + \frac{4n}{\delta^2 N}$.

3.2. Approximation of unbiased convex classification error rate estimator

Unbiased convex error rate estimator can be expressed as:

$$Bias[\hat{\varepsilon}_N] = a\ E[\hat{\varepsilon}_N^{(1)}] + b\ E[\hat{\varepsilon}_N^{(2)}] - E[\varepsilon_N] = 0\ .(12)$$

If estimator $\hat{\varepsilon}_N^{(1)}$ is repeated 2-fold cross-validation and estimator $\hat{\varepsilon}_N^{(2)}$ is repeated half sample resubstitution then $E[\hat{\varepsilon}_N^{(1)}] \approx E[\varepsilon_{0.5N}]$, $E[\hat{\varepsilon}_N^{(2)}] \approx E[\varepsilon_{0.5N}^R]$ and bias of the convex error rate estimator can be written as:

$$Bias[\hat{\varepsilon}_N] \approx a\ E[\varepsilon_{0.5N}] + b\ E[\varepsilon_{0.5N}^R] - E[\varepsilon_N] \approx 0\ .(13)$$

From (13) and the fact that $a + b = 1$ we have

$$a \approx \frac{E[\varepsilon_N] - E[\varepsilon_{0.5N}^R]}{E[\varepsilon_{0.5N}] - E[\varepsilon_{0.5N}^R]} \tag{14}$$

$$b \approx \frac{E[\varepsilon_N] - E[\varepsilon_{0.5N}]}{E[\varepsilon_{0.5N}^R] - E[\varepsilon_{0.5N}]}\ . \tag{15}$$

Now, suppose that the following preconditions are met:

1. Classifier deals with two multivariate Gaussian pattern classes;
2. the covariance matrix is the same for all classes;
3. the covariance matrix is proportional to the identity matrix, i.e. $\mathbf{\Sigma} = \sigma^2\mathbf{I}$;
4. class prior probabilities are equal;
5. the training set has the same number of patterns from each class;
6. Mahalanobis distance is constant;
7. the dimensionality n is fixed and very large;
8. training set size $N \rightarrow \infty$.

Then from (10) and (11) we get that the coefficients are

$$a \approx \lim_{\frac{n}{N}\to 0} \frac{E[\varepsilon_N] - E[\varepsilon_{0.5N}^R]}{E[\varepsilon_{0.5N}] - E[\varepsilon_{0.5N}^R]} \approx 0.75 \tag{16}$$

$$b \approx \lim_{\frac{n}{N}\to 0} \frac{E[\varepsilon_N] - E[\varepsilon_{0.5N}]}{E[\varepsilon_{0.5N}^R] - E[\varepsilon_{0.5N}]} \approx 0.25\,. \tag{17}$$

Note that the derivation of coefficient values is based on the Taylor series expansion of $E[\varepsilon_N]$, $E[\varepsilon_{0.5N}^R]$ and $E[\varepsilon_{0.5N}]$. The proposed convex error rate estimator (PCE) is defined as:

$$\hat{\varepsilon}_N^{(PCE)} = \frac{1}{r}\sum_{i=1}^{r}(a \cdot \hat{\varepsilon}_N^{(CV)} + b \cdot \hat{\varepsilon}_N^{(R)}) \tag{18}$$

where r is the number of repetitions, $\hat{\varepsilon}_N^{(CV)}$ is 2-fold cross-validation error estimate, and $\hat{\varepsilon}_N^{(R)}$ is the resubstitution error estimate.

4. Simulation study

4.1. Experimental setup

Our experiments consist of two parts: synthetic experiments with Gaussian data and experiments with real world data sets. The error estimators studied are repeated 10-fold cross-validation, leave-one-out, 0.632 bootstrap, random subsampling (*h*=*N*-0.7*N*), resubstitution, proposed convex estimator and optimal unbiased estimator. To get a fair comparison, we made the number of classifiers built for each estimator equal, i.e. 320 (except resubstitution and leave-one-out estimators where the number of induced classifiers is fixed and cannot be changed). Therefore, in random subsampling and 0.632 bootstrap the number of runs (*r*) is set to 320, in proposed convex estimator and optimal unbiased estimator the number of runs is set to 160, in repeated 10-fold cross-validation the number of runs is set to 32. The coefficients of optimal unbiased estimator are computed using expressions (10), (11), (14), (15), and this estimator is used only in the case of Gaussian data with equal class prior probabilities.

4.2. Synthetic data

Our set of synthetic simulations is composed of 96 experiments. In all cases, we use two-class Gaussian data model with common identity covariance matrix and class means located at $\mathbf{M}_1 = (m,\ m,...,m)^T$ and $\mathbf{M}_2 = (-m,\ -m,...,-m)^T$. The class prior probabilities are: 1) P_1 = 0.5, P_2 = 0.5; 2) P_1 = 0.6, P_2 = 0.4; 3) P_1 = 0.7, P_2 = 0.3; 4) P_1 = 0.8, P_2 = 02. For each pair of P_1 and P_2, we choose four values of *m* such that Bayes error is from 0.05 to 0.20. In each of these sixteen cases, 10000 independent samples of size N=20, N=40, N=60, N=80, N=100, N=120 are generated (in all cases n=10).

Fig. 1-12 display results of synthetic simulations (where P_1=0.5, P_2=0.5, n=10). The experiments show that resubstitution, repeated 10-fold cross-validation, leave-one-out and random subsampling are outperformed by the proposed convex estimator (in RMS sense). However, the situation with the convex estimators is different. When $\varepsilon_{Bayes} = 0.20$ and N=20, the best error estimation method is the proposed method, in all other cases all three convex estimators perform similarly. The experiments also show that repeated 10-fold cross-validation, leave-one-out and random subsampling are more variable than optimal unbiased estimator, proposed method, 0.632 bootstrap and resubstitution. Among the convex estimators considered, 0.632 bootstrap is least variable, whereas the proposed convex estimator is more variable, but it displays less bias. Additionally, resubstitution, random subsampling and 0.632 bootstrap are more biased than repeated 10-fold cross-validation, leave-one-out, proposed estimator and optimal unbiased estimator. Also, when N=20, the proposed method is more biased than optimal unbiased estimator and when N > 20, both convex estimators perform almost identically.

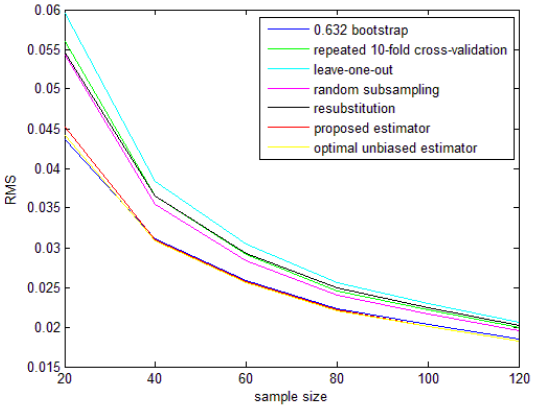

Figure 1. RMS results, ε_{Bayes} = 0.05

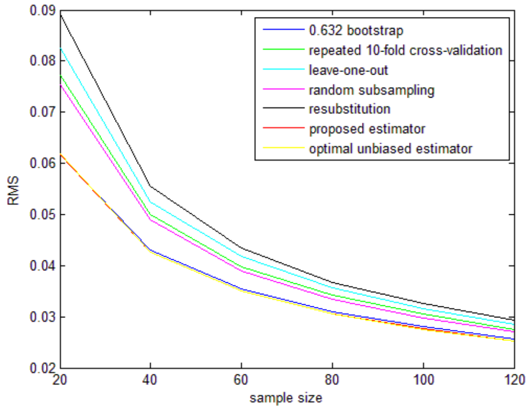

Figure 2. RMS results, ε_{Bayes} = 0.1

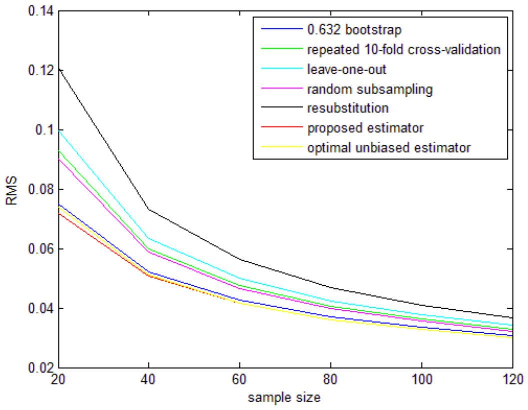

Figure 3. RMS results, ε_{Bayes} = 0.15

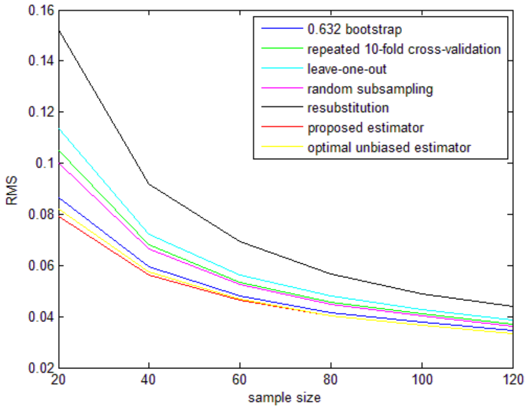

Figure 4. RMS results, ε_{Bayes} = 0.2

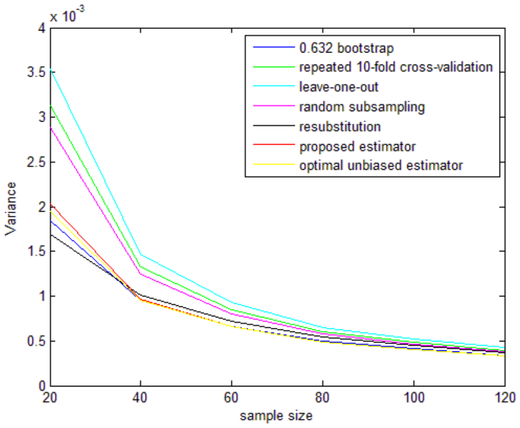

Figure 5. Variance results, ε_{Bayes} = 0.05

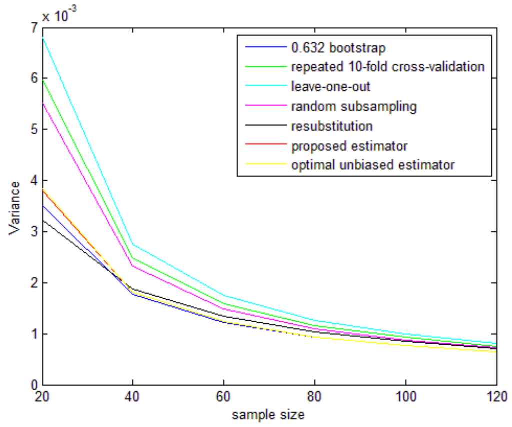

Figure 6. Variance results, ε_{Bayes} = 0.1

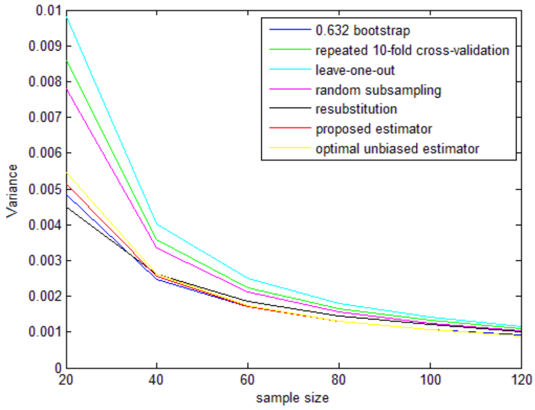

Figure 7. Variance results, $\varepsilon_{\text{Bayes}}$ = 0.15

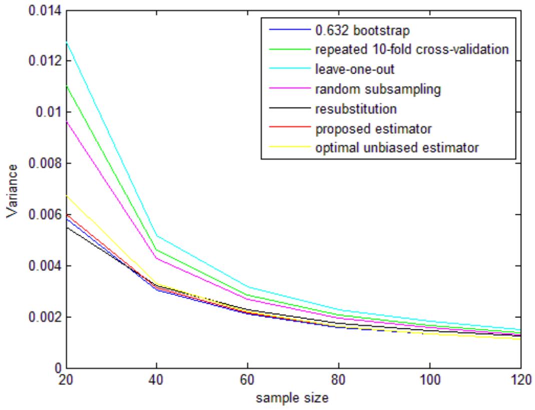

Figure 8. Variance results, $\varepsilon_{\text{Bayes}}$ = 0.2

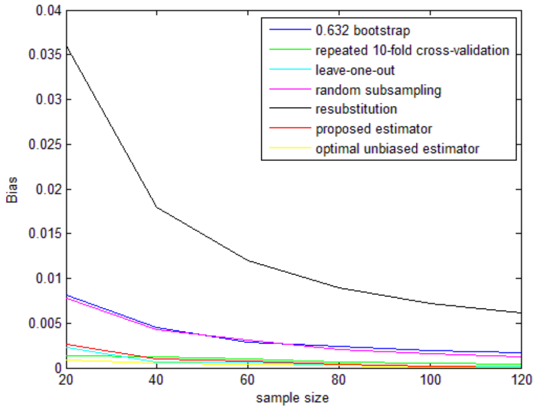

Figure 9. Bias results (absolute values), $\varepsilon_{\text{Bayes}}$ = 0.05

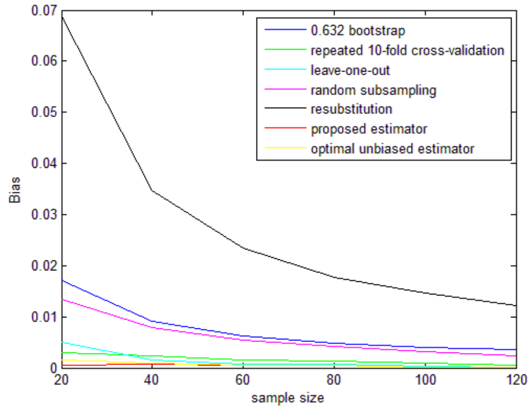

Figure 10. Bias results (absolute values), $\varepsilon_{\text{Bayes}}$ = 0.1

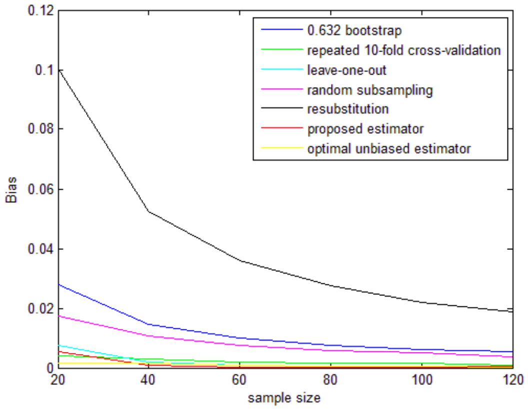

Figure 11. Bias results (absolute values), $\varepsilon_{\text{Bayes}}$ = 0.15

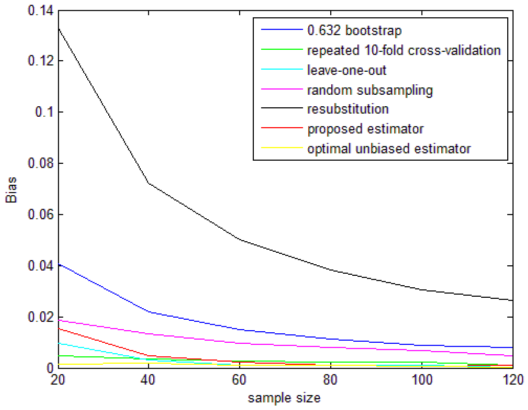

Figure 12. Bias results (absolute values), $\varepsilon_{\text{Bayes}}$ = 0.2

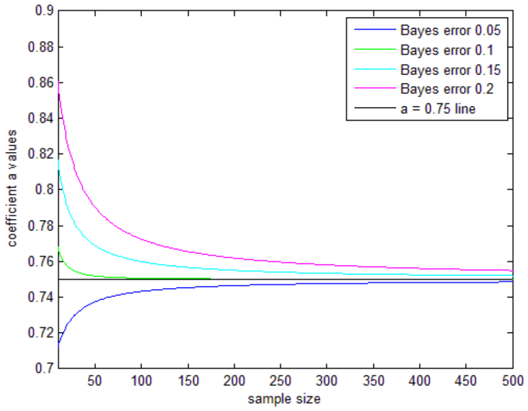

Figure 13. Theoretical coefficients

The RMS and variance results for P_1 > 0.5 (not shown here explicitly) are similar to those for P_1 = 0.5. However, the situation with bias is slightly different. When P_1 = 0.8, the proposed convex estimator is better than resubstitution and 0.632 bootstrap, but worse than repeated 10-fold cross-validation, leave-one-out and random subsampling. Analogical situation is when P_1 = 0.6 or P_1 = 0.7 and $\varepsilon_{Bayes} = 0.20$. In the remaining cases, the bias of the proposed convex estimator is similar to the bias of leave-one-out.

Fig. 13 shows theoretical coefficients for different Bayes errors and sample sizes when $P_1 = 0.5$, $P_2 = 0.5$ and $n = 10$. It is clear that in all cases theoretical coefficients converge to 0.75 as sample size increases. However, the rate of convergence for different Bayes error rates is different. Note that the coefficients are computed using expressions (10), (11) and (14).

4.3. Real data

The experiments are designed in the following way: the data set is randomly split into two sets and one set is used for both, classifier training and error estimation, while the other (much larger set) is used to approximate conditional PMC. This procedure is repeated 10000 times. Real world experiments were conducted using the following data sets:

Pima Indian Diabetes data [1]. It consists of 768 instances that are diabetes positive (268) or diabetes negative (500). The number of features is 8. We consider three training/error estimation samples of size 32, 40 and 48.

QSAR biodegradation database [13]. This data set describes 1055 molecules that are ready biodegradable (356) or not ready biodegradable (699). The number of features is 41. Training/error estimation sample size is 60, 80, 100 and 120.

Spambase data set [1]. This datebase is composed of 4601 instances of which 1813 are spam and 2788 are non-spam. The number of features is 57. We use three training/error estimation samples of size 60, 80 and 100.

Banknote authentication data [1]. This data set consists of 1372 instances of which 762 are classified as genuine and 610 are classified as forged. The number of attributes is 4. Training/error estimation sample size is 20, 30 and 40.

Fig. 14-25 show the experimental results of all six error estimation methods on four real world data sets. Here we can see that the proposed method outperforms all non-convex error counting estimators (in RMS sense). The closest competitor to the proposed method is 0.632 bootstrap. This method performs better in Banknote authentication data set and is comparable to the proposed convex estimator in Spambase

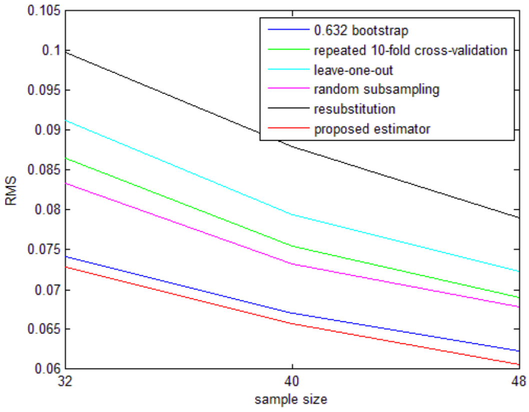

Figure 14. RMS results, Pima Indian Diabetes data

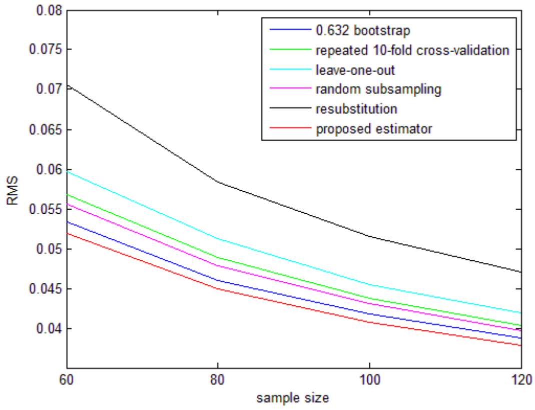

Figure 15. RMS results, QSAR biodegradation data

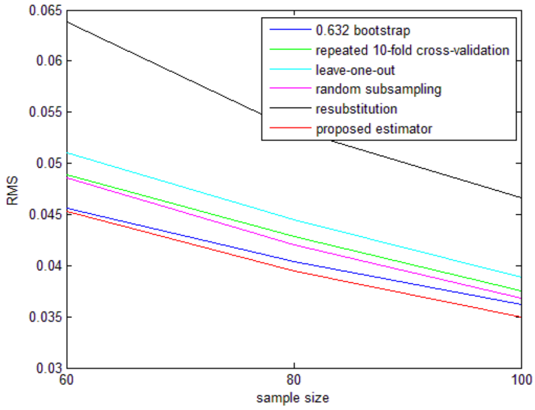

Figure 16. RMS results, Spambase data

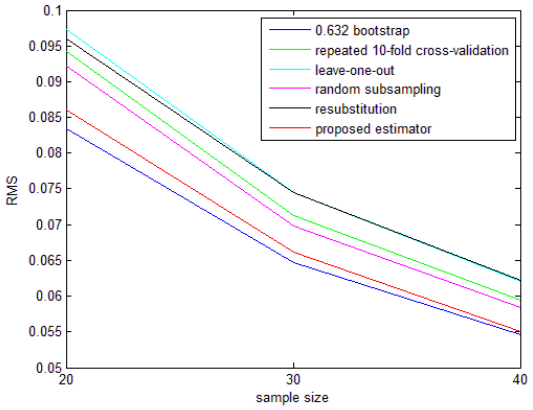

Figure 17. RMS results, Banknote authentication data

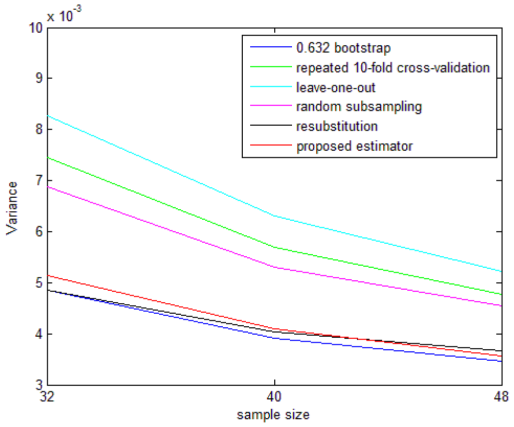

Figure 18. Variance results, Pima Indian Diabetes data

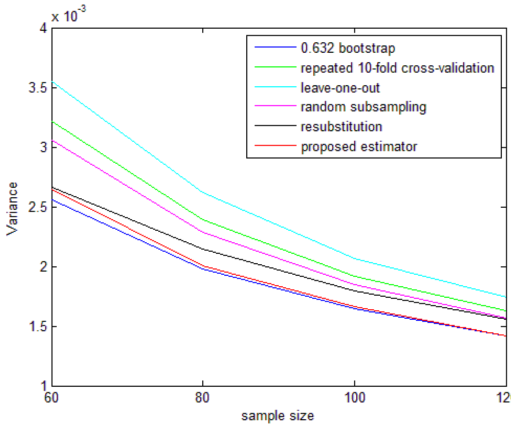

Figure 19. Variance results, QSAR biodegradation data

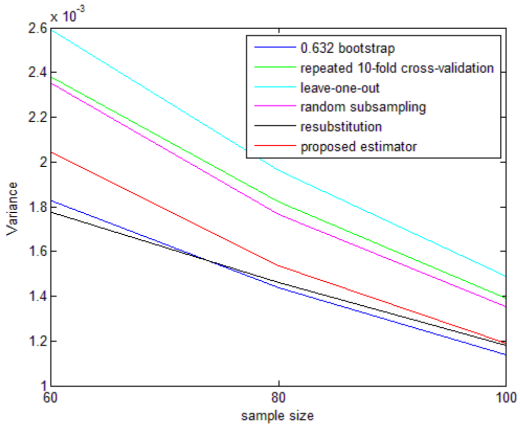

Figure 20. Variance results, Spambase data

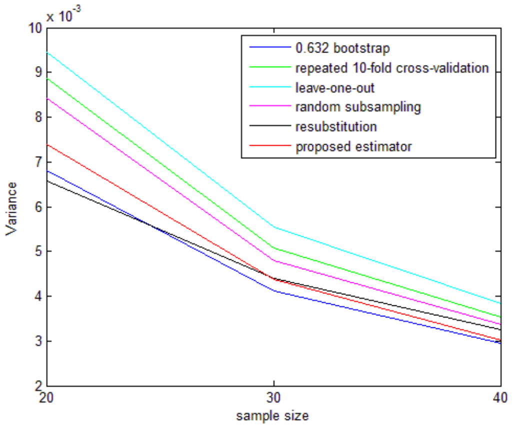

Figure 21. Variance results, Banknote authentication data

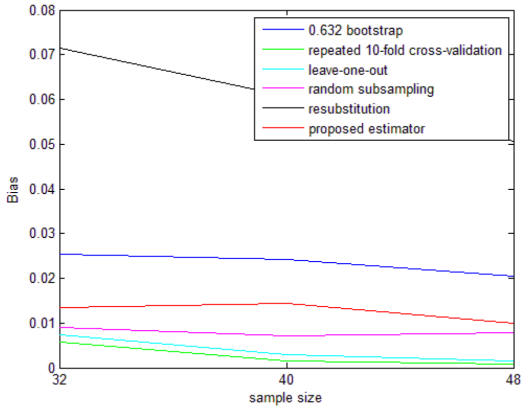

Figure 22. Bias results, Pima Indian Diabetes data

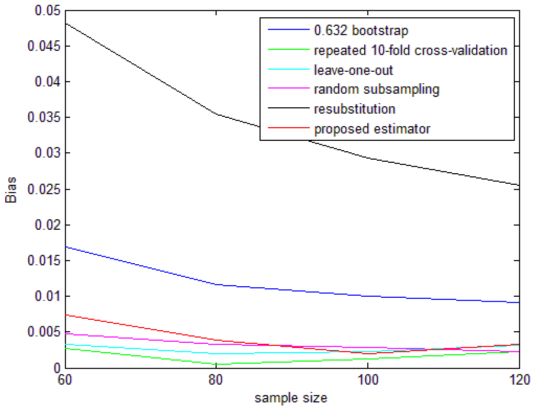

Figure 23. Bias results, QSAR biodegradation data

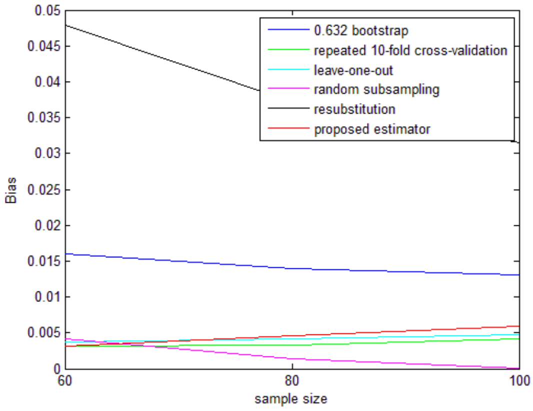

Figure 24. Bias results, Spambase data

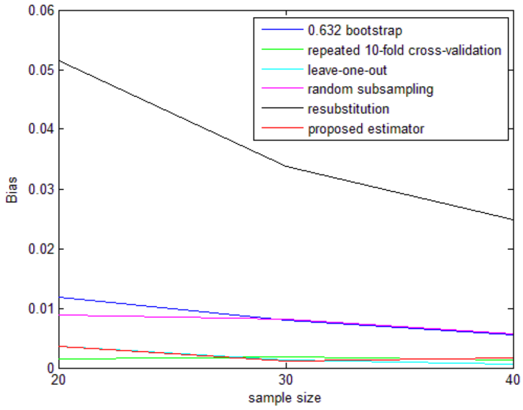

Figure 25. Bias results, Banknote authentication data

data set. In all other cases, the proposed method performs better than 0.632 bootstrap. The experiments also show that the variance of the proposed convex error rate estimator lies between the variances of 0.632 bootstrap and random subsampling. Additionally, the proposed method is less biased than 0.632 bootstrap and resubstitution.

5. Conclusion

In this paper, we have proposed a new convex error estimation method, which approximates unbiased convex classification error rate estimator. Experiments with real world and synthetic data sets show that resubstitution, repeated 10-fold cross-validation, leave-one-out and random subsampling are outperformed by the proposed convex estimator (in RMS sense). The closest competitor to the proposed convex estimator is 0.632 bootstrap, however, contrary to the proposed method, bootstrap is more biased.

References

[1] **K. Bache, M Lichman.** UCI Machine Learning Repository [Online]. Available at: http://archive.ics.uci.edu/ml.

[2] **U. Braga-Neto, E. Dougherty.** Is cross-validation valid for small sample microarray classification? *Bioinformatics*, 2004, Vol. 20, No. 3, 374-380.

[3] **O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee.** Choosing multiple parameters for support vector machines. *Machine Learning,* 2002, Vol. 46, Issue 1, 131–159.

[4] **E. Dougherty, C. Sima., J. Hua., B. Hanczar, U. Braga-Neto**. Performance of error estimators for classification. *Current Bioinformatics*, 2010, Vol. 5, No. 1, 53-67.

[5] **G. Dzemyda, L. Sakalauskas**. Large-Scale Data Analysis Using Heuristic Methods. *Informatica*, 2011, Vol. 22, No. 1, 1-10.

[6] **B. Efron.** Bootstrap Methods: Another look at the jackknife. *Annals of Statistics*, 1979, Vol. 7, No. 1, 1-26.

[7] **B. Efron.** Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association,* 1983, Vol. 78, No. 382, 316–331.

[8] **B. Efron, R. Tibshirani.** Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association,* 1997, Vol. 92, Vol. 438, 548-560.

[9] **R. Kohavi.** A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995, pp. 1137-1143.

[10] **R. Kohavi, G. John.** Wrappers for feature selection. *Artificial Intelligence,* 1997, Vol. 97, Issue 1-2, 273-324.

[11] **W.J. Krzanowski, D.J. Hand.** Assessing error rate estimators: the leaving-one-out method reconsidered. *Australian Journal of Statistics*, 1997, Vol. 39, No. 1, 35-46.

[12] **P. Lachenbruch, R. Mickey.** Estimation of error rates in discriminant analysis. *Technometrics*, Vol. 10 (1), 1968, pp. 1-11.

[13] **K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, V. Consonni**. Quantitative Structure - Activity Relationship models for ready biodegradability of chemicals. *Journal of Chemical Information and Modeling*, 2013, Vol. 53, No. 4, 867-878.

[14] **G.J. McLachlan**. A note on the choice of a weighting function to give an efficient method for estimating the probability of misclassification. *Pattern Recognition*, 1977, Vol. 9, 147-149.

[15] **S. Raudys.** Statistical and Neural Classifiers, An Integrated Approach to Design. *Springer-Verlag, London*, 2001.

[16] **S. Raudys, D. M. Young.** Results in statistical discriminant analysis: A review of the former Soviet Union literature. *Journal of Multivariate Analysis*, 2004, Vol. 89, Issue 1, 1–35.

[17] **C. Sima, E. Dougherty.** Optimal convex error estimators for classification. *Pattern Recognition*, 2006, Vol. 39, No. 9, 1763-1780.

[18] **C. Smith.** Some examples of discrimination. *Annals of Eugenics*, 1947, Vol. 18, 272–282.

[19] **G. Toussaint, P. Sharpe.** An efficient method for estimating the probability of misclassification applied to a problem in medical diagnosis. *Computers in Biology and Medicine*, Vol. 4, 1975, 269–278.

[20] **D.H. Wolpert.** Stacked generalization. *Neural Networks*, 1992, Vol. 5, No. 2, 241-259.

Received April 2015.