Discussion of “Adaptive Confidence Intervals for the Test Error in Classification”, by E.B. Laber and S.A. Murphy

Susan Wei* and Andrew B. Nobel†

March 28, 2011

1 Introduction

Laber and Murphy [8] address the important and interesting question of how to accurately assess the performance of a classifier that is produced from a fixed data set. Performance is measured by the conditional test error of the classifier, that is, the probability that the given classifier will mislabel a future observation drawn from the same distribution as the training data. (In what follows, we will use the terms conditional test error and test error interchangeably.) In practical situations, where the distribution of the data is unknown, the ideal way to assess test error is by means of a sufficiently large test set that is independent of the training set used to produce the classifier. The paper focuses on problems in which the size of the training data set is small, and obtaining independent test samples is impossible, or impractical.

Laber and Murphy advocate the use of interval estimates rather than point estimates of the test error. In this regard, it should be mentioned that there are a number of good point estimates for the *unconditional* test error of a classification procedure, such as K-fold cross validation. There are also good point estimators for the *conditional* test error, such as the .632+ bootstrap estimator of Efron and Tibshirani [3]. In either case, using point estimates to produce confidence intervals for the conditional test error is problematic. Using a point estimator for the unconditional test error as a point estimate of the conditional test error is

*Susan Wei is with the Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260. Corresponding author. Email: susanwe@email.unc.edu

†Andrew Nobel is with the Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260. Corresponding author. Email: nobel@email.unc.edu

ill-advised if the latter has high variance, compounding the already difficult task of interval construction (cf. Nadeau and Bengio [9], Jiang et al. [5] and Braga-Neto and Dougherty [2]). On the other hand, point estimators of the conditional test error often involve many layers of randomness and dependence, making it difficult to assess their variance. For instance, the authors in [3] point out it is difficult to obtain a standard error estimate for the .632+ estimator and do not study it.

The problem of forming a confidence interval for the test error of a given classifier differs from more routine interval estimation problems, in that the parameter of interest is itself random, varying from training set to training set. The need for improved interval estimates is illustrated in [8] by a simulation study, and the analysis of several real data sets. It is a central thesis of the paper that the failure of existing confidence interval procedures, such as the centered percentile bootstrap and normal approximation, is due to their neglect of the excess variation resulting from the randomness of the test error. The adaptive confidence interval (ACI) method proposed by Laber and Murphy attempts to address the effects of added variation. Roughly speaking, the ACI divides the available data points into those that are close to, and those that are far from, the optimal linear decision boundary. Upper and lower confidence bounds are then generated by considering the largest and smallest misclassification rates of the points that are close to the boundary. The empirical results of the paper suggest that the ACI is a good alternative to existing confidence interval procedures.

It is worth noting that Laber and Murphy do not solve the general problem of finding confidence intervals for the conditional test error of classification rules produced by an arbitrary classification procedure. Indeed, their method and results are restricted to the case of procedures that produce a linear classification rule by minimizing a convex loss function. While this setting is general enough to include the important special case of support vector machines (SVM) with linear kernels, it does not include popular classification procedures such as nearest centroid, nearest neighbor, and decision trees. Moreover, linear classifiers are often better suited to high dimensional problems, a setting which the ACI cannot currently handle. In low dimensional problems, linear classifiers may lack the flexibility to capture the potentially complicated geometry of Bayes decision boundaries. Nevertheless, the ACI method provides a promising first step towards a solution to the general problem. Developing related methods for other low- and high-dimensional classification procedures would be of both practical and theoretical interest.

1.1 Overview

The next section outlines the technical framework of the confidence interval problem and briefly reviews the centered percentile bootstrap. An overview and discussion of the ACI is given in Section 3. Section 4 outlines a modification of the ACI while Section 5 presents an extension of the ACI to high dimensional low sample settings. A preliminary simulation study comparing these methods with the ACI and CPB is given in Section 6.

2 Preliminaries

2.1 Technical Framework

Let $\mathcal{T}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a training set whose elements $(X_i, Y_i) \in \mathbb{R}^p \times \{-1, 1\}$ are independent and identically distributed with distribution P, and let $P_n = n^{-1} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$ denote the empirical measure associated with $\mathcal{T}_n$. The ACI method is built around a classification procedure that minimizes the P_n-expectation of a loss-type function $L(x, y, \beta)$ over vectors β in $\mathbb{R}^p$. Formally, $L(x, y, \beta)$ is, for each x, y, a convex function of β. In practice, it is a convex upper bound on the 0/1-loss $L'(x, y, \beta) = I(x^t \beta y < 0)$ of a linear classifier $\text{sign}(x^t \beta)$ associated with the normal vector β. Let

$$\hat{\beta}_n = \arg\min_{\beta \in \mathbb{R}^p} P_n L(X, Y, \beta) \qquad \beta^* = \arg\min_{\beta \in \mathbb{R}^p} P L(X, Y, \beta).$$

be the directions that minimize the empirical L-loss and expected L-loss, respectively. The test error is the probability $\tau(\hat{\beta}_n) := PI(YX^t \hat{\beta}_n < 0)$ that the classifier $\text{sign}(x^t \hat{\beta}_n)$ will mislabel a new feature. The test error depends on the training set $\mathcal{T}_n$, and is therefore random. For fixed $\alpha \in (0, 1)$, upper and lower confidence bounds correspond to functions $\hat{a}$ and $\hat{b}$ of the training data $\mathcal{T}_n$ such that

$$\mathbb{P}(\hat{a} \leq P\mathbb{I}\{YX^t \hat{\beta}_n\} \leq \hat{b}) = 1 - \alpha,$$

where $\mathbb{P}$ denotes expectation with respect to training data. It is evident from the last display that the bounds $\hat{a}$ and $\hat{b}$ should account for the variance of the test error.

2.2 The Centered Percentile Bootstrap

A natural starting point for an interval estimate of the test error is the centered percentile bootstrap (CPB) confidence interval. The CPB interval is formed by bootstrapping the quantity

$$\sqrt{n}(P_n - P)\, \mathbb{I}\{X^t \hat{\beta}_n Y < 0\}. \tag{1}$$

Let $\hat{u}$ and $\hat{l}$ be, respectively, the $1-\alpha/2$ and $\alpha/2$ percentiles of the numbers

$$\sqrt{n}(P_n^b - P_n)\,\mathbb{I}\{X^t\hat{\beta}_n^b Y < 0\}, \quad b = 1, \ldots, B, \tag{2}$$

where b indexes bootstrap samples of the training set. The $1-\alpha$ CPB confidence interval is given by

$$[\,P_n\mathbb{I}\{X^t\hat{\beta}_n Y < 0\} - \hat{u}/\sqrt{n},\; P_n\mathbb{I}\{X^t\hat{\beta}_n Y < 0\} - \hat{l}/\sqrt{n}\,]. \tag{3}$$

In their empirical studies, Laber and Murphy demonstrate that the CPB confidence intervals, and confidence intervals based on a normal approximation of the test error, exhibit marked under coverage (anti-conservative behavior) in small samples. Clearly the performance of these methods will improve as sample size increases, but for small samples the CPB fails to capture the additional variation in the test error due to the non-smoothness of the 0-1 loss. Through simulations, Laber and Murphy demonstrate the improved performance of the CPB that can result from replacing the 0-1 loss with a smooth surrogate.

3 The ACI Method

3.1 Boundary Points

The first step in constructing the ACI is identifying points in the training set that are close to the decision boundary $x^t\beta^* = 0$ of the linear classifier minimizing the expected L-risk. To do this, the ACI method performs, for each covariate vector X_i, a test of the hypothesis $X_i^t\beta^* = 0$. The test accepts when X_i is contained in the set $B_n = \{x : (x^t\hat{\beta}_n)^2 \leq a_n^{-1}x^t\Sigma x\}$, where $a_n = o(n)$ is a fixed threshold, and Σ is (an estimate of) the asymptotic covariance matrix of $\hat{\beta}_n$.

Figure 1 shows the results of a simple simulation in $\mathbb{R}^5$, in which the class conditional distributions are Gaussian with unit covariance, and differ only in the first coordinate of their mean vectors, +1.2 and -1.2, respectively. Following [8], a linear classifier was constructed via the squared error loss from a training sample of size $n = 30$ and size $n = 100$. Notice what the hypothesis test deems as boundary points matches our intuition closely, that is, the projection magnitudes of these accepted points are among the smallest in the data set. However, not all points with a small projection value are classified as boundary points.

In addition to choice of the threshold a_n, the ACI hypothesis test requires an estimate of the asymptotic covariance matrix Σ of $\hat{\beta}_n$. While this is straightforward for the squared error loss, consideration of more general (and more common) loss functions may be problematic.

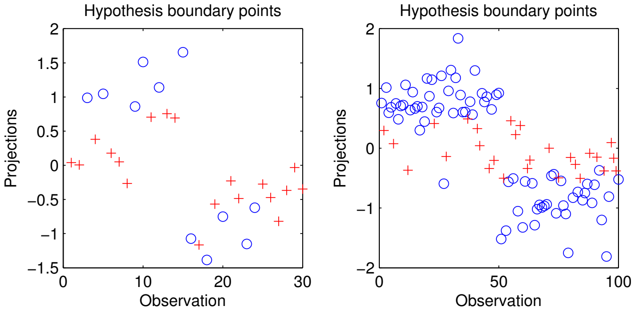

Figure 1: Each point represents an element (x_i, y_i) of the training set. The y-axis displays the projections $x_i^t\hat{\beta}_n$. Points colored in red (the plus symbols) are accepted by the ACI hypothesis test. The accepted points have among the smallest projection values, in absolute terms.

For example, the asymptotic covariance for the hinge loss was only recently derived in Koo et al. [7].

3.2 Upper and Lower Confidence Bounds

The ACI method is closely related to the CPB confidence interval. The ACI can be viewed as a corrected version of the CPB interval that accounts for the additional variation of the conditional test error due to the non-smoothness of the 0-1 loss. The ACI upper bound is based on a bootstrap percentile of the quantity

$$\inf_{u\in\mathbb{R}^p} G_n\mathbb{I}\{X \in B_n\}\mathbb{I}\{X^t u Y < 0\} + G_n\mathbb{I}\{X \in B_n^c\}\mathbb{I}\{X^t\hat{\beta}_n Y < 0\} \tag{4}$$

where the boundary points B_n are defined in Section 3.1. The ACI lower bound is based on bootstrap percentiles of an analogous quantity, in which the infimum is replaced by a supremum.

From [8], we gather that variation in the conditional test error arises, in part, from non-smoothness of the 0-1 loss in conjunction with data points that are close to the boundary of the optimal linear classifier sign$(x^t\beta^*)$. The ACI attempts to capture this additional variation by considering the largest and smallest classification error rates on a training set that consists only of the boundary points. If they are produced from the same bootstrap samples, the ACI always contains the CPB confidence interval, and reduces to it if there are no boundary points. It is worth noting that, while both intervals depend critically on the training error, neither is centered at this quantity.

The asymptotic analysis of the ACI method in [8] considers two situations. The first situation encompasses regular cases in which $P(X^t\beta^* = 0) = 0$, and the second encompasses non-regular cases in which $P(X^t\beta^* = 0) > 0$. In regular cases, the CPB interval and the ACI provide the correct asymptotic coverage, though the ACI always contains the CPB interval. In non-regular cases, the CPB is potentially inconsistent, but the ACI is guaranteed to be conservative. Non-regular cases are of interest, as they encompass situations in which the variance of the test error does not tend to zero with increasing sample size.

The regularity condition $P(X^t\beta^* = 0) = 0$ is satisfied whenever the distribution of X has a density with respect to Lebesgue measure, which is a reasonable assumption in practical situations where some level of homogeneous noise is present. The theoretical results guarantee good asymptotic properties of the ACI under general assumptions. However, they do not fully explain the improved performance of ACI in finite samples. In the regular case, for example, the CPB interval is consistent, and the theoretical results suggest that the consistency of the ACI holds in spite of, not because of, its extremal consideration of boundary points. One obvious reason for the improvement of ACI over the typically anti-conservative CPB intervals in finite samples is that the ACI expands those of the CPB. A less conservative procedure, lying between the CPB interval and the ACI is discussed below.

4 Modifications of the ACI

We investigated several modest changes to the ACI. Based on preliminary simulations, these modifications appear to offer some improvements.

4.1 Motivation

The optimization used to obtain $\hat{\beta}_n$ may potentially lead to overfitting. For small samples, $P_nL(X, Y, \hat{\beta}_n)$ tends to be less than than $PL(X, Y, \hat{\beta}_n)$. Indeed, an elementary argument shows that

$$P_nL(X, Y, \hat{\beta}_n) \ = \ PL(X, Y, \hat{\beta}_n) - \Delta_n + (P_n - P)L(X, Y, \beta^*)$$

where $\Delta_n = (P_nL(X, Y, \beta^*) - P_nL(X, Y, \hat{\beta}_n)) + (PL(X, Y, \hat{\beta}_n) - PL(X, Y, \beta^*))$. Both Δ_n and $(P_n - P)L(X, Y, \beta^*)$ will tend to zero with increasing n, but the former term will usually do so more slowly. (See Hastie et al. [4] for a more detailed discussion of overfitting.) We expect that in many cases an analogous relationship will hold for the 0-1 loss (with $\hat{\beta}_n$ still optimized with respect to the convex loss L), namely, the training error will be less than the test error with high probability.

Overfitting is likely one reason for the poor reported performance of the normal approximation based confidence interval. The training error tends to underestimate the conditional test error, and therefore the plug-in variance estimate in [8] is too small, leading to undercoverage. This might be remedied with a better point estimate of the test error, but the difficulty in obtaining such estimates makes this approach less desirable.

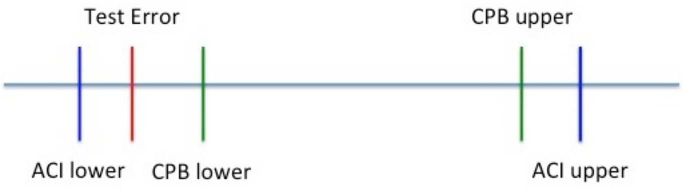

Figure 2: CPB upper bound does not cover but ACI does cover.

Overfitting also has implications for the ACI lower bound. In small sample settings, the CPB confidence interval tends to undercover. We are interested in situations in which the ACI delivers the correct coverage, but the CPB confidence interval fails to do so. This can happen in one of two ways, illustrated in Figure 2 and 3. Simulations indicate undercoverage of the CPB can be largely attributed to its upper bound being too small, rather than its lower bound being too large.

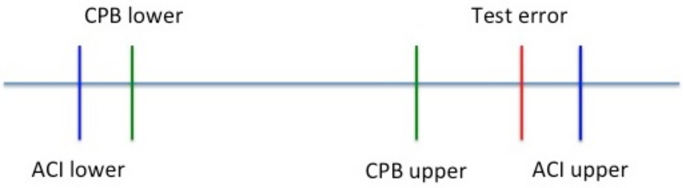

Figure 3: CPB lower bound does not cover but ACI does cover.

As discussed in the beginning of this section, the training error is likely to be less than the test error. Moreover, the CPB lower bound is likely to be less than the training error. This is because the $1-\alpha/2$ percentile of the quantities in (2), $\hat{u}$, is typically positive. To see this, note that in order for $\hat{u}$ to be positive, it only takes very few bootstrap values of the quantity in (2) to be positive. This has the implication that the situation depicted in Figure 3 is more likely to be observed in practice than the situation in Figure 2. Thus the CPB lower bound is a good candidate for a lower bound on the conditional test error, while the smaller ACI lower bound is often too conservative. The ACI upper bound is also subject to some degree of over-conservatism due to the exhaustive search over all of $\mathbb{R}^p$.

4.2 Proposed Modifications

Based on the observations above, we considered two simple improvements to the ACI. The first improvement is to replace the ACI lower bound by the greater (less conservative) CPB lower bound. The second improvement is to replace the ACI upper bound by a smaller (less conservative) bound that lies between it and the CPB upper bound. We describe the upper bound in more detail. Recall that the ACI upper bound is based on bootstrap percentiles of the quantity (4). The infimum captures the best case training error of the boundary points over all linear classifiers. However, in order to address variation of the test error, we are primarily interested in classifiers that might have arisen from other training data arising from the same distribution. Subsamples of the training data provide an obvious proxy for new training data.

Let $\Omega \subset 2^{\mathcal{T}_n}$ be a collection of subsamples of the original training data $\mathcal{T}_n$. Consider a collection of classifiers $\{\hat{\beta}_n^S : S \in \Omega\}$ obtained by minimizing the average of $L(x, y, \beta)$ over subsamples S of the training data.. Accordingly, we replaced the infimum over $\mathbb{R}^p$ in (4) by a minimum over a family of directions $\mathcal{F}_n := \hat{\beta}_n \cup \{\hat{\beta}_n^S : S \in \Omega\}$. (Notice $\mathcal{F}_n$ is data-dependent). The resulting quantity was then bootstrapped, as in the ACI. Details of the choice of Ω are given in Section 6. As the modified ACI always lies between the CPB confidence interval and the standard ACI, it follows immediately from the theory in [8] that the interval resulting from our modifications is consistent when $P(X^t\beta^* = 0) = 0$.

5 An Upper Limit for High Dimensional Data

Currently, the ACI is restricted to the low dimensional setting. Nevertheless, many applications in hand are int he high dimensional setting. This is due to the fact that an exhaustive search over $\mathbb{R}^p$ would produce extreme conservatism, rendering the resulting confidence interval useless. Laber and Murphy suggest a remedy by replacing the exhaustive search with a restricted set of classifiers. Here we briefly suggest one such approach.

In high dimension, low sample size settings, overfitting is more pronounced. In most cases, the training error and CPB lower bound are both equal to zero. As a preliminary step, in this setting we take zero as the lower confidence bound, though we expect that this simple choice can be improved.

Another feature of the high dimension, low sample size setting is that most data points will be close to the empirical decision boundary $x^t\hat{\beta}_n = 0$. For support vector machines, this phenomena is readily observed as "data piling": the projections of most points onto the direction vector $\hat{\beta}_n$ are close to those of the support vectors. See Ahn and Marron [1] for a discussion of this phenomena. With these considerations in mind, we considered an upper confidence bound obtained by bootstrapping the uncentered quantity

$$\sup_{u \in \mathcal{F}_n} P\mathbb{I}\{YX^t u < 0\}. \tag{5}$$

As in the low dimensional case, $\mathcal{F}_n$ contains $\hat{\beta}_n$, and additional direction vectors $\hat{\beta}_n^S$ obtained by subsampling the training data.

6 Preliminary Simulation Results

We applied the methods proposed in Sections 4.2 and 5 to several real and simulated data sets of varying dimension. Many of the data sets we considered here were also used by Laber and Murphy in [8]. Gaussian data (LD for low dimension, and HD for high dimension) was generated in the same fashion as the boundary point simulation in Section 3.1. In the case of data sets taken from the UCI machine learning repository, the true generative model is unknown. Following [8], we substitute the empirical distribution function of the data set as the true generative model. A summary of the data sets considered is given in Table 1.

Name	Features	Instances	Source	$\overline{\tau(\hat{\beta}_n)}, n = 30$	$\overline{\tau(\hat{\beta}_n)}, n = 100$
Quad	2	NA	Simulated	.1088	.1131
Gaussian LD	5	NA	Simulated	.1508	.1241
Liver	7	345	UCI	.3859	.3593
Balance	5	576	UCI	.0800	.0518
Gaussian HD	50	NA	Simulated	.3493	.2329
Sonar	60	208	UCI	.3476	.2198
Spam	57	4601	UCI	.3669	.2231

Table 1: Test data sets used to evaluate confidence interval performance. The last two columns show the average test error for a linear classifier based on the squared error loss and 30 or 100 samples.

For the method proposed in Section 4.2, the family $\mathcal{F}_n$ consisted of $\hat{\beta}_n$ and direction vectors produced from 50 random subsamples of sizes $.1n, .2n, \ldots, .9n$ (a total of 451 directions), using squared error loss. The tuning parameter a_n was chosen following the guidelines in Section 3.4 of [8]. For the upper confidence limit proposed in Section 5, the family $\mathcal{F}_n$ consisted of $\hat{\beta}_n$ and classifiers produced from 100 random subsamples of sizes $.6n, \ldots, .9n$ (a total of 401 directions). In this case, we did not include subsamples of size $.1n$ to $.5n$ because in high dimensional settings, classifiers constructed from such a small portion of the data are not good candidates for possible classifiers constructed from the original training data of size n.

Due to time constraints, we implemented a "lazy" version of the ACI, in which the infimum and supremum were taken over 500 vectors in $\mathbb{R}^p$ uniformly distributed on the $p-1$ unit sphere. It is clear that the standard ACI method will always yield confidence intervals that contain the lazy ACI. Tables 2 and 3 contain a comparison of the modified ACI set out in Section 4.2, the lazy ACI, and the CPB. For the three high dimensional data sets, Table 4 compares the coverage of the upper confidence limit set out in Section 5 and the CPB upper bound, and Table 5 compares the interval widths. There are no reported values of the lazy ACI for the high dimensional data sets because the standard ACI is not applicable in the high dimensional low sample size setting. All results are based on 1000 Monte Carlo iterations, with 100 bootstrap resamples per iteration. Comparisons between the methods are based on the same bootstrap resamples.

	n=30			n=100		
	lazy ACI	**modified ACI**	**CPB**	**lazy ACI**	**modified ACI**	**CPB**
Quad	0.99	0.96	0.83	0.99	0.97	0.91
Gaussian LD	1.00	0.95	0.80	1.00	0.97	0.88
Liver	0.96	0.96	0.83	0.99	0.93	0.87
Balance	1.00	0.96	0.87	1.00	0.99	0.91

Table 2: Low dimensional data sets: coverage comparison between lazy ACI, modified ACI, and CPB for squared error loss. Target coverage is .95.

	n=30			n=100		
	lazy ACI	**modified ACI**	**CPB**	**lazy ACI**	**modified ACI**	**CPB**
Quad	0.29	0.27	0.17	0.20	0.15	0.11
Gaussian LD	0.37	0.32	0.19	0.25	0.17	0.12
Liver	0.52	0.36	0.27	0.32	0.23	0.17
Balance	0.29	0.26	0.14	0.16	0.14	0.08

Table 3: Low dimensional data sets: interval widths comparison between lazy ACI, modified ACI, and CPB for squared error loss. Target coverage is .95.

As the tables show, the modified ACI provides comparable coverage, with intervals that are generally smaller than those of the lazy ACI. Both methods outperform CPB in terms of better coverage. In the high dimensional examples, the upper confidence limit is somewhat conservative, while the CPB severely undercovers. The coverage values for the high dimensional data sets in Table 4 make it clear that the CPB upper bound is falling short of the mark, while the upper confidence limit tends to overshoot.

	n=30		n=100	
	Upper Limit	**CPB**	**Upper Limit**	**CPB**
Gaussian HD	0.97	0.02	1.00	0.38
Sonar	0.99	0.06	1.00	0.23
Spam	0.96	0.08	1.00	0.73

Table 4: High dimensional data sets: coverage comparison between the upper confidence limit proposed in Section 5, and CPB upper bound for squared error loss. Target coverage is .95.

	n=30		n=100	
	Upper Limit	**CPB**	**Upper Limit**	**CPB**
Gaussian HD	0.13	-0.12	0.18	-0.01
Sonar	0.15	-0.09	0.13	-0.03
Spam	0.14	-0.10	0.21	0.03

Table 5: High dimensional data sets: comparison of interval widths between the upper confidence limit proposed in Section 5, and CPB upper bound for squared error loss. Target coverage is .95. Width is defined to be upper bound minus test error.

Acknowledgements

The authors would like to acknowledge extensive and helpful discussions with Andrey Shabalin, who also provided assistance with some of the simulations. In addition, the authors would like to thank Eric Laber for helpful discussions and feedback. The work presented in this paper was supported in part by NSF grant DMS-0907177, NSF Graduate Fellowship, and NIH grant T32 GM067553-05S1.

References

[1] Ahn, J., and Marron, J.S. (2009) The maximal data piling direction for discrimination. *Biometrika*

[2] Braga-Neto, U.M. and Dougherty, E.R.(2004) *Bioinformatics*, 20, 374-380

[3] Efron, B. and Tibshirani, R. (1997) Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association*, 92, 548-560.

[4] Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning* Springer.

[5] Jiang, B., Zhang, X., and Cai,T. (2008) Estimating the Confidence Interval for Prediction Errors of Support Vector Machine Classifiers. *Journal of Machine Learning Research*

[6] Jiang, W., Varma, S., and Simon, R. (2008) Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology*, 7.

[7] Koo, J.-Y., Lee, Y., and Park, C. (2008) A Bahadur Representation of the Linear Support Vector. *Journal of Machine Learning Research*

[8] Laber, E.B. and Murphy, S.A. (2010) Adaptive confidence intervals for the test error in classification. *Preprint*

[9] Nadeau, C., and Bengio, Y. (2003) Inference for the Generalization Error. *Journal of Machine Learning Research*

[10] Yang, Y. (2006) Comparing Learning Methods for Classification. *Statistica Sinica*, 16, 635-657.