*LETTER* Communicated by Hector Corrada Bravo

# Blocked 3×2 Cross-Validated *t*-Test for Comparing Supervised Classification Learning Algorithms

**Wang Yu**
*wangyu@sxu.edu.cn*
**Wang Ruibo**
*wangruibo@sxu.edu.cn*
*Computer Center of Shanxi University, Taiyuan 030006, P.R.C.*

**Jia Huichen**
*jiahuichen@sxu.edu.cn*
*School of Mathematical Sciences, Shanxi University, Taiyuan 030006, P.R.C.*

**Li Jihong***
*lijh@sxu.edu.cn*
*Computer Center of Shanxi University, Taiyuan 030006, P.R.C*

**In the research of machine learning algorithms for classification tasks, the comparison of the performances of algorithms is extremely important, and a statistical test of significance for generalization error is often used to perform it in the machine learning literature. In view of the randomness of partitions in cross-validation, a new blocked 3×2 cross-validation is proposed to estimate generalization error in this letter. We then conduct an analysis of variance of the blocked 3×2 cross-validated estimator. A relatively conservative variance estimator that considers the correlation between any two two-fold cross-validations, and was previously neglected in 5×2 cross-validated *t* and *F*-tests is put forward. A corresponding test using this variance estimator is presented to compare the performances of algorithms. Simulated results show that the performance of our test is comparable with that of 5×2 cross-validated tests but with less computation complexity.**

## 1 Introduction

In typical supervised classification learning, generalization error is a major measure of the performance of learning algorithm. To compare the performances of algorithms, we usually need to use statistical tests of significance to determine which algorithm will be most preferable due to the existence of

*Corresponding author. Wang Yu and Li Jihong contributed equally to this work.

*Neural Computation* **26, 208–235** (2014)
doi:10.1162/NECO_a_00532

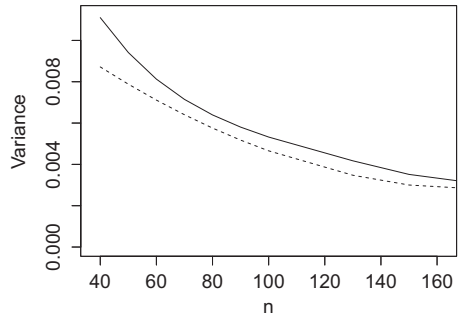


Figure 1: Variance estimator given by Dietterich (1998), shown by the dashed curve, versus the real variance, shown by the solid curve.

random error. Currently generalization error is usually estimated through various forms of cross-validation. For example, Nadeau and Bengio (2003) estimated generalization error by repeated learning-testing, whereas Bengio and Grandvalet (2004), Markatou, Tian, Biswas, and Hripcsak (2005), and Grandvalet and Bengio (2006) used standard *K*-fold cross-validation. For testing the significance of differences between two algorithms, Dietterich (1998) proposed a new 5×2 cross-validated *t*-test and demonstrated that its performance is better than ten-fold cross-validation by simulated experiments. Alpaydin (1999) constructed a variant of 5×2 cross-validated *t*-test, combined 5×2 cross-validated *F*-test, and showed that it has higher power than the 5×2 cross-validated *t*-test. Similar to Alpaydin (1999), Yildiz (2013) gave a combined 5×2 cross-validated *t*-test and exhibited that it has higher power (lower type II error) and lower type I error compared to the 5×2 cross-validated *t*-test. The tests of significance for comparing the performances of algorithms based on other performance measures such as AUC (area under the receiver operating characteristic curve) and *F* value refer to Chen, Gallas, and Yousef (2012) and Yang and Liu (1999).

In terms of the method given by Dietterich (1998), Alpaydin (1999) and Yildiz (2013), we know that the 5×2 cross-validation is the average of random samplings conducted five times. They used the average of five independent sample variances as the estimation of variance in constructing test statistics. However, an underestimation of real variance is illustrated by the following example:

**Example 1.** Comparison of the real variance of $p_1^{(1)}$ and the variance estimator given by Dietterich (1998) in the 5×2 cross-validated *t*-test (see Figure 1).

We have $Z = (X, Y)$ with $P(Y = 1) = P(Y = 0) = \frac{1}{2}$, $X|Y = 0 \sim N(0, I_{10})$, and $X|Y = 1 \sim N(1, 2I_{10})$. Here, the learning algorithm is a classification tree. The real variance of $p_1^{(1)}$ from 5×2 cross-validation is estimated on 100,000 independent experiments.

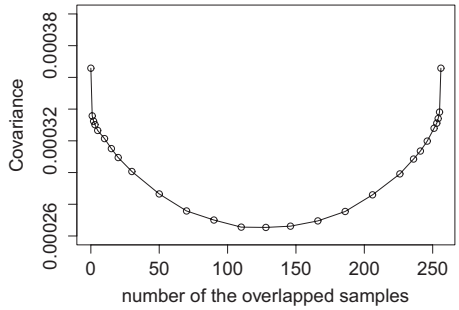


Figure 2: Curve of covariance versus number of overlapped samples.

Although random partitions for 5×2 cross-validation are independent, training sets (test sets) from any two independent partitions contain common samples regardless of how the data are split. Thus, cross-validation estimators for different data partitions are actually not independent. The 5×2 cross-validated $F$-test in Alpaydin (1999) evidently does not consider the correlation among the five two-fold cross-validations: $p_i^{(j)}$ and $p_{i'}^{(j')}$, $i \neq i', i, i' = 1, \ldots, 5$, $j, j' = 1, 2$. This observation can be problematic assuming that the numerator $(\sum_{i=1}^{5}\sum_{j=1}^{2}(p_i^{(j)})^2)$ of the test statistic follows the $\sigma^2\chi^2(10)$ distribution and the test statistic follows the $F(10, 5)$ distribution. Despite the relatively large sample size, the proportion of covariance component to the total variance remains significant. This observation is also evident from the simulated experiment 2 and real example 1 in section 3. However, results may lead to a liberal 5×2 cross-validated $t$- or $F$-test.

Furthermore, the dependence between any two two-fold cross-validations is related to the number of overlapped samples between training sets. The impact of number of overlapped samples on covariance is also demonstrated in this study.

**Example 2.** Consider five independent predictors, all standard normal. The conditional probability function is

$$p(x) = \frac{\exp(1 + \sum_{i=1}^{5} 1.5x_i)}{1 + \exp(1 + \sum_{i=1}^{5} 1.5x_i)}$$

with a probability of $Y = 1$. We took the sample size to be 512. The learning algorithm is a support vector machine (SVM).

Figure 2 shows that the initial decrease and subsequent increase in the covariance of any two two-fold cross-validated estimators correspond to an increase in the number of overlapped samples. The covariance reaches the minimum when the number of overlapped samples is $\frac{n}{4}$ ($n$ is the sample

size). However, the 5×2 cross-validation in Dietterich (1998), Alpaydin (1999), and Yildiz (2013) resulted from five random partitions, which facilitated a larger variance. In particular, Markatou et al. (2005) theoretically proved that the number of overlapped samples from any two training sets follows a hypergeometric distribution and that the mathematical expectation is $\frac{n}{4}$ ($n$ is the sample size). Overlapped samples with a larger deviation from $\frac{n}{4}$ can result in a large covariance, which is unavoidable for a random 5×2 cross-validation. The assumption that an identical covariance exists for different groups of random 5×2 cross-validation is not reasonable despite the fact that samples are independent and identically distributed when theoretically analyzing the covariance between different groups. This observation also leads to difficulty in the theoretical analysis of a variance of a random 5×2 cross-validated estimator.

Thus, for the data partitions of cross-validation, we should split the data such that the ideal value ($\frac{n}{4}$) can be achieved for each partition through preliminary design; then the resulting training and test sets have a better sample balance. By doing this, we aim to mitigate the effect of the number of overlapped samples such that the covariance between different groups becomes theoretically identical. Intuitively, the resulting cross-validation estimator is supposed to have good properties such as smaller variance (see the simulated experiments 1).

Alpaydin (2010) devoted a chapter of his book to the importance of the design of machine learning experiments. In this study, we also regard data partition as an experimental design—designing the data partitions according to the purpose of the experiment. For example, the basic task of machine learning experimental design is to find out how it can use fewer data partitions and fewer experiments to achieve the goal of the experiment.

Based on the above analysis, when implementing cross-validation, the data should first be split into four balanced blocks and then take two as a training set and the other two as a test set to implement the two-fold cross-validation. However, such combinations have three groups altogether, so three replications of two-fold cross-validations are performed (instead of performing a four-fold cross-validation). We call this blocked 3×2 cross-validation. For blocked 3×2 cross-validation, any two data sets between different groups (either the training or the test sets) obviously have the same number of overlapped samples and better sample balance, thus resulting in variance estimation that should have good properties (better variance estimation is the premise of a comparison of algorithm performances). Then a relatively conservative variance estimator that considers the correlation between any two two-fold cross-validations that was previously neglected in 5×2 cross-validated $t$ and $F$-tests is put forward. A corresponding test using this variance estimator is presented to compare the algorithms' performances. Section 6 shows that the performance of our test is comparable to that of 5×2 cross-validated tests. In theory, more replications can help yield better variance estimation, but the experiments by design can achieve

the goal in fewer replicated experiments and improve the efficiency of the variance estimation (i.e., the same accuracy estimator can be obtained even with fewer number of experiments).

Based on the same idea, Li, Wang, Wang, and Li (2010) applied this method to the practical application of natural language processing (NLP). The method of balancing data into four blocks essentially reflects the idea of statistical experimental design, which requires that the collected data should be designed beforehand. This partition method may be easier to implement in NLP.

The remainder of this paper is organized as follows. Section 2 defines the measures of performance for algorithms and their estimation using $K$-fold cross-validation and blocked $3\times2$ cross-validation. A theoretical analysis of the variance of blocked $3\times2$ cross-validated estimator for generalization error is given in section 3. The developed variance estimators are presented in section 4, and section 5 describes the corresponding test statistics. Section 6 discusses the simulated experiments that show how the proposed statistic behaves compared with statistics that are already in use. Section 7 concludes the letter.

## 2 Blocked $3\times2$ Cross-Validation Estimator of Generalization Error

**2.1 Definition of Generalization Error.** Formally, we assume that a data set $D = \{z_1, z_2, \ldots, z_n\}$, $z_i \in \mathcal{Z}$ is independently sampled from an unknown distribution $P$, where $z_i = (x_i, y_i)$, $x_i$ is an input vector and $y_i$ is an output variable. If $f = A(D)$ denotes the prediction function returned by algorithm $A$ on the data set $D$, and $L(f(x), y) = I[f(x) \neq y]$($\{0, 1\}$-loss) represents a measure of discrepancy between the prediction and the observation, then the generalization error of classification learning algorithm is defined by

$$\mu(n) = EPE(n) = E[L(A(D), z)], \tag{2.1}$$

where the expectation is taken with respect to $D$ and $z$, $z$ sampled from $P$ is independent of $D$. The expectation of equation 2.1 means that we are interested in the general performance of a classification algorithm, not the performance of a specific data set.

**2.2 $K$-Fold Cross-Validation Estimator.** $K$-fold cross-validation is probably the simplest and most widely used method for estimating generalization error in limited sample cases. It uses all available examples as training and test examples and mimics $K$ training and test sets by using some of the data to fit the model and some to test it. Afterward, the generalization error is estimated by combining the $K$ results.

Table 1: The Blocked 3×2 Cross-Validation.

| Group | Training Set | Test Set | $\hat{\mu}_k^{(i)}$ |
|---|---|---|---|
| 1 | $D_1^{(1)} = (P_1, P_2)$ | $T_1^{(1)} = (P_3, P_4)$ | $\hat{\mu}_1^{(1)}$ |
| 1 | $D_2^{(1)} = (P_3, P_4)$ | $T_2^{(1)} = (P_1, P_2)$ | $\hat{\mu}_2^{(1)}$ |
| 2 | $D_1^{(2)} = (P_1, P_3)$ | $T_1^{(2)} = (P_2, P_4)$ | $\hat{\mu}_1^{(2)}$ |
| 2 | $D_2^{(2)} = (P_2, P_4)$ | $T_2^{(2)} = (P_1, P_3)$ | $\hat{\mu}_2^{(2)}$ |
| 3 | $D_1^{(3)} = (P_1, P_4)$ | $T_1^{(3)} = (P_2, P_3)$ | $\hat{\mu}_1^{(3)}$ |
| 3 | $D_2^{(3)} = (P_2, P_3)$ | $T_2^{(3)} = (P_1, P_4)$ | $\hat{\mu}_2^{(3)}$ |

First, the data set $D$ is split into $K$ disjoint and equal-sized blocks, denoted as $T_k, k = 1, 2, \ldots, K$. Let $D_k$ be the training set obtained by removing the elements in $T_k$ from $D$. Then the cross-validation estimator is

$$\hat{\mu}_K = \frac{1}{K} \sum_{k=1}^{K} \hat{\mu}_k, \tag{2.2}$$

where $\hat{\mu}_k = \frac{1}{m} \sum_{z_i \in T_k} L(A(D_k), z_i)$ and $m \approx n/K$. When we want to compare the performances of algorithms $A_1$ and $A_2$, $\hat{\mu}_k = \frac{1}{m} \sum_{z_i \in T_k} (L(A_1(D_k), z_i) - L(A_2(D_k), z_i))$.

Note that the $\hat{\mu}_K$ estimates $EPE(n - m)$, not $EPE(n)$. It estimates $EPE(\frac{n}{2})$ in the case of two-fold cross-validation.

**2.3 Blocked 3×2 Cross-Validation Estimator.** The blocked 3×2 cross-validation first proposed by Li et al. (2010) in the practical application of natural language processing is a method used for corpus partitioning. They proposed that data set $D$ should first be split into four balanced blocks and then take either the two as the training set and the other two as a test set to implement the two-fold cross-validation. However, such combinations have three groups altogether, so three replications of two-fold cross-validations are performed.

In practice, data set $D$ is split into four disjoint and equal-sized blocks, denoted as $P_j$, $j = 1, 2, 3, 4$, respectively. The combination of any two $P_j$s results in three groups and six different combinations (see Table 1). Here, $D_k^{(i)}, i = 1, 2, 3, k = 1, 2$ denotes the training set, and $T_k^{(i)}, i = 1, 2, 3, k = 1, 2$ denotes the test set. They serve as a training or testing set with each other; thus, $D_1^{(i)} = T_2^{(i)}, D_2^{(i)} = T_1^{(i)}, i = 1, 2, 3$. The blocked 3×2 cross-validation is defined as the average of errors in all three groups:

$$\hat{\mu}_{3\times 2} = \frac{1}{3} \sum_{i=1}^{3} \hat{\mu}^{(i)} = \frac{1}{3} \sum_{i=1}^{3} \frac{1}{2} \sum_{k=1}^{2} \hat{\mu}_k^{(i)}, \tag{2.3}$$

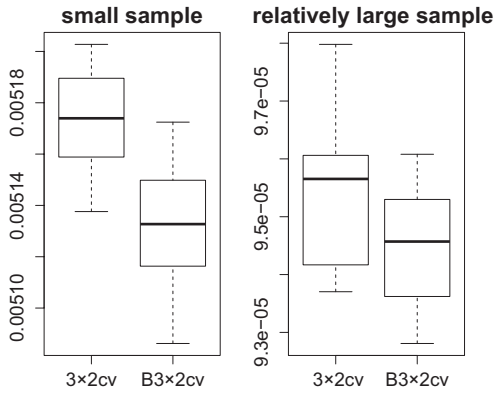


Figure 3: Box plots for simulated experiment 1, where B 3×2 cv (3×2 cv) refers to blocked 3×2 cross validation (random 3×2 cross validation)

where $\hat{\mu}_k^{(i)} = \frac{2}{n} \sum_{z_j \in T_k^{(i)}} L(A(D_k^{(i)}), z_j)$. Similar to the $K$-fold cross-validation, when comparing the performances of algorithms $A_1$ and $A_2$, $\hat{\mu}_k^{(i)} = \frac{2}{n} \sum_{z_j \in T_k^{(i)}} (L(A_1(D_k^{(i)}), z_j) - L(A_2(D_k^{(i)}), z_j))$ .

The structure of blocked 3×2 cross-validation based on four-block data partition is very special; that is, there is one, and only one, overlapped block in any two combinations between different groups: $\#\{D_k^{(i)} \bigcap D_{k'}^{(i')}\} = n/4$, $\#\{T_k^{(i)} \bigcap T_{k'}^{(i')}\} = n/4$, $\#\{D_k^{(i)} \bigcap T_{k'}^{(i')}\} = n/4$ for $i \neq i'$, $k$ and $k'$ randomly drawn, $i, i' = 1, 2, 3$, $k, k' = 1, 2$. This balance reduces errors resulting from different numbers of overlapped samples between different groups in random 3×2 cross-validation obtained from three random partitions (see Figure 3). Refer to simulated experiment 1:

**Simulated experiment 1:** At the sample size $n = 40$, for 20 observations with $Y = 1$, we generate three independent random variables $X_1, X_2, X_3$, all standard normal. For the remaining 20 observations with $Y = 0$, we generate the three predictors that are also independent, but with $N(0.4, 1)$, $N(0.3, 1)$, and $N(0, 1)$ distributions, respectively. Then $X_3$ is not useful for classifying Y. The learning algorithm is a classification tree. We then compare variances of the blocked 3×2 cross-validation and random 3×2 cross-validation. In addition to this setup, we also consider another one with relatively large sample sizes. At sample size $n = 1024$, for 512 observations with $Y = 1$, we generate 300 independent random variables, all standard normal. For the remaining 512 observations with $Y = 0$, we generate 300 predictors that are also independent, but with $N(0.4, I_{100})$, $N(0.3, I_{100})$, and $N(0, I_{100})$ distributions, respectively.

## 3 Theoretical Analysis of the Variance of $\hat{\mu}_{3\times 2}$

For the convenience of the readers, a lemma in Nadeau and Bengio (2003) is listed first.

**Lemma 1.** *Let $U_1, U_2, \ldots, U_K$ be random variables with common mean $\beta$ and the following covariance structure:*

$$\mathrm{Var}(U_k) = \delta, \forall\, k, \mathrm{Cov}(U_k, U_{k'}) = \gamma, \forall\, k \neq k'.$$

*Let $\pi = \frac{\gamma}{\delta}$ be the correlation between $U_k$ and $U_{k'}$. Let $\bar{U} = \frac{1}{K}\sum_{k=1}^{K} U_k$ and $S_U^2 = \frac{1}{K-1}\sum_{k=1}^{K}(U_k - \bar{U})^2$ be the sample mean and sample variance, respectively. Then:*

1. $\mathrm{Var}(\bar{U}) = \gamma + \frac{\delta-\gamma}{K}$.
2. *If the stated covariance structure holds for all K (with $\gamma$ and $\delta$ not depending on K), then $\gamma \geq 0$.*
3. $E(S_U^2) = \delta - \gamma$.

To study $\mathrm{Var}(\hat{\mu}_{3\times 2})$, we need to define the following covariances.

**Definition 1.**

- *Let $\sigma_1^2 = Var(\hat{\mu}_k^{(i)})$ for $i = 1, 2, 3,\ k = 1, 2$.*
- *Let $\sigma_2^2 = Cov(\hat{\mu}_k^{(i)}, \hat{\mu}_{k'}^{(i)})$ for $k \neq k'$, that is, the covariance within group.*
- *Let $\sigma_3^2 = Cov(\hat{\mu}_k^{(i)}, \hat{\mu}_{k'}^{(i')})$, with $i \neq i'$, $k$ and $k'$ randomly drawn, that is, the covariance between different groups. The implicit assumption is made that the covariances of any two blocks in different groups are all the same. This is reasonable for the blocked 3×2 cross-validation. However, it may not be reasonable for the (random) 3×2 or 5×2, because the covariance of any two two-fold cross-validated estimators decreases (or increases) with an increase in the number of overlapped samples.*

**Proposition 1.** *The mean and variance of $\hat{\mu}_{3\times 2}$ have the following forms:*

$$E(\hat{\mu}_{3\times 2}) = \mu(n/2), \tag{3.1}$$

$$Var(\hat{\mu}_{3\times 2}) = \frac{1}{6}\sigma_1^2 + \frac{1}{6}\sigma_2^2 + \frac{2}{3}\sigma_3^2 = \frac{1}{6}\sigma_1^2(1 + \rho_1 + 4\rho_2), \tag{3.2}$$

*where $\rho_1 = \sigma_2^2/\sigma_1^2 = Corr(\hat{\mu}_k^{(i)}, \hat{\mu}_{k'}^{(i)})$ for $k \neq k'$, $\rho_2 = \sigma_3^2/\sigma_1^2 = Corr(\hat{\mu}_k^{(i)}, \hat{\mu}_{k'}^{(i')})$ for $i \neq i'$, $k = k'$ or $k \neq k'$.*

**Proof.** Concerning expectations, we obviously have $E(\hat{\mu}_k^{(i)}) = \mu(n/2)$, and thus $E(\hat{\mu}_{3\times 2}) = \mu(n/2)$.

From lemma 1, we have

$$\sigma_1' \equiv \mathrm{Var}(\hat{\mu}^{(i)}) = \frac{1}{2}(\sigma_1^2 + \sigma_2^2).$$

For $i \neq i'$, we have

$$\sigma_2' \equiv \mathrm{Cov}(\hat{\mu}^{(i)}, \hat{\mu}^{(i')}) = \frac{1}{4}\sum_{k=1}^{2}\sum_{k'=1}^{2}\mathrm{Cov}(\hat{\mu}_k^{(i)}, \hat{\mu}_{k'}^{(i')}) = \sigma_3^2,$$

and therefore (using lemma 1 again)

$$\mathrm{Var}(\hat{\mu}_{3\times 2}) = \sigma_2' + \frac{\sigma_1' - \sigma_2'}{3} = \frac{1}{6}\sigma_1^2 + \frac{1}{6}\sigma_2^2 + \frac{2}{3}\sigma_3^2 = \frac{1}{6}\sigma_1^2(1 + \rho_1 + 4\rho_2).$$

From equation 3.2, we know that the variance of blocked 3×2 cross-validation can decompose in three components. In detail, it is a linear combination of three moments: $\sigma_1^2, \sigma_2^2, \sigma_3^2$. In truth, $\sigma_2^2$ and $\sigma_3^2$ cannot be negligible. Next, we illustrate this using the simulated and real data.

**Simulated experiment 2:** *Classification problem with two classes on simulated data*.

We have $Z = (X, Y)$ with $P(Y = 1) = P(Y = 0) = \frac{1}{2}$, $X|Y = 0 \sim N(0, I_5)$, and $X|Y = 1 \sim N(1, 2I_5)$. Here, the learning algorithm is a logistic regression. We now look at the variance of $3 \times 2$ cross-validation and decompose in the three components for $n = 150, 300, 400, 500, 1000$.

From Figure 4, we know that $\sigma_2^2$ has little effect, but the contribution of $\sigma_3^2$ to $\mathrm{Var}(\hat{\mu}_{3\times 2})$ is of the same order as the one of $\sigma_1^2$, and even greater. Thus, neglecting the effect of $\sigma_2^2$ and $\sigma_3^2$ will introduce a large bias.

**Real example 1:** Classification problem on the letter data of UCI database.

A data set for identifying the letters of the roman alphabet comprises 20,000 examples described by 16 features. The 26 letters represent 26 categories, similar to Nadeau and Bengio (2003), who turned it into a two-class (A–M versus N–Z) classification problem. The learning algorithm is classification tree. Then look at the variance of the blocked 3×2 cross-validation estimator of generalization error and decompose in the three components for $n = 20, 40, 80, 160, 400, 800, 2000$. Results are shown in Figure 5.

Note that accurate estimations of $\mathrm{Var}(\hat{\mu}_{3\times 2})$ and the decomposition of its three components require many independent replicated experiments. This was achieved by independent sampling from 20,000 examples with replacement.

With the changes in the capacity of sample set, $\sigma_1^2$ is responsible for only 30% to 60% of the total variance. Findings further prove that the

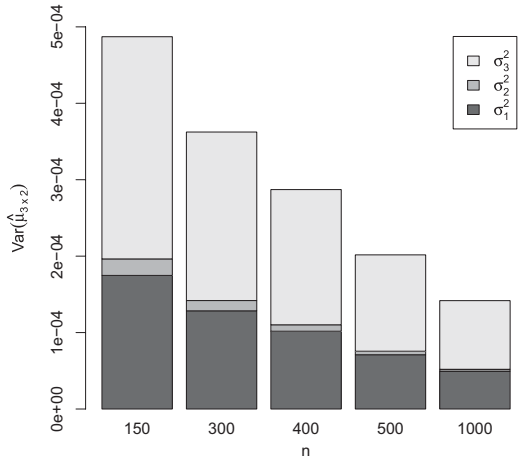


Figure 4: Bar plots of the contributions to $\text{Var}(\hat{\mu}_{3\times2})$ due to $\sigma_1^2, \sigma_2^2, \sigma_3^2$ versus the number of examples $n$ for simulated experiment 2.

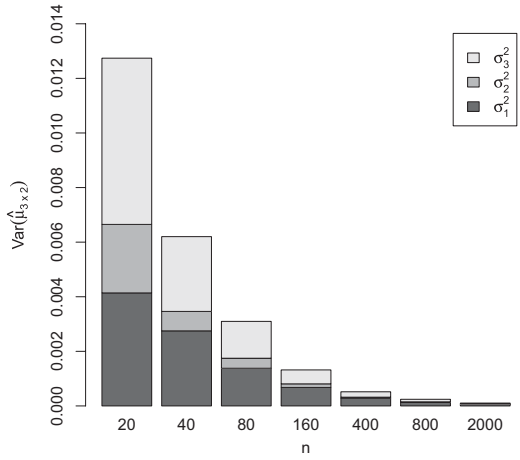


Figure 5: Bar plots of the contributions to $\text{Var}(\hat{\mu}_{3\times2})$ due to $\sigma_1^2, \sigma_2^2, \sigma_3^2$ versus the number of examples $n$ for real example 1.

correlations between blocks should not be negligible when considering $\text{Var}(\hat{\mu}_{3\times2})$. Similar to what Bengio and Grandvalet (2004) pointed out, the estimation of variance of $K$-fold cross-validation should indeed take into account the correlations of errors.

Our findings also demonstrate that it is not appropriate to simply use the sample variance to estimate true variance because it seriously underestimates the true variance.

## 4 Estimation of $\mathrm{Var}(\hat{\mu}_{3\times 2})$

**4.1 Estimator Proposed by Li (2010) (the *L* Estimator).** If the three groups in the blocked 3×2 cross-validation are independent of each other, $\mathrm{Var}(\hat{\mu}_{3\times 2})$ should be expressed as

$$\frac{1}{6}\sigma_1^2(1+\rho_1),$$

where $\rho_1 = \mathrm{Corr}(\hat{\mu}_k^{(i)}, \hat{\mu}_{k'}^{(i)}), k \neq k', i = 1, 2, 3.$

Then from lemma 1, we know that

$$\frac{1}{9}\sum_{i=1}^{3}\left(\frac{1}{2}+\frac{\rho_1}{1-\rho_1}\right)\sum_{k=1}^{2}(\hat{\mu}_k^{(i)}-\hat{\mu}^{(i)})^2$$

is an unbiased estimator of $\mathrm{Var}(\hat{\mu}_{3\times 2})$. However, the problem is that $\rho_1$ is unknown and difficult to estimate. Li (2010) recommends the use of $\hat{\rho}_1 = 0.5$ as a estimator of $\rho_1$, such that the estimator of $\mathrm{Var}(\hat{\mu}_{3\times 2})$ can be written as

$$\begin{aligned}\widehat{\mathrm{Var}}_1(\hat{\mu}_{3\times 2}) &= \frac{1}{9}\sum_{i=1}^{3}\left(\frac{1}{2}+\frac{\frac{1}{2}}{1-\frac{1}{2}}\right)\sum_{k=1}^{2}(\hat{\mu}_k^{(i)}-\hat{\mu}^{(i)})^2 \\ &= \frac{1}{6}\sum_{i=1}^{3}\sum_{k=1}^{2}\left(\hat{\mu}_k^{(i)}-\hat{\mu}^{(i)}\right)^2. \end{aligned} \tag{4.1}$$

We call this estimator the *L* estimator in this paper. Since there are three restrictions in equation 4.1, the degrees of freedom of *L* estimator are 3.

Note that

$$\begin{aligned} E(\widehat{\mathrm{Var}}_1(\hat{\mu}_{3\times 2})) &= \frac{1}{9}\sum_{i=1}^{3}\left(\frac{1}{2}+\frac{\rho_1}{1-\rho_1}+1-\frac{\rho_1}{1-\rho_1}\right)\sum_{k=1}^{2}(\hat{\mu}_k^{(i)}-\hat{\mu}^{(i)})^2 \\ &= \frac{1}{6}\sigma_1^2(1+\rho_1)+\left(1-\frac{\rho_1}{1-\rho_1}\right)\frac{\frac{1}{6}\sigma_1^2(1+\rho_1)}{\frac{1}{2}+\frac{\rho_1}{1-\rho_1}} \\ &= \frac{1}{6}\sigma_1^2(1+\rho_1)+\frac{2}{6}\sigma_1^2(1-2\rho_1), \end{aligned}$$

$$\mathrm{Var}(\hat{\mu}_{3\times 2}) = \frac{1}{6}\sigma_1^2+\frac{1}{6}\sigma_2^2+\frac{2}{3}\sigma_3^2 = \frac{1}{6}\sigma_1^2(1+\rho_1)+\frac{2}{6}\sigma_1^2(2\rho_2).$$

Thus, to determine if the $L$ estimator underestimates or overestimates $\mathrm{Var}(\hat{\mu}_{3\times 2})$, we only need to compare the $1-2\rho_1$ and $2\rho_2$. If $1-2\rho_1 < 2\rho_2$, the estimator $\widehat{\mathrm{Var}}_1(\hat{\mu}_{3\times 2})$ is said to be liberal; otherwise, it is conservative.

Similar to Nadeau and Bengio (2003), we present the following proposition:

**Proposition 2.** *Assume that $L(A(D_k^{(i)}), z_j)$ depends on only test samples $z_j$ and sample size (i.e., the loss function does not depend on the actual examples $D_k^{(i)}$ and the underlying algorithm), $\rho_2 = 0.5$.*

**Proof.** Under the assumption of proposition 2, the loss function can be simply written as the function of test samples. Indeed, when $L(A(D_k), z_j) = f(z_j)$, we have

$$\hat{\mu}_k^{(1)} = \frac{2}{n}\sum_{z_j \in T_k} L(A(D_k), z_j) = \frac{2}{n}\sum_{z_j \in T_k} f(z_j),$$

and

$$\hat{\mu}_{k'}^{(2)} = \frac{2}{n}\sum_{z'_j \in T_{k'}} L(A(D_{k'}), z_{j'}) = \frac{2}{n}\sum_{z_{j'} \in T_{k'}} f(z_{j'}).$$

From the independence of $z_j$, we obviously have

$$\mathrm{Var}(\hat{\mu}_k^{(1)}) = \frac{4}{n^2}\sum_{z_j \in T_k} \mathrm{Var}(f(z_j)) = \frac{2}{n}\mathrm{Var}(f(z_j)) = \mathrm{Var}(\hat{\mu}_{k'}^{(2)}),$$

$$\mathrm{Cov}(\hat{\mu}_k^{(1)}, \hat{\mu}_{k'}^{(2)}) = \frac{4}{n^2}\sum_{z_j \in T_k}\sum_{z_{j'} \in T_{k'}} \mathrm{Cov}(f(z_j), f(z_{j'})).$$

Then, from the definition of $T_k$, we know that half of the observations are the same between any two blocks of different groups. Thus, without loss of generality, let $T_k = (P_1, P_2)$, $T_{k'} = (P_1, P_3)$. We then have

$$\begin{aligned}\mathrm{Cov}(\hat{\mu}_k^{(1)}, \hat{\mu}_{k'}^{(2)}) &= \frac{4}{n^2}\sum_{z_j \in P_1} \mathrm{Cov}(f(z_j), f(z_j)) = \frac{4}{n^2}\frac{n}{4}\mathrm{Var}(f(z_j)) \\ &= \frac{1}{n}\mathrm{Var}(f(z_j)).\end{aligned}$$

Thus, the correlation between $\hat{\mu}_k^{(1)}$ and $\hat{\mu}_{k'}^{(2)}$ is

$$\rho_2 = \frac{\mathrm{Cov}(\hat{\mu}_k^{(1)}, \hat{\mu}_{k'}^{(2)})}{\mathrm{Var}(\hat{\mu}_k^{(1)})} = \frac{1}{2}.$$

From Table 4 we know that $L(A(D_k^{(i)}), z_j)$ depends on the actual examples $D_k^{(i)}$ and the underlying algorithm $A$, so $\rho_2 < 0.5$. Moreover, the simulated results in Tables 2 and 3 indicate that the $L$ estimator underestimates the true variance.

**4.2 A New Estimator (the *AL* Estimator).** The $L$ estimator is obtained by neglecting the correlation among different groups; however, we know that $\rho_2$ cannot actually be neglected by proposition 2 and the following simulated experiments. Based on this, we present a new estimation method by adapting the $L$ estimator denoted as

$$\widehat{\mathrm{Var}}_2(\hat{\mu}_{3\times 2}) = \widehat{\mathrm{Var}}_1(\hat{\mu}_{3\times 2}) + \lambda S^2_{\hat{\mu}^{(i)}} \tag{4.2}$$

where $S^2_{\hat{\mu}^{(i)}} = \frac{1}{2}\sum_{i=1}^{3}(\hat{\mu}^{(i)} - \hat{\mu}_{3\times 2})^2$ denotes the sample variance of $\hat{\mu}^{(i)}$, $\hat{\mu}_{3\times 2} = \frac{1}{3}\sum_{i=1}^{3}\hat{\mu}^{(i)}$ and $\lambda$ is the turning parameter. We refer to this method as the $AL$ estimator (Adaptive $L$ estimator). In particular, when $\lambda = 0$, equation 4.2 is the $L$ estimator. When $\lambda = \frac{2}{3}$, the following simple form for the $AL$ estimator is obtained:

$$\begin{aligned}
\widehat{\mathrm{Var}}_2(\hat{\mu}_{3\times 2}) &= \frac{1}{6}\sum_{i=1}^{3}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}^{(i)})^2 + \frac{2}{3}\cdot\frac{1}{2}\sum_{i=1}^{3}(\hat{\mu}^{(i)} - \hat{\mu}_{3\times 2})^2 \\
&= \frac{1}{6}\sum_{i=1}^{3}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}^{(i)})^2 + \frac{1}{6}\sum_{i=1}^{3}\sum_{k=1}^{2}(\hat{\mu}^{(i)} - \hat{\mu}_{3\times 2})^2 \\
&= \frac{1}{6}\sum_{i=1}^{3}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}_{3\times 2})^2.
\end{aligned} \tag{4.3}$$

This is based on $\frac{1}{6}\sum_{i=1}^{3}\sum_{k=1}^{2}(\hat{\mu}^{(i)} - \hat{\mu}_{3\times 2})(\hat{\mu}_k^{(i)} - \hat{\mu}^{(i)}) = \frac{1}{6}\sum_{i=1}^{3}(\hat{\mu}^{(i)} - \hat{\mu}_{3\times 2})\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}^{(i)}) = 0$.

Obviously, there is only one restriction in equation 4.3, which means that the degrees of freedom of $AL$ estimator are 5. Next, we consider the relationship between $E(\widehat{Var}_2(\hat{\mu}_{3\times 2}))$ and $Var(\hat{\mu}_{3\times 2})$.

Table 2: The True Values of $1-2\rho_1$, $2\rho_2$, $2-\rho_1-2\rho_2$, and $3-4\rho_2$ for Simulated Experiment 2.

| | $n=$ | | | | |
|---|---|---|---|---|---|
| | 150 | 300 | 400 | 500 | 1000 |
| $1-2\rho_1$ | 0.752 | 0.790 | 0.834 | 0.872 | 0.900 |
| $2\rho_2$ | 0.832 | 0.858 | 0.870 | 0.888 | 0.908 |
| $2-\rho_1-2\rho_2$ | 1.044 | 1.037 | 1.047 | 1.048 | 1.042 |
| $3-4\rho_2$ | 1.336 | 1.284 | 1.260 | 1.224 | 1.184 |

Table 3: The True Values of $1-2\rho_1$, $2\rho_2$, $2-\rho_1-2\rho_2$, and $3-4\rho_2$ for Real Example 1.

| | $n=$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | 20 | 40 | 80 | 160 | 400 | 800 | 2000 |
| $1-2\rho_1$ | 0.414 | 0.448 | 0.604 | 0.676 | 0.810 | 0.880 | 0.946 |
| $2\rho_2$ | 0.562 | 0.522 | 0.390 | 0.378 | 0.330 | 0.310 | 0.288 |
| $2-\rho_1-2\rho_2$ | 1.145 | 1.202 | 1.412 | 1.460 | 1.575 | 1.630 | 1.685 |
| $3-4\rho_2$ | 1.876 | 1.956 | 2.220 | 2.244 | 2.340 | 2.380 | 2.424 |

From lemma 1, we have

$$E(S^2_{\hat{\mu}^{(i)}}) = \sigma_1' - \sigma_2' = \sigma_1'(1-\rho) = \frac{1-\rho}{\rho+\frac{1-\rho}{3}}\sigma_1'\left(\rho+\frac{1-\rho}{3}\right)$$
$$= \frac{\sigma_1'(\rho+\frac{1-\rho}{3})}{\frac{1}{3}+\frac{\rho}{1-\rho}} = \frac{\mathrm{Var}(\hat{\mu}_{3\times 2})}{\frac{1}{3}+\frac{\rho}{1-\rho}},$$

where $\rho = \frac{\sigma_2'}{\sigma_1'} = \frac{2\sigma_3^2}{\sigma_1^2+\sigma_2^2} = \frac{2\rho_2}{1+\rho_1}$ is the correlation of $\hat{\mu}^{(i)}$. Then

$$E(\widehat{\mathrm{Var}}_2(\hat{\mu}_{3\times 2})) = E(\widehat{\mathrm{Var}}_1(\hat{\mu}_{3\times 2})) + \lambda E(S^2_{\hat{\mu}^{(i)}})$$
$$= \frac{1}{6}\sigma_1^2(1+\rho_1) + \frac{2}{6}\sigma_1^2(1-2\rho_1) + \lambda\frac{\mathrm{Var}(\hat{\mu}_{3\times 2})}{\frac{1}{3}+\frac{\rho}{1-\rho}}$$
$$= \frac{1}{6}\sigma_1^2(1+\rho_1) + \frac{1}{6}\sigma_1^2(2+3\lambda+(3\lambda-4)\rho_1-6\lambda\rho_2).$$

When $\lambda = \frac{2}{3}$, we only need to compare $2-\rho_1-2\rho_2$ and $2\rho_2$. The true values of $2-\rho_1-2\rho_2$ and $2\rho_2$ in simulated experiment 2 and real example 1 are obtained by conducting 100,000 independent experiments (see Tables 2 and 3, respectively). Results suggest that the *AL* estimator is conservative

Table 4: The True Values of $\rho_1$ and $\rho_2$ for Simulated Experiment 3.

| Classifier | | $n=$ 40 | 80 | 160 | 200 | 400 | 800 | 1200 | 1600 | 2000 |
|---|---|---|---|---|---|---|---|---|---|---|
| NN | $\rho_1$ | 0.328 | 0.288 | 0.248 | 0.237 | 0.199 | 0.158 | 0.135 | 0.120 | 0.103 |
| | $\rho_2$ | 0.332 | 0.310 | 0.283 | 0.276 | 0.254 | 0.251 | 0.261 | 0.271 | 0.276 |
| NB | $\rho_1$ | 0.305 | 0.203 | 0.113 | 0.092 | 0.050 | 0.029 | 0.013 | 0.013 | 0.009 |
| | $\rho_2$ | 0.334 | 0.371 | 0.400 | 0.408 | 0.434 | 0.453 | 0.457 | 0.465 | 0.468 |
| CR | $\rho_1$ | 0.233 | 0.153 | 0.121 | 0.113 | 0.099 | 0.112 | 0.108 | 0.107 | 0.092 |
| | $\rho_2$ | 0.267 | 0.198 | 0.205 | 0.206 | 0.218 | 0.245 | 0.251 | 0.250 | 0.242 |
| SVM | $\rho_1$ | 0.433 | 0.366 | 0.298 | 0.284 | 0.233 | 0.182 | 0.154 | 0.138 | 0.127 |
| | $\rho_2$ | 0.425 | 0.435 | 0.436 | 0.441 | 0.445 | 0.449 | 0.450 | 0.453 | 0.454 |

and becomes more conservative when $\lambda > \frac{2}{3}$. For example, when $\lambda = \frac{4}{3}$, $E(\widehat{\mathrm{Var}}_2(\hat{\mu}_{3\times 2})) = \frac{1}{6}\sigma_1^2(1+\rho_1) + \frac{2}{6}\sigma_1^2(3-4\rho_2)$; thus, we have $3-4\rho_2 > 2\rho_2$ if $\rho_2 < 0.5$, which is conservative. Throughout this letter, the $AL$ estimator refers to the variance estimation when $\lambda = \frac{2}{3}$.

**Simulated experiment 3.** Following the setup of simulated experiment 2, we study the changes of $\rho_1$ and $\rho_2$ with the increase in the number of samples for multiple classifiers, such as neural networks (NN), naive Bayes (NB), classification trees (CR), and support vector machine (SVM). Table 4 shows that $\rho_1$ is almost less than $\rho_2$ in addition to the case that the sample size is 40 and the classifier is SVM, $\rho_1 = 0.433 > \rho_2 = 0.425$, and $\rho_1$ and $\rho_2$ are all greater than 0 in each sample size ($0 < \rho_1 < \rho_2$). As the sample size increases, the difference between $\rho_1$ and $\rho_2$ becomes more obvious. This is because $\rho_1$ gradually decreases, while $\rho_2$ tends toward stability with the increase of $n$.

## 5 Five Tests for Comparing Classification Learning Algorithms

In this section, we present five techniques for performing statistical tests for generalization errors. The first four are methods already discussed in the literature, and the fifth is our proposed method. The test of hypothesis (excluding 5×2 cv $F$-test) has the following general form:

Null hypothesis: $H_0 : \mu(n/2) = \mu_0$ to Alternative hypothesis: $H_1 : \mu(n/2) \neq \mu_0$.

If

$$\left| \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\sigma}^2}} \right| > c, \tag{5.1}$$

reject $H_0$; otherwise, do not reject it. Note that in equation 5.1, $\hat{\mu}$ is an estimator of $\mu(n/2)$, $\hat{\sigma}^2$ is a variance estimator of $\hat{\mu}$, and $c$ is a percentile from Student's $t$ distribution. We choose a significance level of $\alpha = 0.05$ in this letter. The only difference between the four techniques is in the choice of $\hat{\mu}$, $\hat{\sigma}^2$, and $c$.

It should be noted that for a given sample size $n$, those five methods aim at inference for $\mu(n/2)$, not $\mu(n)$. Before we present these test statistics, we first introduce two concepts: a liberal test and a conservative test. A test is liberal if it rejects the null hypothesis with a probability greater than the significance level $\alpha$ whenever the null hypothesis is actually true; if the probability is smaller than the significance level $\alpha$, the test is said to be conservative. To determine whether a test is liberal or conservative, the political ratio is a useful statistic. We call $\frac{\mathrm{Var}(\hat{\mu})}{E(\hat{\sigma}^2)}$ the political ratio because it indicates that the test should be liberal when it is greater than 1 and conservative when it is less than 1. Here, we perform the analysis based on this statistic. Generally, the conservative test is preferable to the liberal test in practical applications.

We now introduce the statistics to be considered in this letter.

**5.1 *K*-Fold cv Paired *t*-Test.** Recall that $\hat{\mu}_K = \frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_k$. The $K$-fold cross-validated paired $t$-test considers $\hat{\mu} = \hat{\mu}_K$ to estimate $\mu(n/K)$ and $\hat{\sigma}^2 = \frac{S^2_{\hat{\mu}_k}}{K}$, where $S^2_{\hat{\mu}_k}$ is the sample variance of $\hat{\mu}_k$. In this letter, we aim at inference for $\mu(n/2)$. For this reason, we take $K = 2$. If we assume that $\hat{\mu}_k$ were drawn independently from a normal distribution, then the cross-validated paired $t$-test can be written as

$$t_{CV} = \frac{\hat{\mu}_2 - \mu_0}{\sqrt{S^2_{\hat{\mu}_k}/2}} \sim t_1.$$

According to Bengio and Grandvalet (2004), the political ratio is expressed as

$$\frac{\mathrm{Var}(\hat{\mu})}{E(\hat{\sigma}^2)} = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 - \sigma_2^2}.$$

Obviously the test is liberal if $\sigma_2^2 > 0$. The simulated results also show that the $\sigma_2^2$s are all greater than 0, which is in accordance with the conclusion of Grandvalet and Bengio (2006). Thus, the results based on the cross-validated $t$-test may be overconfident and lead to wrong conclusions. However, this is one of the most commonly used methods in the literature.

**5.2 5×2 cv Paired *t*-Test.** Dietterich (1998) pointed out that the variance of $K$-fold cv $t$-test can be underestimated due to the overlapping of training

sets. Thus, they proposed a 5×2 cross-validated paired $t$-test based on five replications of two-fold cross-validation. In each replication, the data are randomly partitioned into two equal-sized sets $T_1^{(i)}$ and $T_2^{(i)}, i = 1, \ldots, 5$. Each learning algorithm is trained on each set and tested on the other one. This produces cross-validated estimators $\hat{\mu}_1^{(i)}$ and $\hat{\mu}_2^{(i)}, i = 1, \ldots, 5$. (Here, we use the same mark with the blocked 3×2 cross-validation.) Let $S_i^2 = (\hat{\mu}_1^{(i)} - \hat{\mu}^{(i)})^2 + (\hat{\mu}_2^{(i)} - \hat{\mu}^{(i)})$ be the sample variance computed from the $i$th replication. He then used $\hat{\mu} = \hat{\mu}_1^{(1)}, \hat{\sigma}^2 = \frac{\sum_{i=1}^5 S_i^2}{5}$; under the assumption of normality, the resulting statistic is

$$t_{5\times 2CV} = \frac{\hat{\mu}_1^{(1)} - \mu_0}{\sqrt{\sum_{i=1}^5 S_i^2/5}} \sim t_5.$$

Note that the political ratio is given by

$$\frac{\mathrm{Var}(\hat{\mu}_1^{(1)})}{E(\hat{\sigma}^2)} = \frac{\sigma_1^2}{\sigma_1^2 - \sigma_2^2}.$$

In this work, the effect of the number of overlapped samples between cross-validations on $\mathrm{Var}(\hat{\mu})$ and $E(\hat{\sigma}^2)$ is not considered.

The above leads to liberal inference when $\sigma_2^2 > 0$. Indeed, we know that $\sigma_2^2$ is greater than 0 from the simulation of the previous section, so we think the 5×2 cv $t$-test is liberal. However, Dietterich (1998) showed that this test has an acceptable type I error and is more powerful than the $K$-fold cv $t$-test.

**5.3 Corrected *K*-Fold cv Paired *t*-Test.** Bengio and Grandvalet (2004) showed that the correlation of test blocks cannot be ignored in computing the variance of $K$-fold cross-validation; otherwise, the variance will be grossly underestimated. Based on this, Grandvalet and Bengio (2006) obtained a corrected $K$-fold cross-validated paired $t$-test by correcting the variance of $K$-fold cross-validation. If $\hat{\mu} = \hat{\mu}_K, \hat{\sigma}^2 = \frac{S_{\hat{\mu}_k}^2}{K(1-\rho_0)}$, the Bengio and Grandvalet (2004) test can be expressed as

$$t_{CCV} = \frac{\hat{\mu}_{K=2} - \mu_0}{\sqrt{S_{\hat{\mu}_k}^2/(2(1-\rho_0))}} \sim t_1$$

where $\rho_0 = \frac{\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2)}{\mathrm{Var}(\hat{\mu}_{K=2})}$ is called the correlation of blocks. According to the definition of correlated coefficient of blocks, however, it should be $\rho_1$.

Therefore, we modify this test using the relationship between $\rho_0$ and $\rho_1$, and the modified test takes the form

$$t_{CCV} = \frac{\hat{\mu}_{K=2} - \mu_0}{\sqrt{S^2_{\hat{\mu}_k} / (\frac{1-\rho_1}{1+\rho_1})}} \sim t_1,$$

where $\rho_1$ is also an unknown parameter, for which Grandvalet and Bengio (2006) suggested using $\hat{\rho}_1 = 1/2$ as a surrogate. Based on our simulated analysis, the correlated coefficient of blocks for two-fold cross-validation should be relatively small. Thus, the selection of $\hat{\rho}_1 = 1/2$ will most likely result in a conservative test.

The political ratio is

$$\frac{\mathrm{Var}(\hat{\mu})}{E(\hat{\sigma}^2)} = \frac{\frac{1+\rho_1}{1-\rho_1}}{\frac{1+1/2}{1-1/2}}.$$

Having mentioned earlier that conservative inference is preferable to liberal inference, we therefore expect that $\rho_1 < 1/2$. This is confirmed in the experiment of Grandvalet and Bengio (2006), although there is no theoretical proof for this.

**5.4 5×2 cv Paired *F*-Test.** Alpaydin (1999) pointed out that the numerator of 5×2 cv *t*-test statistic $\hat{\mu}_1^{(1)}$ is arbitrary; actually, there are 10 different values that can be placed in the numerator $\hat{\mu}_i^{(j)}, i = 1, \ldots, 5, \quad j = 1, 2$, leading to 10 possible statistics:

$$t_i^{(j)} = \frac{\hat{\mu}_i^{(j)} - \mu_0}{\sqrt{\frac{1}{5}\sum_{i=1}^{5} S_i^2}}.$$

Alpaydin then proposed a variant of the 5×2 cv *t*-test that combines multiple statistics to get a more robust test.

If $\hat{\mu}_i^{(j)}/\sigma \sim N(0,1)$, then $(\hat{\mu}_i^{(j)})^2/\sigma^2 \sim \chi_1^2$ and $\sum_{i=1}^{5}\sum_{j=1}^{2}(\hat{\mu}_i^{(j)})^2/\sigma^2$ is chi-square with 10 degrees of freedom. Therefore, we have

$$F_{5\times 2CV} = \frac{\frac{1}{10}\sum_{i=1}^{5}\sum_{j=1}^{2}(\hat{\mu}_i^{(j)} - \mu_0)^2}{\frac{1}{5}\sum_{i=1}^{5} S_i^2} \sim F_{10,5}.$$

**5.5 Blocked 3×2 cv *t*-Test.** The analysis presented in section 4.1 indicates that $\widehat{\mathrm{Var}}_1(\hat{\mu}_{3\times 2})$ cannot guarantee that it is always greater than or equal

Table 5: Summary Description of the Test Methods.

| Name | $\hat{\mu}$ | $\hat{\sigma}^2$ | $c$ | $\frac{\mathrm{Var}(\hat{\mu})}{E(\hat{\sigma}^2)}$ |
|---|---|---|---|---|
| $K$-fold cv $t$-test | $\hat{\mu}_{K=2}$ | $S^2_{\hat{\mu}_k}/(2)$ | $t_{1,1-\alpha/2}$ | $\frac{1+\rho_1}{1-\rho_1}$ |
| 5×2 cv $t$-test | $\hat{\mu}_1^{(1)}$ | $\sum_{i=1}^5 S_i^2/5$ | $t_{5,1-\alpha/2}$ | $\frac{1}{1-\rho_1}$ |
| Corrected $K$-fold cv $t$-test | $\hat{\mu}_{K=2}$ | $S^2_{\hat{\mu}_k}/(1/3)$ | $t_{1,1-\alpha/2}$ | $\frac{\frac{1+\rho_1}{1-\rho_1}}{\frac{1+1/2}{1-1/2}}$ |
| Blocked 3×2 cv $t$-test($\lambda=\frac{2}{3}$) | $\hat{\mu}_{3\times2}$ | $\widehat{Var}_2(\hat{\mu}_{3\times2})$ | $t_{5,1-\alpha/2}$ | $\frac{1+\rho_1+4\rho_2}{5-\rho_1-4\rho_2}$ |
| Blocked 3×2 cv $t$-test($\lambda=\frac{4}{3}$) | | | | $\frac{1+\rho_1+4\rho_2}{7+\rho_1-8\rho_2}$ |

Note: $t_{q,p}$ refer to the quantile $p$ of the $t_q$ distribution.

to $\mathrm{Var}(\hat{\mu}_{3\times2})$ , that is, the $L$ estimator may overestimate or underestimate the true variance. For this reason, we propose a new conservative variance estimation by appending a corrected term based on the sample variance of $\hat{\mu}^{(i)}$ to the $L$ estimator. This leads to

$$t_{B3\times2CV} = \frac{\hat{\mu}_{3\times2} - \mu_0}{\sqrt{\widehat{Var}_2(\hat{\mu}_{3\times2})}} \sim t_5.$$

We summarize the tests in Table 5.

## 6 Simulated Experiment Study

In this section, we perform a simulation study to investigate the probability of type I error and the power of the five statistics considered in the previous section. For a given problem, we generate 1000 independent data sets to fully take into account the effect of the randomness of the training set as well as test examples.

**6.1 Classification of Two Gaussian Populations.** Considering the problem of estimating the generalization error in a classification problem with two classes, we thus have $Z = (X, Y)$, with $Prob(Y = 1) = Prob(Y = 0) = \frac{1}{2}$, $X|Y = 0 \sim N(\mu_0, \Sigma_0)$, $X|Y = 1 \sim N(\mu_1, \Sigma_1)$. The classification algorithms are:

- *Regression tree*. We train a least square regression tree, and the decision function is $F_A(Z_S)(X) = I[N_{Z_S}(X) > 0.5]$, where $N_{Z_S}(X)$ is the leaf value corresponding to $X$ of the tree obtained when training on

$Z_S$. Thus, $L_A(j,i) = I[F_A(Z_{S_j})(X_i) \neq Y_i]$ is equal to 1 whenever this algorithm misclassifies example $i$; otherwise, it is 0.

- *Ordinary least squares linear regression*. We perform the regression of $Y$ against $X$, and the decision function is $F_B(Z_S)(X) = I[\hat{\beta}_{Z_S}^T X > 0.5]$, where $\hat{\beta}_{Z_S}$ is the ordinary least squares regression coefficient estimates. Thus, $L_B(j,i) = I[F_B(Z_{S_j})(X_i) \neq Y_i]$ is equal to 1 whenever this algorithm misclassifies example $i$; otherwise, it is 0.

**6.2 Classification of Letters.** We consider the problem of estimating generalization errors in the letter recognition classification problem. The classification algorithms are:

- *Classification tree*. We train a classification tree, and the decision function is $F_A(Z_S)(X)$. Here, the classification loss function $L_A(j,i) = I[F_A(Z_{S_j})(X_i) \neq Y_i]$ is equal to 1 whenever this algorithm misclassifies example $i$; otherwise, it is 0.
- *First nearest neighbor*. We apply the first nearest neighbor rule with a distorted distance metric to perform classification. Specifically, the distance between two vectors of inputs is

$$d(X^{(1)}, X^{(2)}) = \sum_{k=1}^{3} \omega^{2-k} \sum_{i \in C_k} (X^{(1)} - X^{(2)})^2,$$

where $C_1 = \{1, 3, 9, 16\}, C_2 = \{2, 4, 6, 7, 8, 10, 12, 14, 15\}, C_3 = \{5, 11, 13\}$ denote the sets of components that are weighted by $\omega, 1, \omega^{-1}$, respectively.

**6.3 Type I Error Results.** First, from the conclusion of Nadeau and Bengio (2003), we know that the two classification learning algorithms have no statistical significant differences with the setups of Tables 6 and 7, which cannot reject the null hypothesis. Tables 6 and 7 show the results of the type I error rates. The *K*-fold cv *t*-test and the corrected *K*-fold cv *t*-test exhibit somewhat elevated probabilities of type I error. One example is the situation represented by simulation 1 of Table 6. In this case, the probabilities of type I error of $t_{CV}$ and $t_{CCV}$ are 0.08 and 0.07. However, the 5×2 cv *t*-test, the 5×2 cv *F*-test, and blocked 3×2 cv *t*-test have acceptable type I errors. For example, in Table 7, the probability of a type I error of the 5×2 cv *t*-test is 0.05, the 5×2 cv *F*-test is 0.04, and the blocked 3×2 cv *t*-test is 0.02.

**6.4 Power Measure of Tests.** A type I error is not the only important consideration in choosing a statistical test. When the probability of type I error is comparative, the test is always measured by the power function.

Figures 6, 7, 8, 9, 10, and 11 plot the power curves for the five statistical tests. These curves show that the blocked 3×2 cv *t*-test is the most powerful,

Table 6: Probability of Type I Error for Each Statistical Test in the Classification of Two Gaussian Populations.

| | Simulation 1 | Simulation 2 |
|---|---|---|
| n | 200 | 2000 |
| $\mu_0$ | (0,0) | (0,0) |
| $\mu_1$ | (1,1) | (1,1) |
| $\Sigma_0$ | $I_2$ | $I_2$ |
| $\Sigma_1$ | $\frac{1}{6}$ $I_2$ | 0.173 $I_2$ |
| Probability of type I error | | |
| $t_{CV}$ | 0.08 | 0.07 |
| $t_{CCV}$ | 0.07 | 0.05 |
| $t_{5\times2CV}$ | 0.05 | 0.04 |
| $F_{5\times2CV}$ | 0.04 | 0.05 |
| $t_{B3\times2CV}$ | 0.05 | 0.05 |

Table 7: Probability of Type I Error for Each Statistical Test in the Classification of Letters.

| | |
|---|---|
| n | 300 |
| $\omega$ | 25 |
| Probability of type I error | |
| $t_{CV}$ | 0.08 |
| $t_{CCV}$ | 0.06 |
| $t_{5\times2CV}$ | 0.05 |
| $F_{5\times2CV}$ | 0.04 |
| $t_{B3\times2CV}$ | 0.02 |

the corrected $K$-fold cv $t$-test is the least powerful procedure, and the $K$-fold cv, 5×2 cv $t$-test and the 5×2 cv $F$-test are in between. The corrected $K$-fold cv $t$-test reduces the type I error of the $K$-fold cv $t$-test through correction; however, there is a price to be paid for this reduction: lowered power. The blocked 3×2, 5×2 cv $t$-tests and the 5×2 cv $F$-test are much more powerful than the $K$-fold cv and corrected $K$-fold cv $t$-test, although they look similar. Especially for Figures 8 and 11, the power curves of blocked 3×2 cv, 5×2 cv $t$-tests, and the 5×2 cv $F$-test are almost overlapping. We also note that the two-fold cross-validation is executed only three times for the blocked 3×2 cv $t$-test (compared with five times for the 5×2 cv $t$-test and the 5×2 cv $F$-test). Therefore, it can result in valuable savings, especially when the computational cost of running the learning algorithm is high.

**6.5 Classification of Real Data Sets.** In this section, we carry out experiments on six data sets from the UCI repository. Results show that our test also has good performance.

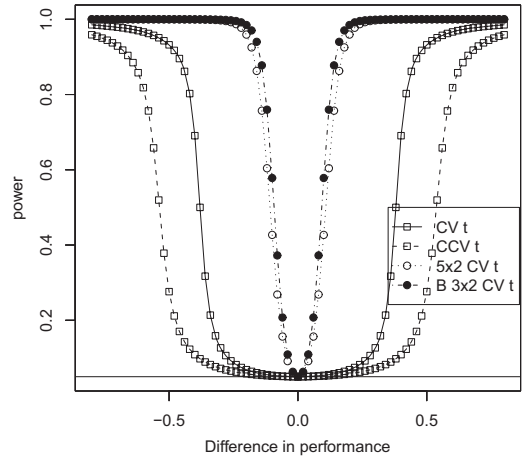


Figure 6: Powers of the tests for the classification of two gaussian populations: simulation 1.

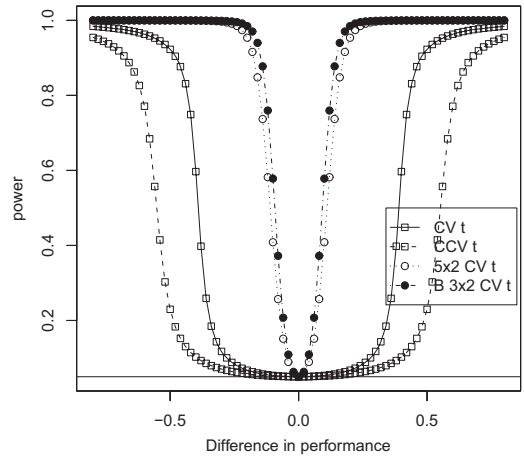


Figure 7: Powers of the tests for the classification of two gaussian populations: simulation 2.

Similar to Alpaydin's (1999) work, to compare type I error of the five tests, we use two MLPs (multilayer perceptrons with one hidden layer) with equal numbers of hidden units. Thus, the null hypothesis is true, and any rejection is a type I error. To compare type II errors of the five tests, we take two classifiers that are different: an LP (single-layer perceptron) and an MLP.

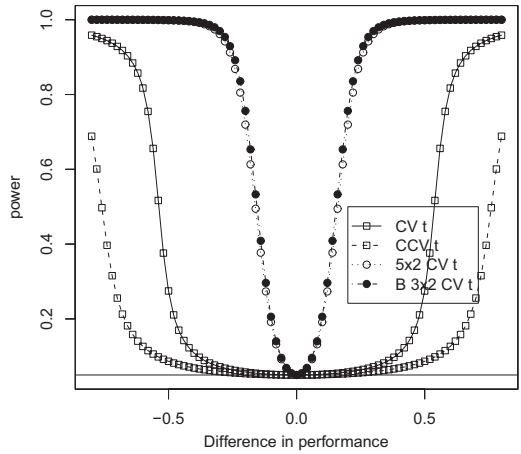


Figure 8: Powers of the tests for the classification of letters.

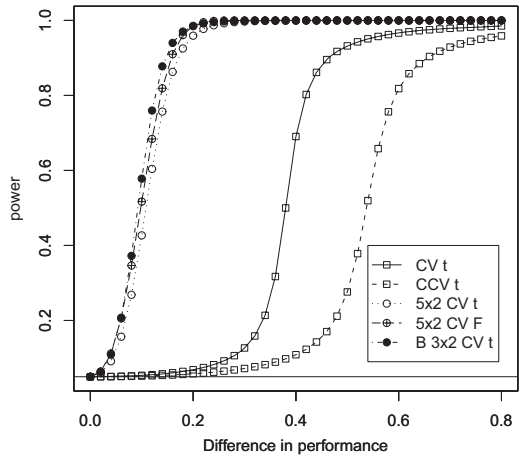


Figure 9: Powers of the tests for the classification of two gaussian populations: simulation 1.

From Tables 8 and 9, we know that in most cases, the blocked $3\times2$ cross-validation has lower type I and type II errors than other commonly used tests.

**6.6 Replicability of Tests.** Bouckaert and Frank (2004) considered the issue that a test may be very sensitive to the random partitioning used in cross-validation. If this is the case, it is possible that when the same data are used—the same learning algorithms $A$ and $B$ and the same significance

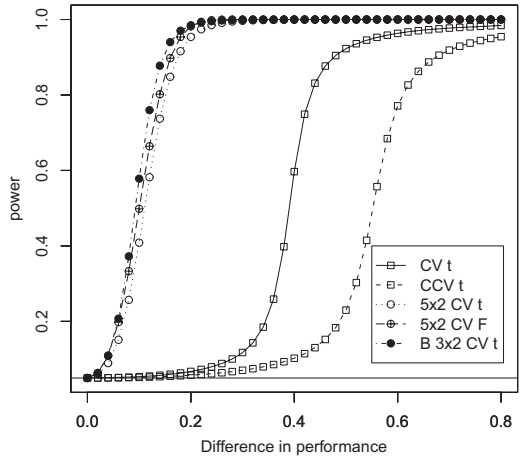


Figure 10: Powers of the tests for the classification of two gaussian populations: simulation 2.

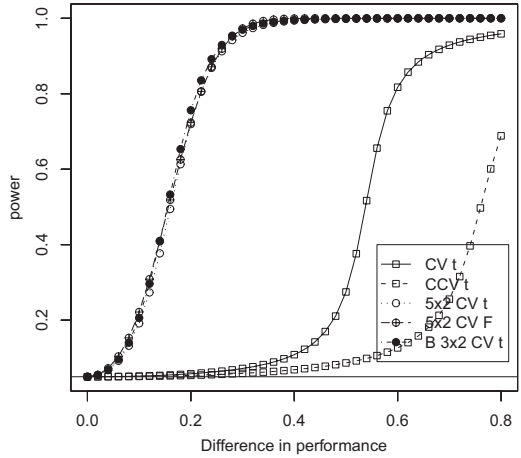


Figure 11: Powers of the tests for the classification of letters.

test—one researcher may find that method *A* is preferable, whereas another may find that the evidence for this is not enough. Lack of replicability can also cause problems when tuning an algorithm: a test may judge favorably on the latest modification purely due to its sensitivity to the particular random number seed used to split the data. And Bouckaert (2005) shows that low replicability of machine learning experiments is not a phenomenon of small data sets. In this section, we study the replicability of five tests in a

Table 8: Probabilities of Type I Error.

| Data Set | Hidden Units | $t_{CV}$ | $t_{CCV}$ | $t_{5\times2CV}$ | $F_{5\times2CV}$ | $t_{B3\times2CV}$ |
|---|---|---|---|---|---|---|
| Iris | 3 | 0.145 | 0.130 | 0.030 | 0.010 | 0.001 |
| | 10 | 0.205 | 0.205 | 0.049 | 0.006 | 0.003 |
| | 20 | 0.244 | 0.244 | 0.051 | 0.005 | 0.004 |
| Vowel | 5 | 0.034 | 0.015 | 0.029 | 0.013 | 0.001 |
| | 10 | 0.050 | 0.026 | 0.037 | 0.010 | 0.002 |
| | 20 | 0.042 | 0.018 | 0.029 | 0.005 | 0.000 |
| Thyroid | 10 | 0.044 | 0.027 | 0.019 | 0.002 | 0.000 |
| Seed | 10 | 0.077 | 0.076 | 0.038 | 0.007 | 0.002 |
| | 20 | 0.081 | 0.081 | 0.029 | 0.008 | 0.001 |
| Heart-statlog | 5 | 0.100 | 0.083 | 0.026 | 0.012 | 0.002 |
| | 10 | 0.050 | 0.032 | 0.021 | 0.001 | 0.001 |
| | 20 | 0.033 | 0.013 | 0.035 | 0.009 | 0.001 |
| Balance | 5 | 0.051 | 0.031 | 0.036 | 0.005 | 0.001 |
| | 10 | 0.056 | 0.039 | 0.030 | 0.007 | 0.001 |
| | 20 | 0.053 | 0.043 | 0.032 | 0.009 | 0.000 |

Table 9: Probabilities of Type II Error.

| Data Set | Hidden Units | $t_{CV}$ | $t_{CCV}$ | $t_{5\times2CV}$ | $F_{5\times2CV}$ | $t_{B3\times2CV}$ |
|---|---|---|---|---|---|---|
| Iris | 3 | 0.142 | 0.127 | 0.076 | 0.029 | 0.000 |
| | 10 | 0.187 | 0.186 | 0.034 | 0.003 | 0.002 |
| | 20 | 0.148 | 0.148 | 0.032 | 0.003 | 0.006 |
| Vowel | 5 | 0.129 | 0.057 | 0.181 | 0.131 | 0.066 |
| | 10 | 0.063 | 0.028 | 0.104 | 0.043 | 0.014 |
| | 20 | 0.042 | 0.016 | 0.052 | 0.012 | 0.006 |
| Thyroid | 10 | 0.075 | 0.053 | 0.085 | 0.037 | 0.011 |
| Seed | 10 | 0.092 | 0.086 | 0.081 | 0.022 | 0.012 |
| | 20 | 0.089 | 0.083 | 0.054 | 0.021 | 0.015 |
| Heart-statlog | 5 | 0.082 | 0.063 | 0.034 | 0.011 | 0.003 |
| | 10 | 0.074 | 0.046 | 0.022 | 0.009 | 0.004 |
| | 20 | 0.067 | 0.032 | 0.033 | 0.011 | 0.021 |
| Balance | 5 | 0.054 | 0.038 | 0.050 | 0.016 | 0.011 |
| | 10 | 0.119 | 0.043 | 0.143 | 0.089 | 0.133 |
| | 20 | 0.147 | 0.049 | 0.259 | 0.198 | 0.312 |

realistic setting based on standard data sets taken from the UCI repository of machine learning problems.

Bouckaert and Frank (2004) first gave the definition of replicability based on the probability that two runs of the test on the same data set will produce the same outcome. If we have performed the test based on $n$ different randomizations for a particular data set, then $\binom{n}{2}$ such pairs are found. Assume that the test rejects the null hypothesis for $k(0 \leq k \leq n)$ of the

Table 10: Replicability for Five Tests.

| | | | NB vs LDA | | | | |
|---|---|---|---|---|---|---|---|
| dataset | #inst. | #atts. | $t_{CV}$ | $t_{CCV}$ | $t_{5\times2CV}$ | $F_{5\times2CV}$ | $t_{B3\times2CV}$ |
| balance_scale | 625 | 4 | 3 | 3 | 2 | 0 | 0 |
| diabetes | 768 | 8 | 4 | 3 | 6 | 1 | 0 |
| glass | 270 | 13 | 7 | 3 | 12 | 11 | 17 |
| heart | 352 | 34 | 1 | 1 | 1 | 0 | 0 |
| ionosphere | 150 | 4 | 2 | 0 | 1 | 0 | 2 |
| irirs | 148 | 18 | 4 | 4 | 1 | 0 | 0 |
| vehicle | 846 | 19 | 28 | 12 | 50 | 50 | 50 |
| wine | 178 | 13 | 5 | 5 | 2 | 0 | 1 |
| yeast | 1484 | 9 | 1 | 0 | 0 | 0 | 0 |
| seed | 210 | 7 | 10 | 2 | 28 | 31 | 44 |
| Replicability | | | 0.817 | 0.883 | 0.863 | 0.913 | 0.921 |
| Replicability: NB versus Tree | | | 0.784 | 0.855 | 0.857 | 0.925 | 0.910 |
| Replicability: Tree versus LDA | | | 0.767 | 0.860 | 0.798 | 0.925 | 0.921 |

Note: #inst. (#atts.) refers to the number of cases (attributes).

randomizations. Then there are $\binom{k}{2}$ rejecting pairs and $\binom{n-k}{2}$accepting ones. Based on this, the above probability can be estimated as $R(k,n) = (\binom{k}{2} + \binom{n-k}{2})/\binom{n}{2}$ . We can use this probability to form a measure of replicability across different data sets. Assume that the number of data sets is $m$, and let $i_k (0 \le k \le n)$ be the number of data sets for which the test agrees $k$ times (i.e., $\sum_{k=0}^{n} i_k = m$). Then we define replicability as $R = \sum_{k=0}^{n} \frac{i_k}{m} R(k,n)$.

To evaluate how replicability affects various tests, we performed experiments on 10 data sets from the UCI repository (see Table 10). We used naive Bayes (NB), classification tree (CR), and linear discriminant analysis (LDA). Each test was run 50 times for each pair of learning schemes, and a 5% significance level was used in all tests.

In our experiments, good replicability was obtained using our test and the 5×2 cross-validation *F*-test. The value of the replicability measure $R$ is above 0.9 for these two tests.

## 7 Conclusion

A new blocked 3×2 cross-validation method is given in our letter. In detail, the method splits a data set into four balanced blocks; two are the training set and the other two are a test set to implement the two-fold cross-validation, resulting in three replications of two-fold cross-validations that are performed. For the six data sets obtained, any two data sets between different groups (either the training sets or test sets) have the same number of overlapped samples and better sample balance. Essentially, through a preliminary design for partitions, this method reduces errors resulting from

random partitions between different groups; thus, the blocked $3\times2$ cross-validation estimator of generalization error has a smaller variance. Then a variance estimator is given. We show using simulations that the $t$-test using this new variance estimator has better performance. We compared the new method with the $5\times2$ cross-validated $t$-test ($F$-test) recommended by Dietterich (1998) and Alpaydin (1999) and found that the performance of our test is comparable to that of $5\times2$ cross-validated tests but with less computational complexity. The computational cost of running the learning algorithm is higher, especially in natural language processing. In this case, in order to obtain the balance of corpus and number of overlapped samples, the corpus partition is easier to implement using blocked $3\times2$ cross-validation than random two-fold partition. This method has been applied to semantic role labeling task in natural language processing and has good results.

Based on this idea, several natural problems are found. If five replications (or more) of two-fold cross-validations are performed, how can data be split such that each overlaps between the data with the same number (i.e., data sets between different groups yield balance property)? Can the balanced cross-validation with minimum variance be proved? For a given data set, how many balanced partitions are there? How can these balanced partition be identified? These questions are worthy of further study.

## Acknowledgments

We thank Wanquan Liu and Chaohua Dong for constructive discussions. This work was supported by National Natural Science Fund of China (60873128). Experiments were supported by High Performance Computing System of Shanxi University.

## References

Alpaydin, E. (1999). Combined $5\times2$ cv $F$ test for comparing supervised classification learning algorithms. *Neural Computation*, *11*, 1885–1892.

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). Cambridge, MA: MIT Press.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of $K$-fold cross-validation. *Journal of Machine Learning Research*, *5*, 1089–1105.

Bouckaert, R. R. (2005). Low replicability of machine learning experiments is not a small data set phenomenon. In *Proceedings of the ICML-2005 Workshop on Meta-learning*.

Bouckaert, R. R., & Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In *Proceedings of the 8th Pacific-Asia Conference* (pp. 3–12). Berlin: Springer.

Chen, W., Gallas, B. D., & Yousef, W. A. (2012). Classifier variability: Accounting for training and testing. *Pattern Recognition*, *45*, 2661–2671.

Dietterich, T. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*, 1895–1924.

Grandvalet, Y., & Bengio, Y. (2006). *Hypothesis testing for cross-validation* (Tech. Rep. 1285). Montreal: University of Montreal.

Li, J. (2010). *Research on techniques of automatic semantic role labeling of Chinese frame-net*. Unpublished doctoral dissertation, Shanxi University.

Li, J., Wang, R., Wang, W., & Li, G. (2010). Automatic labeling of semantic roles on Chinese frame-net. *Journal of Software*, *30*, 597–611.

Markatou, M., Tian, H., Biswas, S., & Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error. *Journal of Machine Learning Research*, *6*, 1127–1168.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*, 239–281.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 42–49). New York: ACM.

Yildiz, O. T. (2013). Omnivariate rule induction using a novel pairwise statistical test. *IEEE Transactions on Knowledge and Data Engineering*, *25*, 2105–2118.

Received July 16, 2012; accepted July 21, 2013.