Choosing Between Two Classification Learning Algorithms Based on Calibrated Balanced $5\times 2$ Cross-Validated $F$-Test

Yu Wang$^1$ · Jihong Li$^{1,2}$ · Yanfang Li$^2$


Abstract 5 × 2 cross-validated $F$-test based on independent five replications of 2-fold cross-validation is recommended in choosing between two classification learning algorithms. However, the reusing of the same data in a 5 × 2 cross-validation causes the real degree of freedom (DOF) of the test to be lower than the $F(10, 5)$ distribution given by (Neural Comput 11:1885–1892, [1]). This easily leads the test to suffer from high type I and type II errors. Random partitions for 5 × 2 cross-validation result in difficulty in analyzing the DOF for the test. In particular, Wang et al. (Neural Comput 26(1):208–235, [2]) proposed a new blocked $3\times 2$ cross-validation, that considered the correlation between any two 2-fold cross-validations. Based on this, a calibrated balanced 5 × 2 cross-validated $F$-test following $F(7, 5)$ distribution is put forward in this study by calibrating the DOF for the $F(10, 5)$ distribution. Simulated and real data studies demonstrate that the calibrated balanced 5 × 2 cross-validated $F$-test has lower type I and type II errors than the 5 × 2 cross-validated $F$-test following $F(10, 5)$ in most cases.

Keywords Test · Type I error · Type II error · Cross-validation · Classification learning algorithm

1 Introduction

In machine learning research, when proposing a new classification learning algorithm, we usually need to compare its performance with the previous best algorithms. However, choosing between two learning algorithms with a data set is not a simple task. The most straightforward approach is to use statistical tests of significance to determine whether a new algorithm performs better than previous ones. Given two classification algorithms and $n$ samples, we first train the two algorithms on the same training data set and then validate whether the difference
between the two algorithms is significant or not on the same test set. If we assume that the two algorithms have no difference, we can construct a test statistic under this assumption and check the probability of rejecting this hypothesis based on the distribution of the test statistic. If this probability is sufficiently high, we accept the hypothesis, otherwise we reject it. The rejected probability $\alpha$ is called the probability of type I error when no difference exists, and the accepted probability $\beta$ is called the probability of type II error when a difference exists.

Cross-validation through training and testing data a number of times is always conducted to estimate the performance of an algorithm and therefore reduce the effect of random error. Numerous tests based on cross-validation used in choosing algorithms have been proposed, such as K-fold cross-validated $t$-test [3–6], $5 \times 2$ cross-validated $t$-test and $F$-test [1,7,8], blocked $3 \times 2$ cross-validated $t$-test [2]. Other related literatures refer to [9–14] and [15].

Dietterich [7] proposed a $5 \times 2$ cross-validated $t$-test based on five group replications of 2-fold cross-validation and demonstrated that its performance is better than 10-fold cross validation by simulated experiments. The numerator of the $5 \times 2$ cross-validated $t$-test statistic is arbitrary, actually there are ten different values that can be placed in the numerator. In addition, changing the numerator corresponds to changing the order of replications or folds and should not affect the test result. However, [1] proved the opposite. Alpaydin then constructed a variant of $5 \times 2$ cross-validated $t$-test, combined $5 \times 2$ cross-validated $F$-test, and demonstrated that the variant was more powerful than the $5 \times 2$ cross-validated $t$-test. The $5 \times 2$ cross-validated test statistic following $F(10, 5)$ distribution is drawn from the independence assumption of replications and folds. However, training (test) sets from any two independent partitions contain common samples regardless of how the data is split. Thus, cross-validation estimators with different data partitions are actually not independent. The $5 \times 2$ cross-validated $F$-test discussed by [1] did not consider the correlation among five 2-fold cross-validations. Bouckaert [16] pointed out that reusing of the same data in cross-validation would make the DOF of the corresponding test statistic lower than the theoretically expected number. This easily leads the test to suffer from high type I and type II errors (see [13]). Random partitions for $5 \times 2$ cross-validation result in difficulty in analyzing the DOF for the test. Thus, the correlation of any two 2-fold cross-validations should be considered in studying the distribution of $5 \times 2$ cross-validated $F$-test statistic such that the distribution can be closer to the true distribution (DOF of distribution) and more accurate analysis of type I and type II errors can be conducted. In particular, [2] proposed a new blocked $3 \times 2$ cross-validation, that considered the correlation between any two 2-fold cross-validations. Furthermore, [13] studied the effect of the correlation between any two 2-fold cross-validations in a $m \times 2$ cross-validation on the performance of algorithm.

In this paper, based on the blocked $3 \times 2$ cross-validation given by [2], we propose a calibrated balanced $5 \times 2$ cross-validated $F$-test following $F(7, 5)$ distribution by calibrating the DOF for the $F(10, 5)$ distribution. Simulated and real data studies demonstrate that the calibrated balanced $5 \times 2$ cross-validated $F$-test has lower type I and type II errors than the $5 \times 2$ cross-validated $F$-test following $F(10, 5)$ in most cases.

## 2 Calibrated Balanced 5 × 2 Cross-Validated $F$-Test

### 2.1 5 × 2 Cross-Validated $F$-Test

Dietterich [7] pointed out that the variance of $K$-fold cross-validated $t$-test could be underestimated because of the overlapping of training sets. The so-called $K$-fold cross-validation
Choosing Between Two Classification Learning Algorithms...

is that the data set is split into \( K \) disjoint and equal-sized subsets, and \( K - 1 \) subsets are used to train and one subset is used to test. This process is replicated \( K \) times, it is obvious that the overlapping of any two training sets are \( K - 2 \). Based on this, Dietterich proposed a \( 5 \times 2 \) cross-validated paired \( t \)-test based on five replications of 2-fold cross-validation. In each replication, the available data is randomly partitioned into two equal-sized sets, \( S_1^{(i)} \) and \( S_2^{(i)} \), \( i = 1, \ldots, 5 \). Each learning algorithm is trained on each set and tested on the other set to produce cross-validated estimators \( \hat{\mu}_1^{(i)} \) and \( \hat{\mu}_2^{(i)} \), \( i = 1, \ldots, 5 \), where

\[
\hat{\mu}_1^{(i)} = \frac{2}{n} \sum_{z_j \in S_2^{(i)}} L(A(S_1^{(i)}), z_j), \quad \hat{\mu}_2^{(i)} = \frac{2}{n} \sum_{z_j \in S_1^{(i)}} L(A(S_2^{(i)}), z_j), \quad L(., .) \text{ represents } \{0, 1\}-loss function, \ n \text{ is sample size. Let } S_i^2 = (\hat{\mu}_1^{(i)} - \hat{\mu}^{(i)})^2 + (\hat{\mu}_2^{(i)} - \hat{\mu}^{(i)})^2 \text{ be the sample variance computed from the } i-th \text{ replication, where } \hat{\mu}^{(i)} = \frac{\hat{\mu}_1^{(i)} + \hat{\mu}_2^{(i)}}{2}. \text{ He then used }
\]

\[
t = \frac{\hat{\mu}}{\sqrt{\hat{\sigma}^2}} = \frac{\hat{\mu}_1^{(i)}}{\sqrt{\sum_{i=1}^5 S_i^2 / 5}} \text{ approximately follows a } t \text{ distribution with DOF } 5.
\]

Alpaydin[1] pointed out that the numerator of \( 5 \times 2 \) cross-validated \( t \)-test statistic \( \hat{\mu}_1^{(1)} \) is arbitrary; actually there are ten different values that can be placed in the numerator \( \hat{\mu}_k^{(i)} \), \( i = 1, \ldots, 5, \ k = 1, 2 \), leading to ten possible statistics

\[
t_k^{(i)} = \frac{\hat{\mu}_k^{(i)}}{\sqrt{\sum_{i=1}^5 S_i^2 / 5}} \quad (1)
\]

Alpaydin then proposed a variant of \( 5 \times 2 \) cross-validated \( t \)-test, that combines multiple statistics to get a more robust test, denoting

\[
F_{5 \times 2CV} = \frac{\sum_{i=1}^5 \sum_{k=1}^2 (\hat{\mu}_k^{(i)})^2 / 10}{\sum_{i=1}^5 S_i^2 / 5} \quad (2)
\]

2.2 Balanced \( 5 \times 2 \) Cross-Validated \( F \)-Test

Alpaydin[1] showed that \( F_{5 \times 2CV} \) followed an \( F \) distribution with DOFs of 10 and 5. However, this conclusion was obtained based on the following assumption: all \( \hat{\mu}_k^{(i)} \)’s were independent for \( i = 1, \ldots, 5, \ k = 1, 2 \). Although random partitions for \( 5 \times 2 \) cross-validation are independent, training (test) sets from any two independent partitions contain common samples regardless of how the data is split. It causes the real DOF of the test to be lower than the \( F(10, 5) \) distribution given by [1]. This would affect the type I and type II errors of the test (see [13]). Random partitions for \( 5 \times 2 \) cross-validation result in difficulty in analyzing the DOF for the test. Wang et al. [2] pointed out that the dependence between any two 2-fold cross-validations is related to the number of overlapped samples between training sets and reaches the minimum when the number of overlapped samples is \( \frac{n}{2} \) (\( n \) is the sample size). Then they proposed a new blocked \( 3 \times 2 \) cross-validation with the same number of overlapped samples. Based on this, a new balanced \( 5 \times 2 \) cross-validated \( F \)-test is proposed in this study, that can easily conduct a theoretical analysis of the correlation and the DOF of the test.

Blocked \( 3 \times 2 \) cross-validation is constructed as follows: the 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 as displayed in Table 1. It is obvious that there is one, and only one overlapped block in any two combinations between different groups. Furthermore, we find that by exchanging the first half of \( P_1 \) for the first half of \( P_2 \) and the first half of \( P_3 \) for the first half of \( P_4 \), four new blocks (\( P_j, \ j = 1, 2, 3, 4 \)
Remark 2 Although our method has a similar form with the Alpaydin’s 5 × 2 cross-validated F-test, the partitions are different for these two tests. A poor partition may result in a large difference in the test statistics, which is not desirable.
Choosing Between Two Classification Learning Algorithms...

variance, as well as large type I and type II errors for $5 \times 2$ cross-validated $F$-test as shown in [13]. However, balanced partitions for our method guarantee that it has minimum variance.

### 2.3 Calibration for the DOF of the Balanced $5 \times 2$ Cross-Validated $F$-Test

Next, we examine the distribution of the $F_{B5\times2CV}$ and calibrate the DOF of the distribution.

**Proposition** If we assume that $\hat{\mu}_{B_k}^{(i)} / \sigma \sim N(0, 1)$, group-in covariance $\text{Cov}(\hat{\mu}_{B_k}^{(i)}, \hat{\mu}_{B_{k'}}^{(i)}) \equiv \sigma^2 \rho_1, \ k \neq k'$, and group-out covariance $\text{Cov}(\hat{\mu}_{B_k}^{(i)}, \hat{\mu}_{B_k}^{(i')}) \equiv \sigma^2 \rho_2, \ i \neq i', \ k = k'$ or $k \neq k'$, $i, i' = 1, \ldots, 5$, $k, k' = 1, 2$, then

$$F = \frac{(1 - \rho_1) \sum_{i=1}^{5} \sum_{k=1}^{2} (\hat{\mu}_{B_k}^{(i)})^2}{\sum_{i=1}^{5} (\hat{\mu}_{B_1}^{(i)} - \hat{\mu}_{B_2}^{(i)})^2} = \frac{1 - \rho_1}{2} \frac{\sum_{i=1}^{5} \sum_{k=1}^{2} (\hat{\mu}_{B_k}^{(i)})^2}{\sum_{i=1}^{5} S_{B_i}^2} \sim F(f, 5), \quad (5)$$

where $0 \leq \rho_1 \leq \rho_2 < 0.50$, $S_{B_i}^2$ is sample variance, $f \equiv \frac{10}{1 + \rho_1^2 + 8 \rho_2^2}$.

The proof is provided in the Appendix.

**Remark 3** Wang et al. [2] shows that the $\rho_1$ is almost less than $\rho_2$, and $\rho_1$ and $\rho_2$ are all greater than 0 and less than 0.5 by simulated experiments with multiple classifiers and sample sizes. This implies that the $0 \leq \rho_1 \leq \rho_2 < 0.50$ is reasonable.

**Remark 4** This proposition indicates that the test statistic $F$ decreases with increasing group-in correlation coefficient $\rho_1$. However, group-in and group-out correlation coefficients $\rho_1$ and $\rho_2$ affect the DOF of the distribution. Table 2 presents the results.

Table 2 shows the change in the DOF of the distribution $f$ with the changes in $\rho_1$ and $\rho_2$ ($\rho_1 \leq \rho_2$) from 0 to 0.50, where “−” represents the condition $\rho_1 \leq \rho_2$ unsatisfied. We can obtain two conclusions from Table 2. First, the DOF of the distribution $f$ reaches the maximum of 10 when $\rho_1 = \rho_2 = 0$ and reaches the minimum when $\rho_1 = \rho_2 = 0.50$. Second, the DOF of $f$ decreases with increasing $\rho_1$ and $\rho_2$, but the decrease with the change

<table>
<thead>
<tr>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.05</td>
<td>9.780</td>
</tr>
<tr>
<td>0.15</td>
<td>8.316</td>
</tr>
<tr>
<td>0.20</td>
<td>7.353</td>
</tr>
<tr>
<td>0.25</td>
<td>6.400</td>
</tr>
<tr>
<td>0.30</td>
<td>5.525</td>
</tr>
<tr>
<td>0.35</td>
<td>4.756</td>
</tr>
<tr>
<td>0.40</td>
<td>4.098</td>
</tr>
<tr>
<td>0.45</td>
<td>3.543</td>
</tr>
<tr>
<td>0.50</td>
<td>3.077</td>
</tr>
</tbody>
</table>
of $\rho_2$ is faster. For example, the DOF of $f$ decreases from 5.814 to 5.525 with increasing $\rho_1$ from 0.00 to 0.30 and $\rho_2 = 0.30$. However, $f$ changes from 7.353 to 3.289 with a change in $\rho_2$ from 0.20 to 0.50 and $\rho_1 = 0.20$. The difference 4.064 is significantly more than 0.289.

The previous analysis and proposition indicated that the test statistic remains unchanged with increasing group-out correlation coefficient. However, the DOF of $f$ with varying group-out correlation coefficient decreases more rapidly. They all decrease with increasing group-in correlation coefficient. This decrease is slower than the change in the DOF of $f$ with varying group-out correlation coefficient. In other words, the group-in correlation coefficient significantly affects the test statistic, whereas the group-out correlation coefficient affects the change in the DOF of the distribution of the test statistic.

For this reason, let $\rho_1 = 0$, $0 < \rho_2 < 0.50$, namely, group-in $\hat{\mu}_{B_k}^{(i)}$s are independent, but group-out correlation coefficients of $\hat{\mu}_{B_k}^{(i)}$s lay between 0 and 0.50. In this case, $f = \frac{10}{1 + 8\rho_2^2}$,

$$F = F_{B_5 \times 2CV} = \frac{1}{2} \sum_{i=1}^{5} \sum_{k=1}^{2} (\hat{\mu}_{B_k}^{(i)})^2 = \frac{1}{2} \sum_{i=1}^{5} S_{B_i}^2 = F(f, 5) \sim F(10, 5). \quad (6)$$

Here, the problem is that $\rho_2$ is unknown and difficult to estimate [2, 6]. Given that $f$ changes with the $\rho_2$, we consider using the mean of $\rho_2$ in interval (0, 0.50) as the value of $f$ as follows:

$$f = \int_{0}^{0.50} \frac{10}{1 + 8\rho_2^2} \frac{d\rho_2}{0.50} = \frac{20}{2 \sqrt{2}} \int_{0}^{\sqrt{2}} \frac{1}{1 + x^2} dx = 5 \sqrt{2} \arctan \sqrt{2} = 6.76 \approx 7 \quad (7)$$

Remark 5 We say that this substitute is gross, but it is at least better than $\rho_2 = 0$. When $\rho_1 = 0$, $\rho_2 = 0$, all $\hat{\mu}_{B_k}^{(i)}$s including group-in and group-out are independent of one another, and therefore lead to $f = 10$ and

$$F = \frac{1}{2} \sum_{i=1}^{5} \sum_{k=1}^{2} (\hat{\mu}_{B_k}^{(i)})^2 = \frac{1}{2} \sum_{i=1}^{5} S_{B_i}^2 = F(10, 5).$$

In this case, the statistic $F$ has a same form with the $5 \times 2$ cross-validated $F$-test, however, they are different for the partitions, as well as the training and test sets.

Remark 6 Choosing $\rho_2 = 0.5$ may be a good substitute when $L(A(D), y)$ does not depend much on the training set $D$ and underlying algorithm $A$, that is when the decision function of the underlying algorithm does not change too much when different training sets are chosen. For instance, the support vector machine classifier (linear kernel) may be robust relative to perturbations in the training set (see [6, 17]).

In this study, we refer to the test following the distribution $F(7, 5)$ as calibrated balanced $5 \times 2$ cross-validated $F$-test. In the following section, type I and type II errors of the $5 \times 2$ cross-validated $F$-test and the calibrated balanced $5 \times 2$ cross-validated $F$-test are compared through simulated experiments.

3 Simulated Experiments

As pointed out by [13], for the Alpaydin’s $5 \times 2$ cross-validated $F$-test, a poor partition may result in large type I and type II errors. In this section, we illustrate this and show that the calibrated balanced $5 \times 2$ cross-validated $F$-test has lower type I and type II errors than

[Springer]
the $5 \times 2$ cross-validated $F$-test with poor partition in most cases. Noting that it is hard to construct a $5 \times 2$ cross-validation with identical number of overlapped samples (except the number of overlapped samples is $\frac{n}{2}$), we then perform the experiments by controlling the maximum number of overlapped samples in five replications of 2-fold cross-validation. The detail is as follows.

First, the data set $D$ is split into two disjoint and equal-sized blocks, denoted as $P_j$, $j = 1, 2$ respectively. By exchanging $k$ elements of $P_1$ for $k$ elements of $P_2$, a new replication of 2-fold cross validation is obtained. The other 3 replications are obtained by random partition. Here, we choose the maximum number of overlapped samples be $\frac{kn}{20}$, i.e., $k = \frac{n}{20}$.

3.1 Simulated Data

Considering the problem of comparing the performances of two algorithms in a classification problem with two classes, we thus have $Z = (X, Y)$, with $\text{Prob}(Y = 1) = \text{Prob}(Y = 0) = \frac{1}{2}$, $X|Y = 0 \sim N(\mu_0, \Sigma_0)$, $X|Y = 1 \sim N(\mu_1, \Sigma_1)$. The classification algorithms are

(A) Regression Tree (RT)

We train a least square RT and the decision function is $F_A(ZS)(X) = I[N(ZS)(X) > 0.5]$, where $N(ZS)(X)$ is the leaf value corresponding to $X$ of the tree obtained when training on $ZS$. Thus, $L_A(j, i) = I[F_A(ZS)(X_i) \neq Y_i]$ is equal to 1 whenever this algorithm misclassifies example $i$; otherwise it is 0.

(B) Ordinary Least Squares Linear Regression (LS)

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

and

(C) Support Vector Machine (SVM)

We train a SVM classifier with a Gaussian kernel of $F_C(ZS)(X)$. Then we use the loss function $L_C(j, i) = I[F_C(ZS)(X_i) \neq Y_i]$ to examine the classification results.

(D) Random Forest (RF)

RF is a classifier containing multiple decision trees that the classification result is decided by voting to the classification results of multiple trees. The obtained classifier is denoted as $F_D(ZS)(X)$. The same $0 - 1$ loss function is used.

(E) Adaboost (AB)

AB can be used in conjunction with some weak learning algorithms to improve their performance. The output of these weak learning algorithms is combined into a weighted sum that represents the final output of the boosted classifier. Here, we use the tree algorithm. Also, we can train the boosted classifier of $F_E(ZS)(X)$ and obtain the classification results based on $0 - 1$ loss function.

Similar to [6], we take $\mu_0 = (0, 0)$, $\Sigma_0 = I_2$, but take multiple $\mu_1$ and $\Sigma_1$, let $n = 200$. We then test whether the two algorithms are different. Table 3 shows the probabilities of rejecting the null hypothesis of the $5 \times 2$ cross-validated $F$-test, the blocked $3 \times 2$ cross-validated $t$-test and the calibrated balanced $5 \times 2$ cross-validated $F$-test in 5000 replicated experiments with the setups shown in Table 3. The replicated experiments are drawn by independent sampling with replacement.

In eight simulated experiments presented in Table 3, the $5 \times 2$ cross-validated $F$-test with poor partition all exhibit high probabilities of rejecting the null hypothesis. For example, in case (3) with the comparisons of RT and LS classifiers, the probability of rejecting the null hypothesis of the $5 \times 2$ cross-validated $F$-test is 0.070, which is higher than 0.05, however, it
Table 3  Experimental results of the cases of different means and variances in simulated data

<table>
<thead>
<tr>
<th>Case</th>
<th>Case (1)</th>
<th>Case (2)</th>
<th>Case (3)</th>
<th>Case (4)</th>
<th>Case (5)</th>
<th>Case (6)</th>
<th>Case (7)</th>
<th>Case (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_0 )</td>
<td>((0, 0))</td>
<td>((0, 0))</td>
<td>((0, 0))</td>
<td>((0, 0))</td>
<td>((0, 0))</td>
<td>((0, 0))</td>
<td>((0, 0))</td>
<td>((0, 0))</td>
</tr>
<tr>
<td>( \mu_1 )</td>
<td>((-\frac{3}{2}, -\frac{1}{2}))</td>
<td>((-\frac{1}{2}, -\frac{1}{2}))</td>
<td>((1, 1))</td>
<td>((1, 1))</td>
<td>((1, 1))</td>
<td>((2, 2))</td>
<td>((2, 2))</td>
<td></td>
</tr>
<tr>
<td>( \Sigma_0 )</td>
<td>(I_2)</td>
<td>(I_2)</td>
<td>(I_2)</td>
<td>(I_2)</td>
<td>(I_2)</td>
<td>(I_2)</td>
<td>(I_2)</td>
<td></td>
</tr>
<tr>
<td>( \Sigma_1 )</td>
<td>(\frac{1}{2}I_2)</td>
<td>(\frac{1}{6}I_2)</td>
<td>(\frac{1}{3}I_2)</td>
<td>(\frac{1}{2}I_2)</td>
<td>(\frac{1}{4}I_2)</td>
<td>(\frac{1}{3}I_2)</td>
<td>(\frac{1}{4}I_2)</td>
<td></td>
</tr>
</tbody>
</table>

where \(F_{5 \times 2CV}\), \(t_{B3 \times 2CV}\) and \(F_{B5 \times 2CV}\) refer to the 5 \(\times\) 2 cross-validated \(F\)-test following \(F(10, 5)\), the blocked 3 \(\times\) 2 cross-validated \(t\)-test following \(t(5)\) and the calibrated balanced 5 \(\times\) 2 cross-validated \(F\)-test following \(F(7, 5)\), respectively.

3.2 Real Data

In this subsection, we further carry out experiments on six data sets from the UCI repository (iris, wine, glass, heart, balance, thyroid gland). The specifications of these data sets are listed as follows.

- **Iris data set** the iris data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. The number of attributes are 4. The sample size is 150.
- **Wine data set** this data set is used to the recognition of three types of wines. It includes 178 samples with class 1 of 59, class 2 of 71 and class 3 of 48, and 13 continuous attributes.
Choosing Between Two Classification Learning Algorithms...

Table 4  Probabilities of type I and type II errors based on different numbers of hidden units of MLP on six data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number of hidden units</th>
<th>Probabilities of type I error</th>
<th>Probabilities of type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$F_{5\times2CV}$ $t_{B3\times2CV}$ $F_{B5\times2CV}$</td>
<td>$F_{5\times2CV}$ $t_{B3\times2CV}$ $F_{B5\times2CV}$</td>
</tr>
<tr>
<td>Iris</td>
<td>3</td>
<td>0.023 0.001 0.018</td>
<td>0.038 0.001 0.042</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.006 0.008 0.008</td>
<td>0.007 0.009 0.009</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.008 0.009 0.007</td>
<td>0.009 0.009 0.008</td>
</tr>
<tr>
<td>Wine</td>
<td>10</td>
<td>0.034 0.078 0.030</td>
<td>0.909 0.938 0.892</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.019 0.020 0.019</td>
<td>0.687 0.701 0.664</td>
</tr>
<tr>
<td>Glass</td>
<td>5</td>
<td>0.018 0.005 0.018</td>
<td>0.635 0.599 0.554</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.022 0.004 0.017</td>
<td>0.214 0.210 0.196</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.019 0.003 0.019</td>
<td>0.059 0.070 0.064</td>
</tr>
<tr>
<td>Heart</td>
<td>5</td>
<td>0.015 0.004 0.015</td>
<td>0.020 0.003 0.022</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.015 0.002 0.018</td>
<td>0.013 0.007 0.013</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.015 0.004 0.013</td>
<td>0.023 0.038 0.022</td>
</tr>
<tr>
<td>Balance</td>
<td>5</td>
<td>0.018 0.004 0.013</td>
<td>0.058 0.039 0.036</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.015 0.006 0.018</td>
<td>0.255 0.251 0.157</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.016 0.007 0.013</td>
<td>0.522 0.545 0.331</td>
</tr>
<tr>
<td>Thyroid gland</td>
<td>10</td>
<td>0.018 0.008 0.016</td>
<td>0.081 0.027 0.074</td>
</tr>
</tbody>
</table>

- **Glass data set** this data set contains 214 samples. 6 types of glass are identified based on 9 attributed variables. The numbers of sample for each class are 70, 17, 76, 13, 9, and 29, respectively.
- **Heart data set** this data set is a database concerning heart disease diagnosis. It contains 270 samples and 13 variables. The classification results are the heart “present” or “absent”.
- **Balance data set:** this data set was generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The sample size and the number of variable are 625 and 4, respectively. The class distributions for three class are 46.08%, 46.08%, and 7.84%, respectively.
- **Thyroid gland data set** the thyroid gland data for identifying thyroid gland diseases (normal, hypo or hyper) comprise 215 examples described by 5 attributes. The numbers of samples for each class are 150, 35 and 30, respectively.

First, similar to Alpaydin’s [1] work, to compare type I error of the two tests, we use two multilayer perceptrons (MLP with one hidden layer) with equal numbers of hidden units. Thus, the null hypothesis is true, and any reject is a type I error. To compare type II error of the two tests, we take two classifiers that are different: an LP (single-layer perceptron) and an MLP. A MLP is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron with a nonlinear activation function. The LP is the simplest neural network. It only contains input and output layers. Tables 4 shows the results of the type I and type II error rates based on 5000 replicated experiments. This is achieved by considering the data set to be the population, from which training and test samples are drawn by independent partitioning to the data set.
Table 5  Probabilities of rejecting the null hypothesis on real data and SVM, RF and LDA classifiers

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Iris</th>
<th>Wine</th>
<th>Glass</th>
<th>Heart</th>
<th>Balance</th>
<th>Thyroid gland</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM versus RF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(F_{5\times2CV})</td>
<td>0.039</td>
<td>0.040</td>
<td>0.988</td>
<td>0.029</td>
<td>0.079</td>
<td>0.968</td>
</tr>
<tr>
<td>(t_{B3\times2CV})</td>
<td>0.102</td>
<td>0.036</td>
<td>0.999</td>
<td>0.025</td>
<td>0.060</td>
<td>0.986</td>
</tr>
<tr>
<td>(F_{B5\times2CV})</td>
<td>0.032</td>
<td>0.035</td>
<td>0.987</td>
<td>0.021</td>
<td>0.059</td>
<td>0.958</td>
</tr>
<tr>
<td>SVM versus LDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(F_{5\times2CV})</td>
<td>0.242</td>
<td>0.034</td>
<td>0.275</td>
<td>0.021</td>
<td>0.144</td>
<td>0.143</td>
</tr>
<tr>
<td>(t_{B3\times2CV})</td>
<td>0.491</td>
<td>0.041</td>
<td>0.410</td>
<td>0.004</td>
<td>0.187</td>
<td>0.112</td>
</tr>
<tr>
<td>(F_{B5\times2CV})</td>
<td>0.224</td>
<td>0.043</td>
<td>0.293</td>
<td>0.013</td>
<td>0.123</td>
<td>0.124</td>
</tr>
<tr>
<td>LDA versus RF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(F_{5\times2CV})</td>
<td>0.124</td>
<td>0.008</td>
<td>0.935</td>
<td>0.027</td>
<td>0.378</td>
<td>0.972</td>
</tr>
<tr>
<td>(t_{B3\times2CV})</td>
<td>0.057</td>
<td>0.001</td>
<td>0.850</td>
<td>0.001</td>
<td>0.114</td>
<td>0.957</td>
</tr>
<tr>
<td>(F_{B5\times2CV})</td>
<td>0.108</td>
<td>0.009</td>
<td>0.927</td>
<td>0.025</td>
<td>0.347</td>
<td>0.954</td>
</tr>
</tbody>
</table>

In 23 of 30 cases, the calibrated balanced \(5 \times 2\) cross-validated \(F\)-test has lower type I and type II errors than the \(5 \times 2\) cross-validated \(F\)-test (Table 4). One example is the situation in which the numbers of hidden units is 5 and the data set is balance. In this case, the probabilities of type I and type II errors of the \(5 \times 2\) cross-validated \(F\)-test are 0.018 and 0.058 respectively, whereas those of the calibrated balanced \(5 \times 2\) cross-validated \(F\)-test are 0.013 and 0.036, respectively. For the blocked \(3 \times 2\) cross-validated \(t\)-test, it exhibits good performance in the probabilities of type I error. It has lower type I errors than the calibrated balanced \(5 \times 2\) cross-validated \(F\)-test in 10 of 15 cases. However, for the probabilities of type II error, only 4 out of 15 cases has a superior performance over our proposed \(F\)-test.

Furthermore, we also compare these three tests based on the SVM, RF, and linear discriminant analysis (LDA) classifiers on six real data sets. As shown in Table 5, in 15 of 18 cases, our method has lower probabilities of rejecting the null hypothesis than that of the \(5 \times 2\) cross-validated \(F\)-test. In the comparisons of SVM versus RF and SVM versus LDA, our method is superior to the blocked \(3 \times 2\) cross-validated \(t\)-test in 9 of 12 cases. However, in the comparison of LDA versus RF, the blocked \(3 \times 2\) cross-validated \(t\)-test exhibits better performances than the \(5 \times 2\) and calibrated balanced \(5 \times 2\) cross-validated \(F\)-tests in almost all cases (5 out of 6 cases). This implies that the conservative variance estimation in the blocked \(3 \times 2\) cross-validated \(t\)-test results in good performance in some cases. This finding provides a research direction toward further improving the performance of calibrated balanced \(5 \times 2\) cross-validated \(F\)-test by improving its variance estimation.

4 Conclusion

Noting that \(5 \times 2\) cross-validation is the result of five replications of 2-fold cross-validation, training (test) sets from any two independent partitions contain common samples regardless of how the data is split. Thus, cross-validation estimators for different data partitions are actually not independent, that was previously neglected in \(5 \times 2\) cross-validated \(t\) and \(F\) tests. In this study, based on the blocked \(3 \times 2\) cross-validation considering the correlation between any two 2-fold cross-validations given by [2], we propose a balanced \(5 \times 2\) cross-validated \(F\)-test.
Then, we analyze and discuss the distribution of this test statistic under the assumptions of same group-in correlations and same group-out correlations. The assumptions are reasonable for our balanced 5 × 2 cross-validation from the sample balance (the identical number of overlapped samples in the five groups). However, it may not be reasonable for the (random) 5 × 2 cross-validation. We find that the test statistic remains unchanged with increasing group-out correlation coefficient. However, the DOF of the distribution $f$ decreases more rapidly. They all decrease with increasing group-in correlation coefficient. This decrease is slower than the change in the DOF $f$ with varying group-out correlation coefficient. Thus, we calibrate the DOF of the $F(10, 5)$ distribution. We believe that following $F(7, 5)$ is more reasonable. Simulated and real data studies also demonstrate that the calibrated balanced $F$-test has lower type I and type II errors than the 5 × 2 cross-validated $F$-test following $F(10, 5)$ in most cases.

Nonetheless, the calibrated test statistic cannot ensure the independence between the numerator and the denominator of the test statistic. Further study is being conducted.

**Acknowledgements** This work was supported by National Natural and Social Science Funds of China (61503228, 16BTJ304), Natural Science Fund of Shanxi Province (201601D011046) and Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase).

**Appendix**

**Proof of proposition** Denoting $U = (\hat{\mu}_{B_1}^{(1)}, \hat{\mu}_{B_2}^{(1)}, \hat{\mu}_{B_1}^{(2)}, \hat{\mu}_{B_2}^{(2)}, \ldots, \hat{\mu}_{B_1}^{(5)}, \hat{\mu}_{B_2}^{(5)})^T$, we have $U \sim N(0, \sigma^2 \Sigma)$ from the assumption of Proposition, where

$$
\Sigma = \begin{pmatrix}
1 & \rho_1 & \rho_2 & \cdots & \rho_2 & \rho_2 \\
\rho_1 & 1 & \rho_2 & \cdots & \rho_2 & \rho_2 \\
\vdots & \vdots & \ddots & \cdots & \vdots & \vdots \\
\rho_2 & \rho_2 & \rho_2 & \cdots & 1 & \rho_1 \\
\rho_2 & \rho_2 & \rho_2 & \cdots & \rho_1 & 1
\end{pmatrix}_{10 \times 10}
$$

The eigenvalues of $\Sigma$ are obtained easily from $|\lambda I - \Sigma| = 0$: $\lambda_1 = 1 - \rho_1$ with multiplicity 5, $\lambda_6 = -2\rho_2 + \rho_1 + 1$ with multiplicity 4, and $\lambda_{10} = \rho_1 + 8\rho_2 + 1$. Thus, we can conclude that real symmetric matrix $\Sigma$ represents a positive definite matrix when $0 \leq \rho_1 \leq \rho_2 < 0.50$. We can also conclude that $\Sigma^{\frac{1}{2}}$ is a positive definite matrix from the decomposition $\Sigma = \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}}$.

Let $U_1 = U/\sigma$, $Z = \Sigma^{-\frac{1}{2}} U_1$, then we have $U_1 \sim N(0, \Sigma)$, $Z \sim N(0, I_{10})$ and obviously $U_1^T U_1 = Z^T \Sigma Z$.

An orthogonal matrix exists for each $n$ order real symmetric matrix such that the matrix can be diagonalized. Thus, an orthogonal matrix $T$ exists such that $T \Sigma T^T = \Lambda$, i.e., $\Sigma = T^T \Lambda T$, where $\Lambda$ is a diagonal matrix, and its element is the eigenvalue of $\Sigma$.

We know that $TZ \sim N(0, I_{10})$ from the properties of the orthogonal matrix, then

$$
U_1^T U_1 = Z^T \Sigma Z = Z^T T^T \Lambda T Z = \sum_{i=1}^{10} \lambda_i \eta_i^2.
$$

Thus, $U_1^T U_1$ approximately follows an $C \chi^2(f)$ distribution because $\sum_{i=1}^{10} \lambda_i \eta_i^2$ approximately follows $C \chi^2(f)$ distribution, where $\lambda_i$ denotes the eigenvalue of $\Sigma$, $\eta_i$ is the $i$-th element of matrix $TZ$, and

$$
C = \frac{\sum_{i=1}^{10} \lambda_i^2}{\sum_{i=1}^{10} \lambda_i} = 1 + \rho_1^2 + 8\rho_2^2, \quad f = \frac{(\sum_{i=1}^{10} \lambda_i)^2}{\sum_{i=1}^{10} \lambda_i^2} = \frac{10}{1 + \rho_1^2 + 8\rho_2^2}
$$

(see [18,19]).
Note $\sum_{i=1}^{10} \lambda_i = 10$, $\sum_{i=1}^{10} \lambda_i^2 = 10(1 + \rho_1^2 + 8\rho_2^2)$, $fC = \sum_{i=1}^{10} \lambda_i = 10$.

Moreover, we have $\text{Var}(\hat{\mu}_B^{(i)} - \bar{\mu}_B^{(i)}) = 2(1 - \rho_1)\sigma^2$, $\text{Cov}(\hat{\mu}_B^{(i)} - \bar{\mu}_B^{(i)}, \bar{\mu}_B^{(i)} - \bar{\mu}_B^{(i)})$

$= (\rho_2 - \rho_2 - \rho_2 + \rho_2)\sigma^2 = 0$ for $i \neq i'$. If letting $U_2 = ((\hat{\mu}_B^{(1)} - \bar{\mu}_B^{(1)})/\sigma, \ldots, (\hat{\mu}_B^{(5)} - \bar{\mu}_B^{(5)})/\sigma)$, then $U_2 \sim N(0, \Sigma_2)$, and $U_2^T U_2 \sim 2(1 - \rho_1)\chi^2(5)$, where

$$\Sigma_2 = \begin{pmatrix}
2(1 - \rho_1) & 0 & \cdots & 0 \\
0 & 2(1 - \rho_1) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 2(1 - \rho_1)
\end{pmatrix}_{5 \times 5}$$

These result in

$$\frac{\sum_{i=1}^{5} \sum_{k=1}^{2} (\hat{\mu}_B^{(i)} - \bar{\mu}_B^{(i)})^2 / fC}{\sum_{i=1}^{5} (\hat{\mu}_B^{(i)} - \bar{\mu}_B^{(i)})^2 / (2(1 - \rho_1))} = \frac{U_2^T U_2}{fC} \frac{10(1 - \rho_1)}{(1 - \rho_1) U_1^T U_1} \sim F(f, 5).$$

We know that

$$S_{B_k}^2 = \sum_{k=1}^{2} (\hat{\mu}_B^{(i)} - \bar{\mu}_B^{(i)})^2 = (\hat{\mu}_B^{(i)} - (\frac{\hat{\mu}_B^{(i)} + \bar{\mu}_B^{(i)}}{2}))^2 + (\hat{\mu}_B^{(i)} - (\frac{\hat{\mu}_B^{(i)} + \bar{\mu}_B^{(i)}}{2}))^2

= \frac{(\hat{\mu}_B^{(i)} - \bar{\mu}_B^{(i)})^2}{2},$$

where $\hat{\mu}_B^{(i)} = \frac{\hat{\mu}_B^{(i)} + \hat{\mu}_B^{(i)}}{2}$.

Therefore,

$$F = \frac{(1 - \rho_1) \sum_{i=1}^{5} \sum_{k=1}^{2} (\hat{\mu}_B^{(i)} - \bar{\mu}_B^{(i)})^2}{\sum_{i=1}^{5} (\hat{\mu}_B^{(i)} - \bar{\mu}_B^{(i)})^2} = \frac{1 - \rho_1}{2} \sum_{i=1}^{5} \sum_{k=1}^{2} S_{B_k}^2 \sim F(f, 5),$$

where $f = \frac{10}{1 + \rho_1^2 + 8\rho_2^2}$. \hfill \Box

References

Choosing Between Two Classification Learning Algorithms...