Pattern Recognition Letters 65 (2015) 211–217

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Measure for data partitioning in $m \times 2$ cross-validation☆

Yu Wang[a], Jihong Li[a,*], Yanfang Li[b]

[a] Computer Center of Shanxi University, Taiyuan 030006, PR China
[b] School of Mathematical Sciences, Shanxi University, Taiyuan 030006, PR China

ARTICLE INFO

Article history:
Received 19 December 2014
Available online 14 August 2015

Keywords:
Data partitioning
measure
cross-validation
small probability event

ABSTRACT

An $m \times 2$ cross-validation based on m half–half partitions is widely used in machine learning. However, the cross-validation performance often relies on the quality of the data partitioning. Poor data partitioning may cause poor inference results, such as a large variance and large Type I and II errors of the corresponding test. To evaluate the quality of the data partitioning, we propose a statistic based on the difference between the observed and expected numbers of overlapped samples of two training sets in an $m \times 2$ cross-validation. The expectation and variance of the proposed statistic are also given. Furthermore, by studying the quantile of the distribution of the statistic, we find that the occurrence of poor data partitioning is not a small probability event. Thus, data partitioning should be predesigned before conducting $m \times 2$ cross-validation experiments in machine learning such that the number of overlapped samples observed is equal or as close as possible to the number expected.

1. Introduction

In machine learning research, a cross-validation method is commonly used in model assessment and selection, as well as in the estimation of generalization error. To date, many versions of cross-validation have been developed, including Repeated Learning-Testing(RLT), standard K-fold cross-validation, Monte-carlo cross-validation, 5×2 cross-validation and blocked 3×2 cross-validation (Hastie et al. [7]; Nadeau and Bengio [9]; Bengio and Grandvalet [3]; Markatou et al. [8]; Arlot and Celisse [2]; Dietterich [4]; Alpaydin [1]; Yildiz [12]; Wang et al. [10]).

In particular, standard 2-fold cross-validation has received considerable attention because it exhibits some good properties, such as a low computational cost, consistency in selection used to model selection for the classification situation of Yang [11], and use in variance estimation in the ultra-high dimensional linear regression model of Fan [5]. A 2-fold cross-validation splits the data into two equal-sized data sets, i.e., training and test sets, and each 2-fold cross-validation corresponds to one half–half partition. In practice, to be able to eliminate the effect by chance (e.g., variance due to changes in the training set), typically, one does 2-fold cross-validation a number of times. A cross-validation based on m replications of 2-fold cross-validation is called an $m \times 2$ cross-validation.

An $m \times 2$ cross-validation based on m data partitions is widely used in machine learning. Dietterich [4], Alpaydin [1] and Yildiz [12] demonstrated the superiority of a 5×2 cross-validation when comparing algorithms for $m = 5$. Hafidi and Mkhadri [6] provided a large sample property of an $m \times 2$ cross-validation (called Repeated Half Sampling in their paper) in the model selection of linear regression. Unfortunately, the performance of cross-validation often relies on the quality of the data partitioning. However, the data partitioning cannot be directly evaluated. It should be noted that different partitions generate different training and test sets in an $m \times 2$ cross-validation, and that training sets (test sets) from any two independent partitions contain common samples regardless of how the data are split. The number of common samples in training sets (test sets) from two data partitions is defined as the number of overlapped samples. In fact, Markatou et al. [8] theoretically proved that the number of overlapped samples from any two training sets follows a hypergeometric distribution, and that the mathematical expectation is $n/4$ (where n is the sample size). The following two examples illustrate the impact of the number of overlapped samples on the performance of an $m \times 2$ cross-validation.

Example 1. At the sample size $n = 40$, we predict the response variable Y with $Y \in \{0, 1\}$ based on the predictive variables X_1, X_2, X_3. 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 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 classification tree. Here, we will examine the impact of the

☆ This paper has been recommended for acceptance by Egon L. van den Broek.

* Corresponding author. Tel.: +86- 351 -701 -1017; fax: +86- 351 -701 -1017.
E-mail addresses: wangyu@sxu.edu.cn (Y. Wang), lijh@sxu.edu.cn (J. Li), liyanfang@sxu.edu.cn (Y. Li).

http://dx.doi.org/10.1016/j.patrec.2015.08.002
0167-8655/

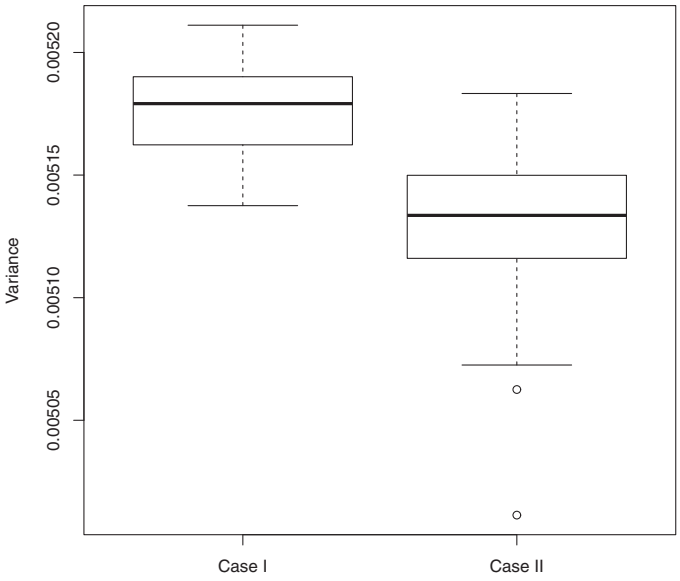

Fig. 1. Box plots for Example 1, where Case I and Case II refer to the cases of the difference of maximum number of overlapped samples and $n/4$ being 2 and 0, respectively.

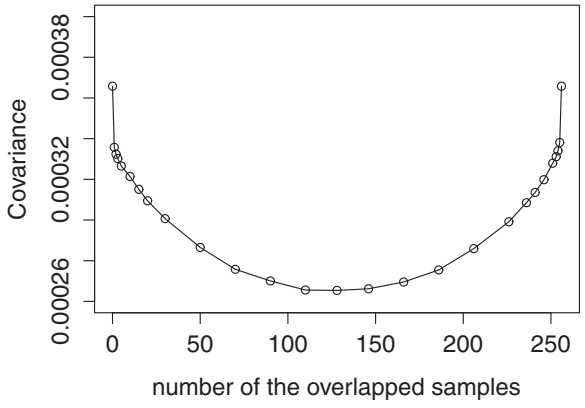

Fig. 2. Curve of covariance vs. number of overlapped samples.

number of overlapped samples on the variance of the generalization error estimation in an $m \times 2$ cross-validation at $m = 3$.

Fig. 1 shows the results. The variances of a 3×2 cross-validation with the number of overlapped samples equal to $n/4$ (Case II) in 100 replications are all smaller than that of a 3×2 cross-validation where the difference of the maximum number of overlapped samples and $n/4$ is 2 (Case I). This implies that a large variance may be caused by data partitioning when the number of overlapped samples is not equal to $n/4$. Furthermore, Wang et al. [10] had used a simulation to show that a 3×2 cross-validation with the expected number of overlapped samples of $n/4$ had a minimum variance. See the example shown in Fig. 2 of the change of covariance with the number of overlapped samples in any two replications of 2-fold cross-validation for $n = 512$ and support vector machine classifier.

Example 2 shows the impact of the number of overlapped samples on Type I and II errors of the corresponding test based on a 5×2 cross-validation and a similar conclusion is obtained.

Note: It is hard to construct an $m \times 2$ cross-validation with the same number of overlapped samples (except when the number of overlapped samples is $n/4$). So, we perform the experiments by controlling the maximum number of overlapped samples in m replications of a 2-fold cross-validation.

Table 1
Probabilities of Type I and II errors for Example 2, where 90 and 59 refer to the maximum numbers of overlapped samples in a 5×2 cross-validation.

n	200	200
μ_0	(0,0)	(0,0)
μ_1	(1,1)	(1,1)
Σ_0	I_2	I_2
Σ_1	$\frac{1}{6} I_2$	$\frac{1}{2} I_2$
	Probability of Type I error	Probability of Type II error
$F_{5\times 2CV}(90)$	0.070	0.086
$F_{5\times 2CV}(59)$	0.055	0.061
$F_{5\times 2CV}(\frac{n}{4} = 50)$	0.037	0.050

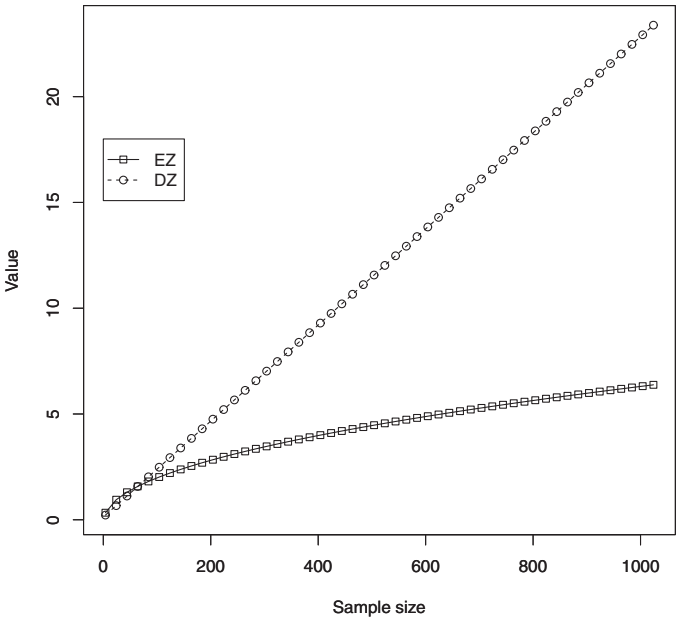

Fig. 3. The change in *EZ* and *DZ* for n from 4 to 1024 .

Example 2. By considering the problem of comparing two algorithms: regression tree and ordinary least squares linear regression, we thus have (X, Y), with $Prob(Y = 1) = Prob(Y = 0) = \frac{1}{2}, X|Y = 0 \sim N(\mu_0, \Sigma_0), X|Y = 1 \sim N(\mu_1, \Sigma_1)$. X is a binary input variable. Y is a output variable. The sample size is $n = 200$. Here we examine the impact of the maximum number of overlapped samples on the performance of a 5×2 cross-validated F-test given by Alpaydin [1].

The results of Table 1 show that the probabilities of Type I and II errors of a 5×2 cross-validated F-test gradually increase correspond to an increase in the maximum number of overlapped samples. The probabilities of Type I and II errors of a 5×2 cross-validated F-test, where the maximum number of overlapped samples are 59 and 90 respectively, are higher than the significance level of 0.05. However, from the conclusion derived by Nadeau and Bengio [9], we can see that these two classification learning algorithms have no statistical significant differences with a setup of $n = 200, \mu_0 = (0, 0), \mu_1 = (1, 1), \Sigma_0 = I_2, \Sigma_1 = \frac{1}{6} I_2$, but do have significant differences with a setup of $n = 200, \mu_0 = (0, 0), \mu_1 = (1, 1), \Sigma_0 = I_2, \Sigma_1 = \frac{1}{2} I_2$. For a 5×2 cross-validation with the number of overlapped samples of $n/4 = 50$, the probabilities are only 0.037 and 0.050. These findings all indicate that the poor data partitioning that occurs when the number of overlapped samples is not equal to $n/4$ results in a poor performance of an $m \times 2$ cross-validation.

However, how can the quality of the data partitioning be measured? Does such poor partitioning always occur when the data is randomly split? Thus, providing a measure for data partitioning in an $m \times 2$ cross-validation is important. In addition, such a measure should be constructed based on the number of overlapped samples

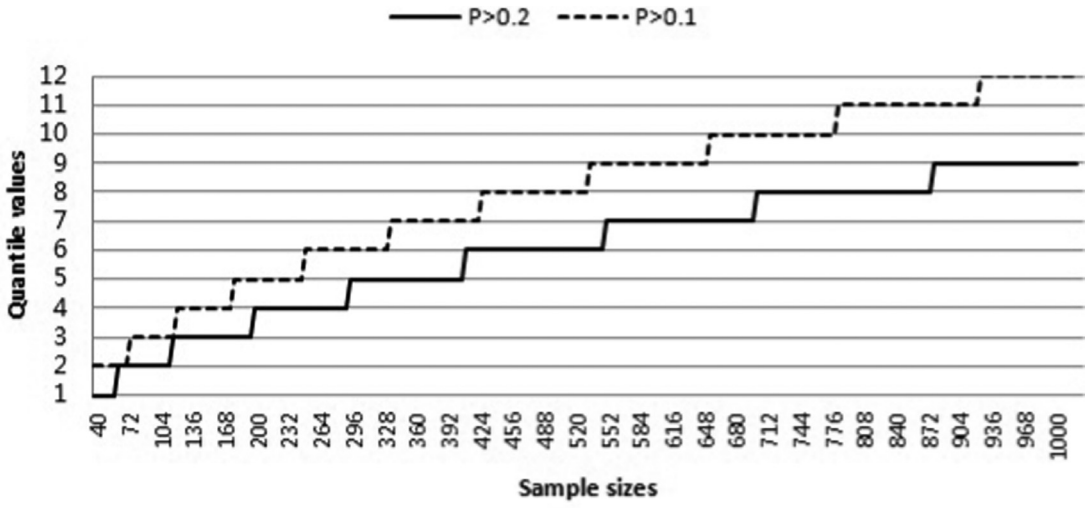

Fig. 4. Quantile values for different sample sizes.

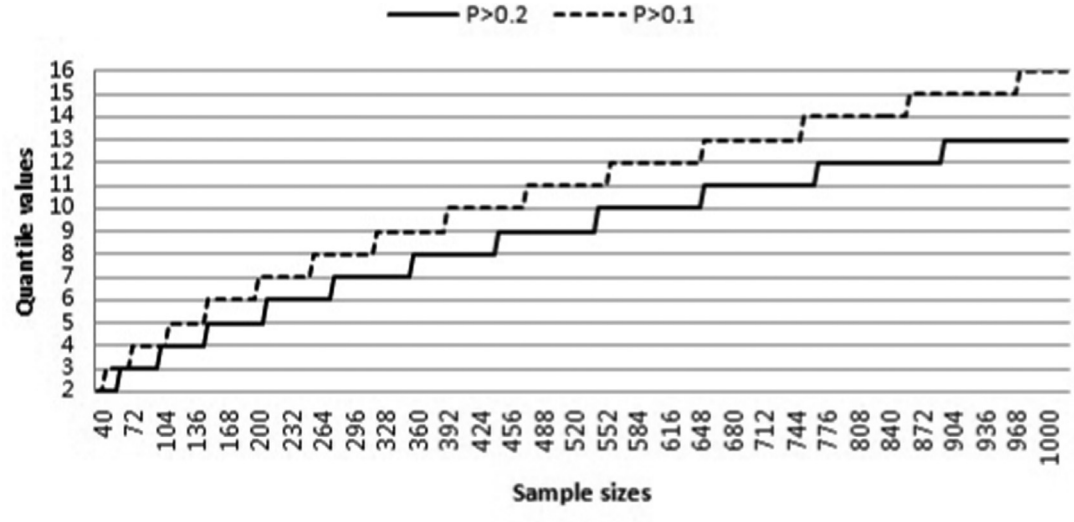

Fig. 5. Quantile values for different sample sizes and an 3 × 2 cross validation.

of the training sets in an $m \times 2$ cross-validation. As such, we propose a statistic based on the difference between the number of overlapped samples in the training sets resulting from any two data partitions of an $m \times 2$ cross-validation and its mathematical expectation. We also show that the occurrence of poor data partitioning is not a small probability event.

The remainder of this study is organized as follows. Section 2 describes the statistic used for measuring data partitioning and its statistical properties. Section 3 demonstrates that the occurrence of poor data partition is not a small probability event. Section 4 concludes the study.

2. Statistic used for measuring data partitioning

2.1. Statistic and its distribution

According to Markatou et al. [8], in an $m \times 2$ cross-validation the number of overlapped samples from any two training sets follows a hypergeometric distribution . This random variable is denoted as X. Thus, a statistic used for measuring the data partitioning based on X and its mathematical expectation can be defined as:

$$Z = |X - EX|, \tag{1}$$

where, $EX = n'$, $n = 4n'$ is sample size.

Then, we consider the distribution of Z. First, from X following hypergeometric distribution $h(2n', n, 2n')$, we have

$$P(X = k) = \frac{\binom{2n'}{k}\binom{2n'}{2n'-k}}{\binom{n}{2n'}}, k = 0, 1, 2, \ldots, 2n'. \tag{2}$$

Then, $P(X = k) = P(X = 2n' - k)$ is obtained from the symmetry properties of the distribution of X. Thus, in the case of $z = 0$

$$\begin{aligned} P(Z = 0) &= P(|X - EX| = 0) = P(|X - 2n'| = 0) \\ &= P(X = n') = \frac{\binom{2n'}{n'}^2}{\binom{n}{2n'}}. \end{aligned} \tag{3}$$

In the case of $z \neq 0, z = 1, 2, \ldots, n'$,

$$P(Z = z) = P(|X - n'| = z) = 2\frac{\binom{2n'}{n'-z}\binom{2n'}{n'+z}}{\binom{n}{n'}} = 2\frac{\binom{2n'}{n'-z}^2}{\binom{n}{2n'}}. \tag{4}$$

Remark 1. From the combination equation $\binom{2n'}{n'-z} = \binom{2n'}{n'-z-1} + \binom{2n'-1}{n'-z-1}$, we have $\binom{2n'}{n'-z} > \binom{2n'}{n'-z-1}$, i.e., $P(Z = z)$ is a decreasing function of z for $z > 0$.

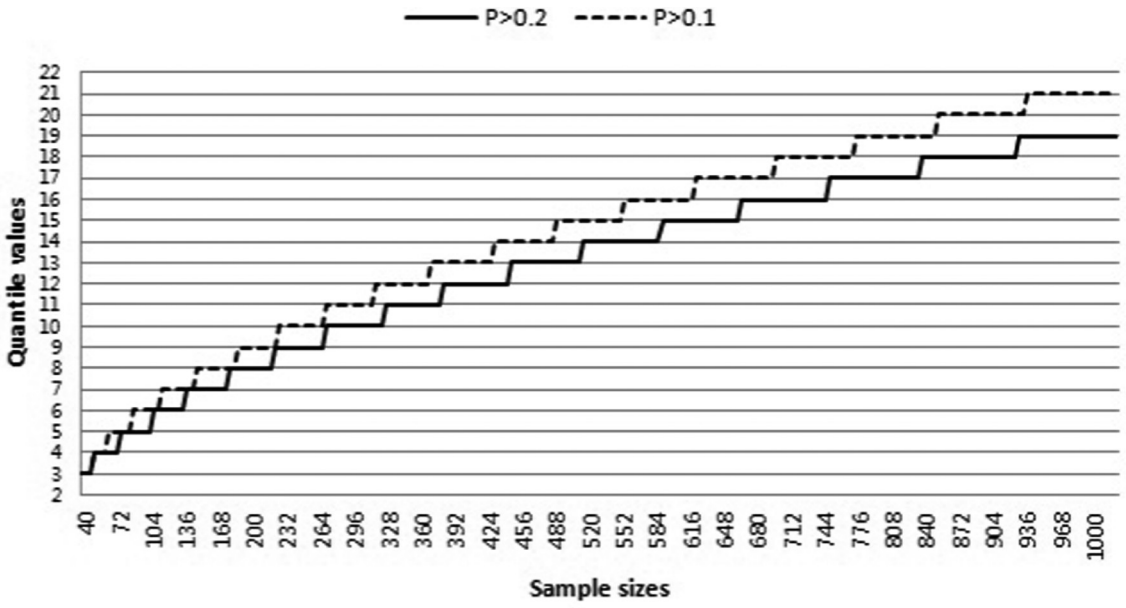

Fig. 6. Quantile values for different sample sizes and an 7×2 cross validation.

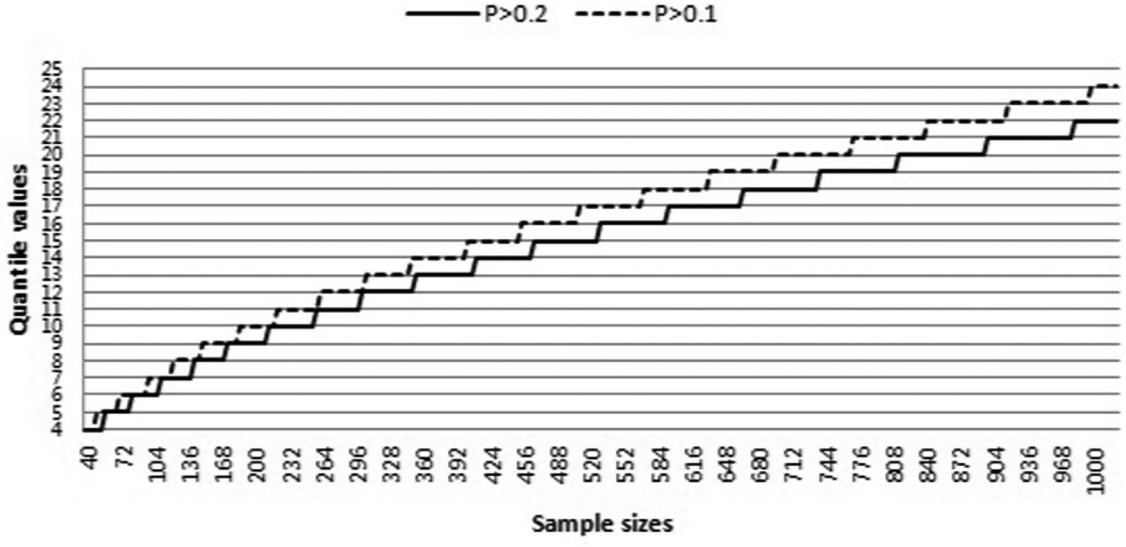

Fig. 7. Quantile values for different sample sizes and an 11×2 cross validation.

Table 2
The numbers of $Z_{max} = 0, 1, \ldots, 14$ in 100 replications of 3×2, 7×2, and 11×2 cross-validations for $n = 200$.

	Z_{max}=0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3×2 cross-validation	9	16	15	13	19	12	10	4	0	2	0	0	0	0	0
7×2 cross-validation	0	0	0	5	5	19	27	19	13	8	3	1	0	0	0
11×2 cross-validation	0	0	0	0	0	0	10	26	31	19	7	3	3	0	1

Table 3
The numbers of $Z_{max} = 0, 1, \ldots, 25$ in 100 replications of 3×2, 7×2, and 11×2 cross-validations for $n = 600$.

	Z_{max}=0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3×2 cross-validation	4	9	14	13	5	10	8	9	10	5	3	0	2	2	2
7×2 cross-validation	0	0	0	0	0	0	2	4	11	14	6	13	12	13	7
11×2 cross-validation	0	0	0	0	0	0	0	0	0	1	5	8	12	10	18
	Z_{max}=15	16	17	18	19	20	21	22	23	24	25				
3×2 cross-validation	2	1	0	0	0	1	0	0	0	0	0				
7×2 cross-validation	5	5	4	1	2	1	0	0	0	0	0				
11×2 cross-validation	13	14	8	5	3	2	0	0	0	0	1				

Table 4
The numbers of $Z_{max} = 0, 1, \ldots, 28$ in 100 replications of 3×2, 7×2, and 11×2 cross-validations for $n = 1000$.

	Z_{max}=0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3×2 cross-validation	1	5	8	9	7	9	12	8	6	5	4	3	6	4	2
7×2 cross-validation	0	0	0	0	0	0	0	1	1	7	10	9	6	12	7
11×2 cross-validation	0	0	0	0	0	0	0	0	0	0	0	0	0	10	5
	Z_{max}=15	16	17	18	19	20	21	22	23	24	25	26	27	28	
3×2 cross-validation	2	3	1	2	0	1	0	0	0	0	0	0	0	0	
7×2 cross-validation	10	6	4	9	7	1	1	3	2	0	1	1	0	2	
11×2 cross-validation	12	8	14	6	13	11	7	6	2	2	3	0	1	0	

2.2. Mathematical expectation and variance of the statistic

An mathematical expectation and variance are very important numerical characteristics of a statistic. A expectation reflects the extent of the average difference between the observed and expected numbers of overlapped samples. This characteristic is of great concern by researchers in terms of the statistic's practical application. In addition to a mathematical expectation, we are also concerned with the variance of the statistic because it reflects the fluctuating range of the statistic's values. To further understand the proposed statistic, we discuss its expectation and variance.

Proposition 1. *The expectation and variance of statistic Z can be expressed as Eqs. (5)*

$$EZ = \frac{n^2}{4(n-1)} \frac{\binom{2n'-1}{n'}\binom{2n'-1}{n'-1}}{\binom{n-2}{2n'-1}} + \frac{n(n-2)}{8(n-1)} \frac{\binom{2n'-1}{n'-1}^2}{\binom{n-2}{2n'-2}} - \frac{n}{4} \frac{\binom{2n'}{n'}^2}{\binom{n}{2n'}} \tag{5}$$

and (6)

$$DZ = \frac{n^2}{16(n-1)} - \left(\frac{n^2}{4(n-1)} \frac{\binom{2n'-1}{n'}\binom{2n'-1}{n'-1}}{\binom{n-2}{2n'-1}} + \frac{n(n-2)}{8(n-1)} \frac{\binom{2n'-1}{n'-1}^2}{\binom{n-2}{2n'-2}} - \frac{n}{4} \frac{\binom{2n'}{n'}^2}{\binom{n}{2n'}} \right)^2 \tag{6}$$

where $n' = n/4$.

Proof. First, from Eqs. (3) and (4), the expectation of Z can be written as

$$\begin{aligned} EZ &= \sum_{k=0}^{n'} |k-n'| P(X=k) + \sum_{k=n'+1}^{2n'} |k-n'| P(X=k) \\ &= n' - n' P(X=n') - 2\sum_{k=1}^{n'-1} k \frac{\binom{2n'}{k}\binom{2n'}{2n'-k}}{\binom{n}{2n'}}. \end{aligned}$$

And

$$\begin{aligned} \sum_{k=1}^{n'-1} k \frac{\binom{2n'}{k}\binom{2n'}{2n'-k}}{\binom{n}{2n'}} &= \frac{n}{4} \sum_{k=0}^{n'-2} \frac{\binom{2n'-1}{k}\binom{2n'}{2n'-k-1}}{\binom{n-1}{2n'-1}} \\ &= \frac{n^2}{8(n-1)} \left[\frac{1}{2} - \frac{\binom{2n'-1}{n'}\binom{2n'-1}{n'-1}}{\binom{n-2}{2n'-1}} \right] \\ &\quad + \frac{n(n-2)}{16(n-1)} \left[1 - \frac{\binom{2n'-1}{n'-1}\binom{2n'-1}{n'-1}}{\binom{n-2}{2n'-2}} \right]. \end{aligned}$$

This imply that

$$EZ = \frac{n^2}{4(n-1)} \frac{\binom{2n'-1}{n'}\binom{2n'-1}{n'-1}}{\binom{n-2}{2n'-1}} + \frac{n(n-2)}{8(n-1)} \frac{\binom{2n'-1}{n'-1}^2}{\binom{n-2}{2n'-2}} - \frac{n}{4} \frac{\binom{2n'}{n'}^2}{\binom{n}{2n'}}.$$

Second, the variance of Z has the following form:

$$\begin{aligned} DZ &= E(Z^2) - (EZ)^2 \\ &= \frac{n^2}{16(n-1)} \\ &\quad - \left(\frac{n^2}{4(n-1)} \frac{\binom{2n'-1}{n'}\binom{2n'-1}{n'-1}}{\binom{n-2}{2n'-1}} + \frac{n(n-2)}{8(n-1)} \frac{\binom{2n'-1}{n'-1}^2}{\binom{n-2}{2n'-2}} - \frac{n}{4} \frac{\binom{2n'}{n'}^2}{\binom{n}{2n'}} \right)^2, \end{aligned}$$

where

$$E(Z^2) = \sum_{k=0}^{2n'} (k-n')^2 P(X=k) = \frac{n(n-2)^2}{16(n-1)} - \frac{n^2}{16} + \frac{n}{4} = \frac{n^2}{16(n-1)}.$$

□

Table 5
Table corresponding to Graph 4, where $k(P > 0.2)$ and $k(P > 0.1)$ refer to the quantile values for the probabilities $P(Z > k) > 0.2$ and $P(Z > k) > 0.1$, respectively.

Sample size (n)	$k(P > 0.2)$	$k(P > 0.1)$	Sample size (n)	$k(P > 0.2)$	$k(P > 0.1)$
40–60	1	2	412–424	6	7
64–72	2	2	428–532	6	8
76–116	2	3	536–548	6	9
120	3	3	552–652	7	9
124–176	3	4	656–700	7	10
180–196	3	5	704–780	8	10
200–248	4	5	784–876	8	11
252–292	4	6	880–924	9	11
296–332	5	6	928–1024	9	12
336–408	5	7			

Table 6
Table corresponding to Graph 5.

Sample size (n)	$k(P > 0.2)$	$k(P > 0.1)$	Sample size (n)	$k(P > 0.2)$	$k(P > 0.1)$
40–44	2	2	360–392	8	9
48–60	2	3	396–444	8	10
64–72	3	3	448–472	9	10
76–100	3	4	476–544	9	11
104–108	4	4	548–556	10	11
112–148	4	5	560–652	10	12
152–200	5	6	656–752	11	13
204–208	5	7	756–768	11	14
212–256	6	7	772–860	12	14
260–276	6	8	864–896	12	15
280–320	7	8	900–972	13	15
324–356	7	9	976–1024	13	16

Remark 2. To clarify the change in the expectation and variance of statistic Z with varying sample sizes n, we examined their values for n from 4 to 1024. Fig. 3 indicates that the values of EZ and DZ all gradually increase with an increase in sample capacity and that the variance changes more quickly. This implies that the difference between the observed and expected numbers of overlapped samples becomes large and that the stability deteriorates with an increase in sample sizes n. This finding further shows the importance of measuring the quality of the data partitioning.

3. Quantile value of Z in which a small probability event does not occur

From Wang et al. [10] and Examples 1 and 2 described in Section 1, we can see that the performance of an $m \times 2$ cross-validation based on poor data partitioning with a large Z deteriorates. Next, we show that the occurrence of poor data partitioning is not a small

Table 7
Table corresponding to Graph 6.

Sample size (n)	$k(P > 0.2)$	$k(P > 0.1)$	Sample size (n)	$k(P > 0.2)$	$k(P > 0.1)$
40–48	3	3	384–428	12	13
52–60	4	4	432–444	12	14
64–72	4	5	448–488	13	14
76–84	5	5	492–512	13	15
88–104	5	6	516–552	14	15
108–112	6	6	556–588	14	16
116–136	6	7	592–620	15	16
140–144	7	7	624–664	15	17
148–176	7	8	668–696	16	17
180–184	8	8	700–748	16	18
188–220	8	9	752–772	17	18
224	9	9	776–836	17	19
228–268	9	10	840–852	18	19
272–316	10	11	856–928	18	20
320–324	10	12	932–936	19	20
328–368	11	12	940–1024	19	21
372–380	11	13			

Table 8
Table corresponding to Graph 7.

Sample size(n)	$k(P > 0.2)$	$k(P > 0.1)$	Sample size (n)	$k(P > 0.2)$	$k(P > 0.1)$
40	3	4	356–400	13	14
44–48	4	4	404–408	13	15
52–56	4	5	412–452	14	15
60–68	5	5	456–464	14	16
72–80	5	6	468–508	15	16
84–92	6	6	512–528	15	17
96–108	6	7	532–568	16	17
112–120	7	7	572–592	16	18
124–140	7	8	596–632	17	18
144–148	8	8	636–664	17	19
152–172	8	9	668–696	18	19
176–184	9	9	700–736	18	20
188–212	9	10	740–768	19	20
216–220	10	10	772–812	19	21
224-256	10	11	816–840	20	21
260	11	11	844–896	20	22
264–300	11	12	900–916	21	22
304	12	12	920–980	21	23
308–348	12	13	984–996	22	23
352	12	14	1000–1024	22	24

probability event. If this partitioning is a small probability event, then finding the partitioning that satisfies the condition in which the number of overlapping samples is equal to $\frac{n}{4}$ is insignificant because a small probability event is unlikely to happen in a single experiment. A small probability event refers to an event with a probability of less than 0.1 or 0.2. We thus consider the quantile values of $P(Z > k) > 0.1$ and $P(Z > k) > 0.2$ for different sample sizes. In view of the expectation of X following a hypergeometric distribution, which is one quarter of the sample size, we always assume that sample size n is a multiple of 4.

3.1. Case of two partitions

For two arbitrarily random partitions, we assume that $X \sim h(\frac{n}{2}, n, \frac{n}{2}), Z = |X - EX|$. Then, $P(Z > k) > 0.1 \Leftrightarrow F_Z(k) \leq 0.9$ and $P(Z > k) > 0.2 \Leftrightarrow F_Z(k) \leq 0.8$ are obtained from $P(Z > k) = 1 - P(Z \leq k) = 1 - F_Z(k)$. Based on Eqs. (3)–(6), Fig. 4 shows the quantile values for n from 40 to 1024.

The quantile values in which a small probability event does not occur are $1(P > 0.2)$ and $2(P > 0.1)$, i.e., $P(Z > 1) > 0.2$ and $P(Z > 2) > 0.1$ when sample size n lies between 40 and 60.

3.2. Case of multiple partitions

Let $X_1, X_2, \ldots, X_l$ be $l = \binom{m}{2}$ random variables for m partitions in an $m \times 2$ cross-validation. For $Z_i = |X_i - EX_i|$ and $Z_{max} = max_i(Z_i)$, $i = 1, 2, \ldots, l$, the quantile values of Z_{max} for n from 40 to 1024 and $m = 3, 7$, and 11 are listed in Figs. 5–7 based on the distribution of Z and its properties.

From Figs. 5–7, we can see that even with a sample size of 1000, the probabilities are larger than 0.2 when the differences between the observed and expected maximum numbers of overlapped samples are only 13, 19, and 22 for 3×2, 7×2 and 11×2 cross-validations, respectively. These probabilities occur easily in a large number of experiments when implementing $m \times 2$ cross-validation. From Examples 1 and 2 of Section 1, we see that poor data partitioning results in a poor performance of an $m \times 2$ cross-validation. All of these results indicate the importance of data partitioning in an $m \times 2$ cross-validation. The probability of poor data partitioning occurring is high (see the description in Tables 2–4). For example, Table 2 shows the numbers of $Z_{max} = 0, 1, \ldots, 14$ in 100 replications of 3×2, 7×2, and 11×2 cross-validations for $n = 200$. With an increase in the value of m, the difference between the observed and expected maximum numbers of overlapped samples gradually increases. For example, in 100 replications of an 11×2 cross-validation, the most frequently occurring numbers of Z_{max} are 7, 8, and 9, and the numbers of replications are 26, 31, and 19, respectively.

4. Conclusions

Considering the finding in this study that poor data partitioning may result in poor inference results, we propose a measure for data partitioning in $m \times 2$ cross-validation. By analyzing the distribution of the statistic, we find that the occurrence of poor data partitioning is not a small probability event. Thus, we should consider data partitioning before analyzing data in practical application. Essentially, this finding reflects the idea of statistical experimental design, which requires that the collected data should be designed beforehand. Thus, when implementing data partitioning, we should control the number of overlapped samples as closely as possible to its mathematical expectation or construct data partitioning with identical number of overlapped samples equal to mathematical expectation similar to that in Wang et al. [10].

In future research, we will attempt to theoretically study the exact function of covariance and the number of overlapped samples, and then provide an exact expression of these variables. Furthermore, providing the test statistic using this exact function serves to the comparisons of algorithms' performance.

Acknowledgements

This work was supported by National Natural Science Fund of China (NNSFC) (61503228). Experiments were supported by High Performance Computing System of Shanxi University.

Appendix A

To clarify the values in the Figs. 4–7, Tables 5–8 corresponding to Figures are also provided in this section.

References

[1] E. Alpaydin, Combined 5 × 2 cv F test for comparing supervised classification learning algorithms, Neural Comput. 11 (8) (1999) 1885–1892.
[2] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Stat. Surv. 4 (2010) 40–79.
[3] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of *K*-fold cross-validation, J. Mach. Learn. Res. 5 (2004) 1089–1105.
[4] T. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. 10 (7) (1998) 1895–1924.
[5] Fan Jianqing, Guo Shaojun, Hao Ning, Variance estimation using refitted cross-validation in ultrahigh dimensional regression, J. R. Stat. Soc.: Ser. B 74 (1) (2012) 37–65.
[6] B. Hafidi, A. Mkhadri, Repeated half sampling criterion for model selection, Indian J. Stat. 66 (3) (2004) 1–16.

[7] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, New York, 2001.
[8] M. Markatou, H. Tian, S. Biswas, G. Hripcsak, Analysis of variance of cross-validation estimators of the generalization error, J. Mach. Learn. Res. 6 (2005) 1127–1168.
[9] C. Nadeau, Y. Bengio, Inference for the generalization error, Mach. Learn. 52 (3) (2003) 239–281.
[10] Wang Yu, Wang Ruibo, Jia Huichen, Li Jihong, Blocked 3 × 2 cross-validated t-test for comparing supervised classification learning algorithms, Neural Comput. 26 (1) (2014) 208–235.
[11] Yang Yuhong, Comparing learning methods for classification, Stat. Sin. 16 (2006) 635–657.
[12] O.T. Yildiz, Omnivariate rule induction using a novel pairwise statistical test, IEEE Trans. Knowl Data Eng. 25 (2013) 2105–2118.