Pattern Recognition 69 (2017) 94–106

Contents lists available at ScienceDirect

Pattern Recognition

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

# Error estimation based on variance analysis of *k*-fold cross-validation

Gaoxia Jiang, Wenjian Wang*

*School of Computer and Information Technology, Shanxi University, Taiyuan 030006, PR China*

ARTICLE INFO

*Article history:*
Received 25 July 2016
Revised 24 February 2017
Accepted 22 March 2017
Available online 14 April 2017

*Keywords:*
Error estimation
*k*-fold cross-validation
Variance analysis
Model selection

ABSTRACT

Cross-validation (CV) is often used to estimate the generalization capability of a learning model. The variance of CV error has a considerable impact on the accuracy of CV estimator and the adequacy of the learning model, so it is very important to analyze CV variance. The aim of this paper is to investigate how to improve the accuracy of the error estimation based on variance analysis. We first describe the quantitative relationship between CV variance and its accuracy, which can provide guidance for improving the accuracy by reducing the variance. We then study the relationships between variance and relevant variables including the sample size, the number of folds, and the number of repetitions. These form the basis of theoretical strategies of regulating CV variance. Our classification results can theoretically explain the empirical results of Rodríguez and Kohavi. Finally, we propose a uniform normalized variance which not only measures model accuracy but also is irrelative to fold number. Therefore, it simplifies the selection of fold number in *k*-fold CV and normalized variance can serve as a stable error measurement for model comparison and selection. We report the results of experiments using 5 supervised learning models and 20 datasets. The results indicate that it is reliable to determine which variance is less before *k*-fold CV by the proposed theorems, and thus the accuracy of error estimation can be promoted by reducing variance. In so doing, we are more likely to select the best parameter or model.

## 1. Introduction

Modeling approaches that use supervised learning typically determine the optimal parameter or model by its generalization ability, which usually measured by the prediction error [1,2]. However, in various real problems, the prediction error cannot be calculated accurately because the underlying probability distribution is unknown. There are several estimators of prediction error, such as resubstitution [3], hold-out [4], *k*-fold cross-validation (CV)[5], repeated *k*-fold cross-validation [6,7] , the simple bootstrap and 0.632 bootstrap estimators [8]. The results in Refs.[9,10] indicate that *k*-fold CV and repeated *k*-fold CV generally produce better performance in model selection.

It is important to measure the uncertainty of prediction error estimators because the accuracy of the model selection is limited by the variance of error estimates [11,12]. A model error estimation can be considered as a random variable as the variability in training or test set [13,14], and its quality is usually measured by means of its bias and variance. The ideal estimator should be an efficient estimator which is unbiased and has the lowest variance. It is known that CV provides an unbiased estimate of the prediction error on the training set [1]. The variance is crucial for the accuracy of CV estimator. As well as being an important indicator for assessing estimators of prediction error, error estimators with low variance are quite interesting in model selection if we assume that the bias term is independent of the considered model [5].

Recent research on the variance of *k*-fold CV has mainly focused on estimation, decomposition, and some empirical studies of the variance.

Variance was estimated in different ways. Dietterich [15] and Alpaydin [16] employed the classical sample variance estimator to complete hypothesis tests for comparing classifiers, although this estimator is biased because of the overlap among training sets or test sets [1,17]. Moreover, Bengio and Grandvalet [1] showed that the bias could not be ignored, otherwise the variance would be grossly underestimated. The approximate variance estimator presented by Markatou [18] identifies all first-order terms in the reciprocal of the size of the training set. An improved estimator depending on the correlation among different group means was developed by Nadeau and Bengio [19], although the correlation is difficult to estimate. Bengio and Grandvalet [1] showed that there is no unbiased and universal estimator of the variance of *k*-fold CV that is valid under all distributions.

Variance decomposition can provide a better understanding of the sources and nature of variance. Bengio and Grandvalet [1] found that the variance of *k*-fold CV is a linear combination

* Corresponding author.
*E-mail addresses:* jianggaoxia@sxu.edu.cn (G. Jiang), wjwang@sxu.edu.cn (W. Wang).

http://dx.doi.org/10.1016/j.patcog.2017.03.025
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

of three moments, namely the variance of errors, within-fold covariance, and between-folds covariance. Rodriguez [20] proposed a novel theoretical decomposition of the variance that considers the sensitivity to changes in the training set and sensitivity to changes in the folds. Moreno-Torres [21] found the difference of variance under different cross-validation schemes, and empirical evidences support the conclusion that 2 × 5 (2-folds iterated five times) experiments converge significantly faster than 5 × 2 (5-folds iterated two times) and 10 × 1 (10-folds iterated once).

Some empirical results about the variance of *k*-fold CV have been summarized. It is assumed that repeated *k*-fold CV stabilizes the error estimation and, therefore, reduces the variance of the *k*-fold CV estimator, especially for small samples [22] . Rodriguez [20] found that the variance decreases with the sample size in all cases. Moreover, repeated *k*-fold CV results in lower variance than the non-repeated version. We can infer that the variance may be related to relevant variables such as the sample size, the number of folds, and the number of repetitions (in the repeated version).

Although there have been some brilliant achievements in *k*-fold CV variance analysis, there are some problems and difficulties. For example, there is no clear relation between the CV variance and prediction error, so it is hard to improve the accuracy of error estimator. Additionally, there are difficulties in providing a universal variance estimation, partly because we cannot theoretically explain the effects of some related variables on the variance. Many studies on CV variance [1,18,23] take a statistical perspective, but have no significant connection with machine learning models.

This paper focuses on variance analysis of *k*-fold CV for error estimation in supervised learning. We explain the effect of CV variance on the accuracy of error estimator by examining the distribution of the true error and estimated error. The relationships between CV variance and relevant variables are deduced from the CV procedure. A novel indicator is proposed to integrate the variances in classification and regression.

The remainder of this paper is organized as follows. In Section 2, we prove the relationship between the accuracy of error estimator and its variance, and it is validated by two examples. In Section 3, the relationships between the variance of *k*-fold CV and relevant variables are derived, and we introduce a new definition called normalized variance which measures model error and is independent of fold number. In Section 4, we present the results of numerical experiments in classification and regression. Section 5 concludes.

## 2. Expected absolute deviation and variance of cross-validation

### 2.1. Notations

*k*-fold CV is usually advocated to measure the prediction error of a learner. The dataset $D = \{(x_i, y_i)\}(i = 1, 2, \cdots, n)$ is partitioned into *k* groups or folds $F_1$, $F_2$, $\cdots$, $F_k$, such that $F_i \bigcap F_j = \varnothing$ for any $i \neq j$. $x_i$ and $y_i$ denote input feature(s) and output variable of *i*-th sample, respectively. $F_{-t} = D - F_t, t = 1, 2, \cdots, k$. For the sake of clarity and without loss of generality, we suppose that *n* is a multiple of *k*, where *n* denotes the size of the dataset. The size of each group is $m = n/k$, and *r* is the number of repetitions in repeated *k*-fold CV.

$e_A(D,\ T)$ denotes the error of a model *A* which is trained on dataset *D* and tested on dataset *T*. $\hat{e}_A^k(D, T)$ is the estimated error by *k*-fold CV, i.e. $\hat{e}_A^k(D, T) = \frac{1}{k} \sum_{t=1}^{k} e_A(F_{-t}, F_t) = \frac{1}{n} \sum_{i=1}^{n} e_A(F_{-t^*}, \{(x_i, y_i)\})$, where $t^* = \arg_t\ (x_i, y_i) \in F_t$. As an estimation of $e_A(D,\ T)$, $\hat{e}_A^k(D, T)$ is independent of *T* from above equations. Here $e_A(F_{-t^*}, \{(x_i, y_i)\})$ is the error on the *i*-th sample with the model trained on all folds except the one including the sample. And $e_A(F_{-t^*}, \{(x_i, y_i)\})$ can be simply written as $e_i (i = 1, 2, \ldots, n)$ which is calculated as the following way.

$$e_i = \begin{cases} I(\hat{y}_i \neq y_i) & \text{in classification} \\ |\hat{y}_i - y_i| & \text{in regression} \end{cases}$$

where $y_i$ denotes the real label in classification or real value in regression, and $\hat{y}_i$ denotes predicted label or value given by a learner on the test set. $I(\cdot)$ is the indicator function.

Let the true prediction error $e_{true} = e_A(D, D_p)$ and estimated (CV) error $e_{est} = \hat{e}_A^k(D, D_p)$, where $D_p$ is usually an unknown population in reality. The true error can be seen as a random variable with respect to the training set or model. The estimated error also can be seen as a random variable with respect to data partition for a given model and training set.

If CV works well, $e_{est}$ should approximate to $e_{true}$. Thus a new definition is given to measure the accuracy of prediction error estimator. Let

$$EAD = E(|e_{est} - e_{true}|) \tag{1}$$

where *EAD* is the expected absolute deviation of *k*-fold CV with respect to data set partition and $E(\cdot)$ is the expectation function. Note that the prediction error measures the accuracy of a model, whereas *EAD* measures the accuracy of CV.

Fig. 1 shows the deviation and variance of *k*-fold CV. Fig. 1(a) is the joint probability density distribution (*PDF*) of estimated and true prediction error *f*($e_{est}$, $e_{true}$), and Fig. 1(b) is the shadow slice in Fig. 1(a). The dash dot line on the floor of Fig. 1(a) denotes the ideal situation, which means $e_{est}$ is exactly equal to $e_{true}$. A dash line and a solid line appear in both sub-figures. The former denotes the situation when $e_{true} = e_{est} = 0.3$, whereas the latter denotes the center of the conditional distribution of $e_{est}$ for a fixed $e_{true}$. The conditional *PDF* in Fig. 1(b) is identical to $f(e_{est}, e_{true} = 0.3)$ except for a normalization factor [24]. Thus, the deviation is the distance between the solid line and the dash line. The distance between the solid line and the dot line $\sqrt{Variance}$ denotes the square root of the variance of $e_{est}$.

### 2.2. Expected absolute deviation and CV variance

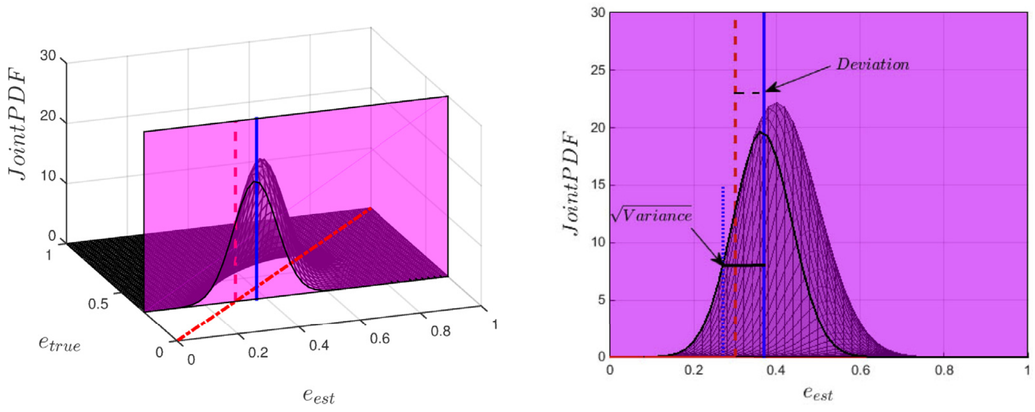
For a given learning problem, more than one true error value may exist due to the variety of model (modeling mechanism or parameter) and data set (size or samples). For a given data set and a training model, also there are kinds of estimated errors with different CV partitions (including partition number and sample attribution). There is a significant correlation between true error and estimated error because they are from the same model and most training samples $(n(k-1)/k)$. Considering the large quantity and correlation of true error and estimated error, we suppose that they are from a bivariate normal distribution.

**Lemma 1** [25]. *Let $X_1$ and $X_2$ follow the bivariate normal distribution whose PDF is $f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}]\}$. The conditional distribution of $X_1$ given that $X_2 = x_2$ is the normal distribution with mean and variance given by*

*$E(X_1|X_2 = x_2) = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2), Var(X_1|X_2 = x_2) = \sigma_1^2(1-\rho^2)$.*

**Theorem 1.** *Suppose that the estimated error $e_{est}$ and true error $e_{true}$ follow the bivariate normal distribution, $(e_{est}, e_{true}) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. For each fixed $e_{true}$, if $E(e_{est}) = e_{true}$, the variance of estimated error $Var(e_{est}) = \frac{\pi}{2(1-\rho^2)} \cdot EAD^2$ and $\frac{d(EAD)}{d(Var(e_{est}))} > 0$, where $\frac{d(\cdot)}{d(\cdot)}$ denotes derivative operation.*

**Proof.** From Lemma 1, $e_{est}|e_{true} \sim N(\mu_0, \sigma_0^2)$, where $\mu_0 = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(e_{true} - \mu_2)$ and $Var(e_{est}|e_{true}) = \sigma_0^2 = \sigma_1^2(1-\rho^2)$.



(a) Joint $PDF$ of estimated and true prediction error $f(e_{est}, e_{true})$

(b) A slice of the joint $PDF$ for a fixed true error $(e_{true} = 0.3)$

**Fig. 1.** Deviation and variance of prediction error estimator.

As $E(e_{est}) = e_{true}$, $e_{est}|e_{true} \sim N(e_{true}, \sigma_0^2)$.

Let $X = \frac{e_{est}-e_{true}}{\sigma_0}$, such that $X \sim N(0, 1)$. Note that $E(|X|) = \sqrt{\frac{2}{\pi}}$,

$$EAD = E(|e_{est} - e_{true}|) = \sigma_0 \cdot E(|X|) = \sqrt{\frac{2(1-\rho^2)}{\pi}} \cdot \sigma_1 \qquad (2)$$

or

$$Var(e_{est}) = \sigma_1^2 = \frac{\pi}{2(1-\rho^2)} \cdot EAD^2. \qquad (3)$$

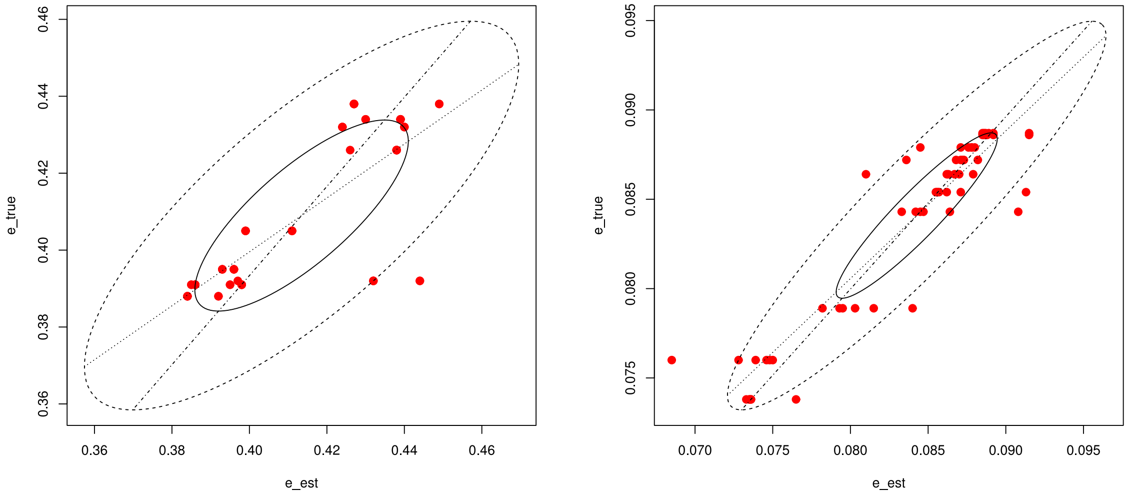
For $EAD > 0$, we have $\frac{d(EAD)}{d(Var(e_{est}))} = \sqrt{\frac{1-\rho^2}{2\pi Var(e_{est})}} > 0$. □

The assumptions and conclusion of Theorem 1 are validated by two traditional supervised learning examples.

Support vector machines (SVMs) and feedforward neural networks (NNs) are trained with $k$-fold CV on two UCI datasets, *EEG Eye State* (*EEG*, 14,980 instances and 15 features) and *Online News Popularity*(*ONP*, 39,644 instances and 60 features) [26] for classification and regression, respectively. The original sets are assumed to be the population $D_p$. Samples in $D_p^{pc}$ are randomly selected from $D_p$ and $pc$ is the proportion of sample size ($pc =$ 5%, 10%, 15%, $\cdots$, 50%). The true error of model A trained on $D_p^{pc}$ can be expressed as $e_{true} = e_A(D_p^{pc}, D_p)$. For selected $D_p^{pc}$ and model, $e_{true}$ is estimated by CVs with different fold numbers ($k =$ 2, 5, 10, 20, 30, 50). $Var(e_{est})$ are estimated by repeating the CV procedure 10 times for each $pc$ and each $k$. Then $E(e_{est})$ can be approximated by $\overline{e_{est}}$, the average of 60 estimations with different $k$ values and different partitions. $EAD$ is obtained by averaging $|e_{est} - e_{true}|$ with respect to partition for each $pc$ and each $k$.

(1) Validations of normality and unbiasedness

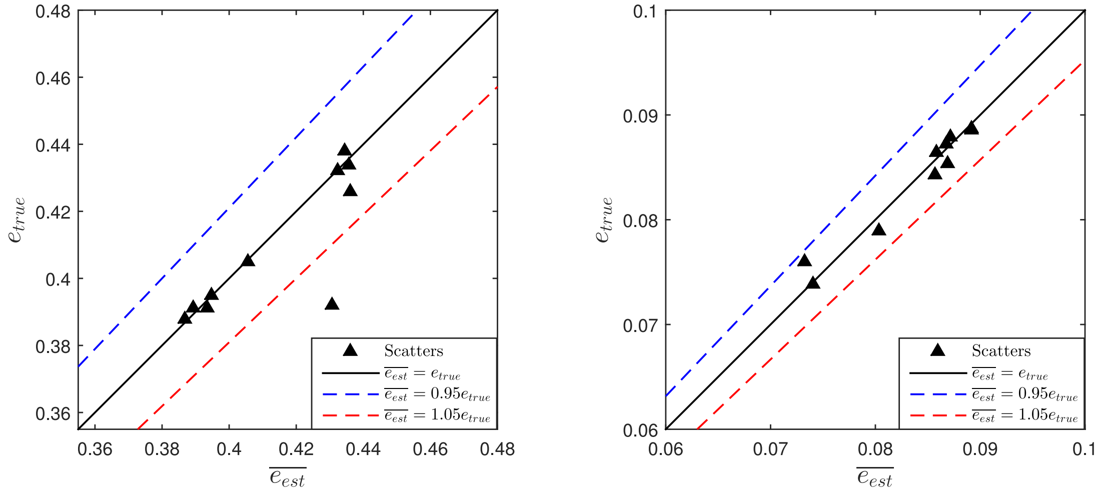
$e_{est}$ and $e_{true}$ are assumed to follow the bivariate normal distribution in Theorem 1. Now the normality assumption is examined on data sets *EEG* and *ONP*. $e_{est}$ and $e_{true}$ can be calculated from the previous paragraph. 120 scatters (6 $k$ values, 10 $pc$ values, 2 data sets) representing $e_{est}$ and $e_{true}$ are plotted in Fig. 2. In each sub-figure, two dashed lines intersect at the center of a distribution. It can be observed that the contours of distributions are almost elipses. Moreover, the correlation coefficients of scatters in the two sub-figures are 0.78 and 0.93, respectively. And thus $e_{est}$ and $e_{true}$ are intuitively from normal distribution with positive correlation for both classification and regression.



(a) Classification on $EEG$

(b) Regression on $ONP$

**Fig. 2.** Distribution of $e_{true}$ and $e_{est}$.



(a) Classification on $EEG$

(b) Regression on $ONP$

**Fig. 3.** Distribution of $e_{true}$ and $\overline{e_{est}}$.

Unbiasedness of $e_{est}$ which is expressed as $E(e_{est}) = e_{true}$ is assumed in Theorem 1. Considering that $E(e_{est})$ is usually approximated by $\overline{e_{est}}$ in real problems, the relationship between $\overline{e_{est}}$ and $e_{true}$ is tested here. Fig. 3 shows 20 scatters (10 $pc$ values, 2 data sets) representing $e_{true}$ and $\overline{e_{est}}$. In each sub-figure, the solid line is diagonal, and dashed lines are biased positively or negatively at 5%. It can be observed that scatters are around the diagonal, and all deviations between $\overline{e_{est}}$ and $e_{true}$ are within 5% of $e_{true}$ except for one outlier in classification. So $E(e_{est})$ should be very close to $e_{true}$, too. Moreover, the frequencies of $e_{est} > e_{true}$ and $e_{est} < e_{true}$ are 27/60 and 33/60 (6 $k$ values, 10 $pc$ values) in classification, respectively. Both of them are 30/60 in regression. Thus it is reasonable to assume the unbiasedness of $e_{est}$ on the two data sets.

It is known that CV error is unbiased for the error learned from $n(k-1)/k$ samples but not $n$ samples. When $n$ and $k$ are large enough, information of $n(k-1)/k$ samples will be very close to that of $n$ samples. Then the error estimator of $k$-fold CV can be thought to be unbiased and Theorem 1 is applicable.

(2) Validation of the conclusion (the relation of $EAD$ and CV variance)

The quadratic relation of $EAD$ and CV variance described in Theorem 1 is also examined on the two data sets. The distributions of $EAD$ and $Var(e_{est})$ in classification and regression are displayed in Figs. 4 and 5, respectively. Scatters representing $EAD$ and $Var(e_{est})$ are denoted as $'*'$. Considering the type of Eq. (3) in Theorem 1, scatters are fitted to a specific quadratic equation $Var(e_{est}) = c \cdot EAD^2$ ($c$ is an undetermined coefficient). The goodness of fit $R^2$ describes how well the quadratic equation fits the scatters.

As shown in Figs. 4 and 5, $R^2$ increases with $k$ on each data set. It means that the larger $k$ is, the better the fit is. In other words, the larger $k$ is, the more evident the quadratic relation of $EAD$ and $Var(e_{est})$ is. The reason may be that the model trained on $k-1$ folds containing $n(k-1)/k$ samples is close to that on the whole set ($n$ samples) when $k$ is large enough. So the assumption of unbiasedness is suitable for large fold number, and the experimental results are also closer to the theoretical result of Theorem 1 for large $k$.

Above experiment results indicate that, CV estimator can be seen as unbiased and $EAD$ and $Var(e_{est})$ are from the quadratic relation when both the size of data set and the number of folds are large enough.

## 3. Variance analysis of $k$-fold CV

### 3.1. Variances

We define three kinds of CV estimators for the prediction error of learners.

- The **group mean** estimates the prediction error with the mean of the errors in a group or fold,

$$\hat{e}^{(t)} = \frac{1}{m}\sum_{(x_i,y_i)\in F_t} e_i, t = 1, 2, \ldots, k. \tag{4}$$

- The **total mean** estimates the prediction error with the mean of the errors in all groups or folds,

$$\hat{e}^{(k)} = \frac{1}{k}\sum_{t=1}^{k}\hat{e}^{(t)} = \frac{1}{k}\sum_{t=1}^{k}\left(\frac{1}{m}\sum_{(x_i,y_i)\in F_t} e_i\right) = \frac{1}{mk}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n} e_i. \tag{5}$$

- The **repeated mean** estimates the prediction error with the mean of the errors in all repeated groups or folds,

$$\hat{e}^{(k,r)} = \frac{1}{r}\sum_{j=1}^{r}\hat{e}^{(k_j)}, \tag{6}$$

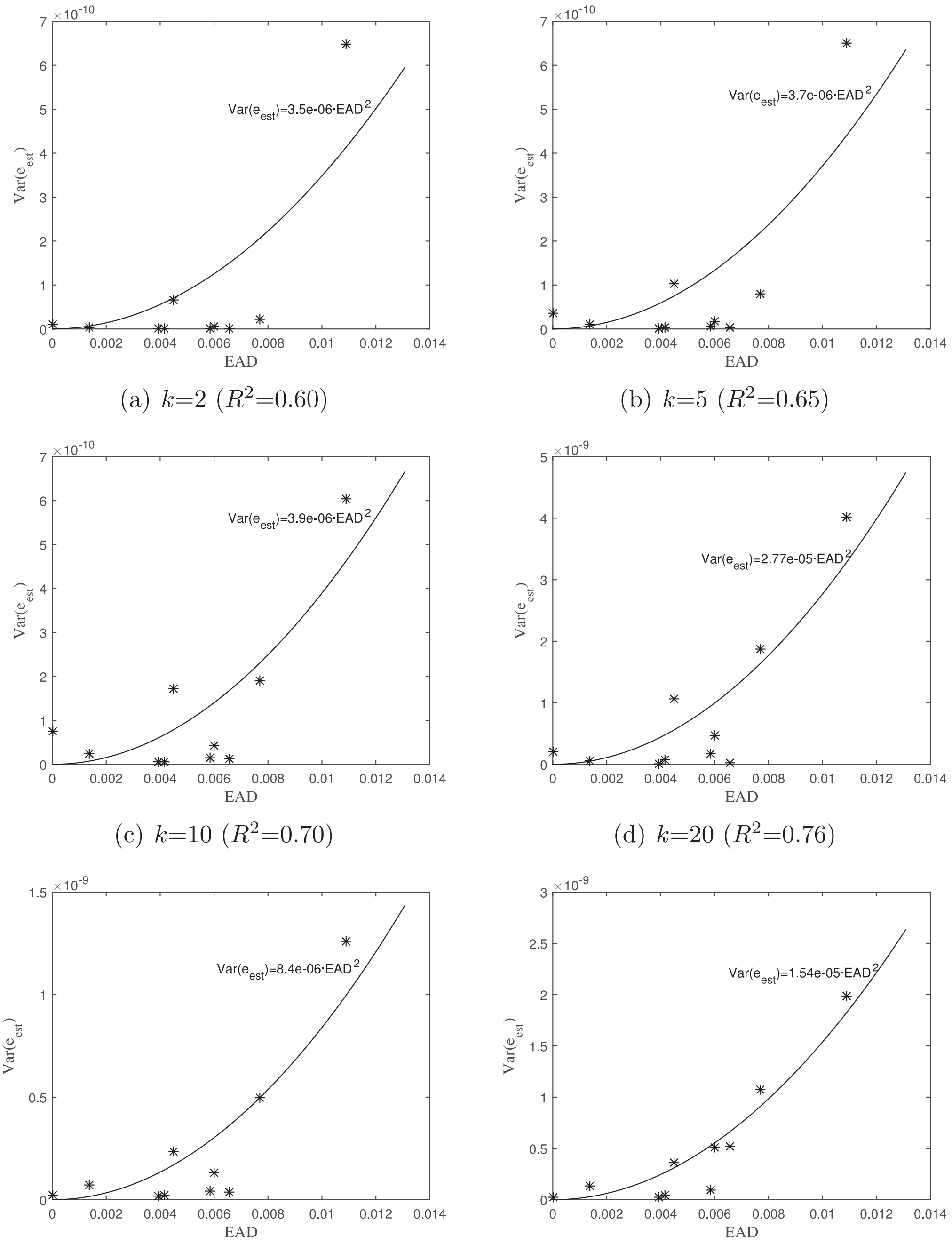
  where $k$ is the number of folds and $r$ is the number of repetitions.

As the partition in CV is not fixed (except for leave-one-out CV), CV estimation is not a determinate value. There are three kinds of CV variances of the above estimators from the variability of data partition.

- The **group variance** (variance from different groups): $Gvar = Var(\hat{e}^{(t)})$;
- The **total variance** (normal variance of $k$-fold CV from different partitions): $Tvar = Var(\hat{e}^{(k)})$;
- The **repeated variance** (variance of repeated $k$-fold CV from different partitions): $Rvar = Var(\hat{e}^{(k,r)})$, where $Var(\cdot)$ is the variance function.

### 3.2. Variance analysis of k-fold CV

This subsection mainly describes two theorems about CV variances in classification and regression.



(a) $k$=2 ($R^2$=0.60)

(b) $k$=5 ($R^2$=0.65)

(c) $k$=10 ($R^2$=0.70)

(d) $k$=20 ($R^2$=0.76)

(e) $k$=30 ($R^2$=0.78)

(f) $k$=50 ($R^2$=0.89)

**Fig. 4.** *EAD* and *Var*($e_{est}$) on *EEG*.

**Theorem 2.** *Let $p(0 < p < 1)$ be the accuracy (the probability of right prediction) of a classification model on someone data set. Suppose the classifier is stable under the perturbations caused by data partition in /repeated k-fold CV [9]. If group means $\hat{e}^{(t)}(t{=}1,2,...,k)$ in /repeated k-fold CV are mutually independent [13], then we have the following approximations:*

$$Tvar = \frac{p(1-p)}{n}, \tag{7}$$

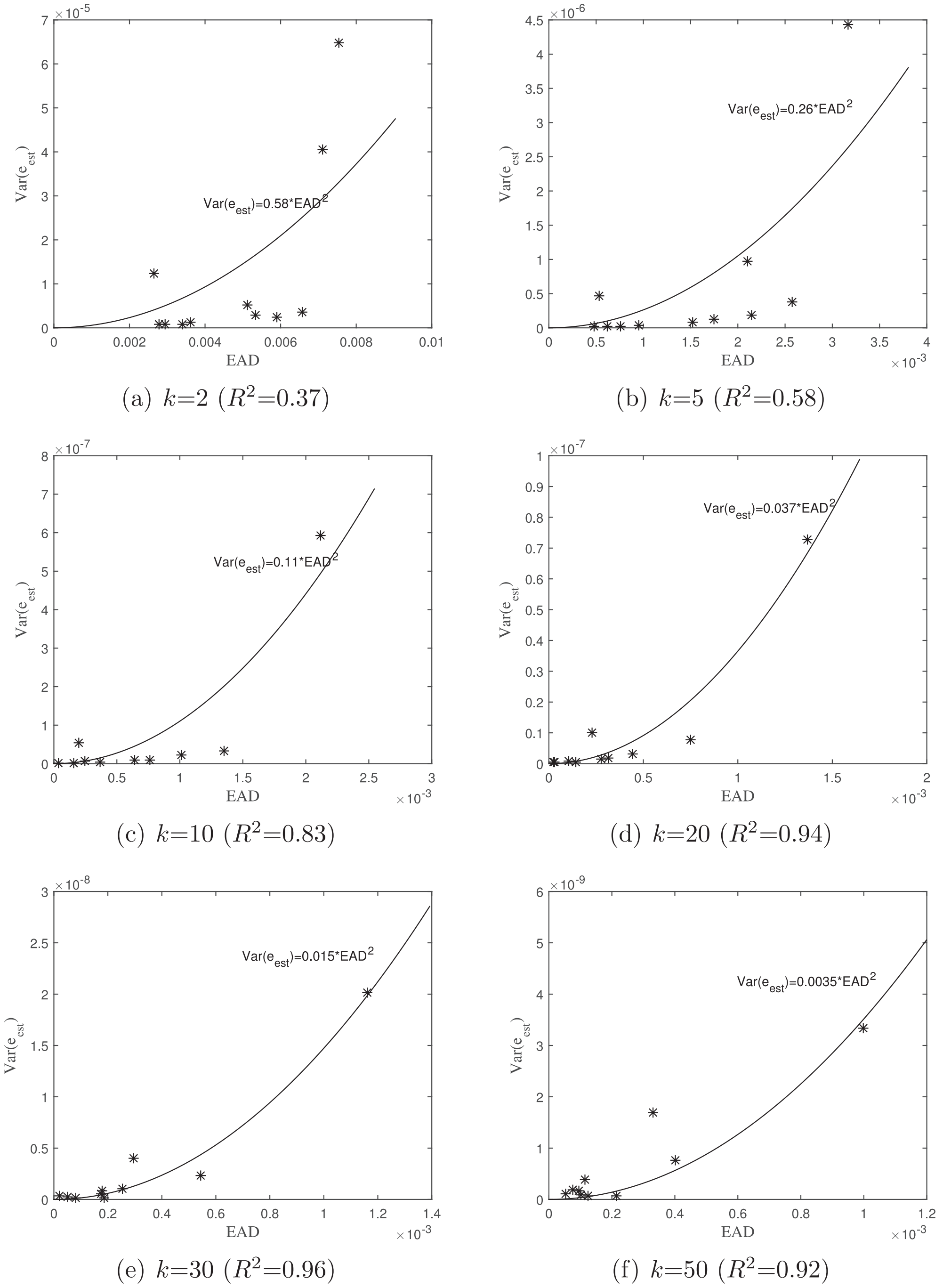
$$Gvar = \frac{kp(1-p)}{n}, \tag{8}$$

$$Rvar = \frac{p(1-p)}{nr}. \tag{9}$$

**Proof.**

(1) Ref. [13] showed that if the numbers of correct and wrong predictions are both not less than five, the total accuracy (frequency of right prediction on the whole data set, $1 - \hat{e}^{(k)}$) can be approximated by a normal distribution with mean $p$ and variance $p(1-p)/n$ according to central limit theorem.
Thus the total mean has the approximate distribution: $\hat{e}^{(k)} \sim N(1-p, p(1-p)/n)$, and $Tvar = \frac{p(1-p)}{n}$.

(2) The group means are identically distributed due to the stability of classifier. By the independence of group mean, $Tvar = Var(\hat{e}^{(k)}) = Var(\frac{1}{k}\sum_{t=1}^{k}\hat{e}^{(t)}) = \frac{k \cdot Var(\hat{e}^{(t)})}{k^2} = \frac{1}{n}p(1-p)$.
Thus $Gvar = Var(\hat{e}^{(t)}) = \frac{kp(1-p)}{n}$.



(a) $k$=2 ($R^2$=0.37)

(b) $k$=5 ($R^2$=0.58)

(c) $k$=10 ($R^2$=0.83)

(d) $k$=20 ($R^2$=0.94)

(e) $k$=30 ($R^2$=0.96)

(f) $k$=50 ($R^2$=0.92)

**Fig. 5.** *EAD* and $Var(e_{est})$ on *ONP*.

(3) By the property of the variance function, $Rvar = Var(\hat{e}^{(k,r)}) = Var(\frac{1}{r}\sum_{j=1}^{r}\hat{e}^{(k_j)}) = \frac{1}{r^2}Var(\sum_{j=1}^{r}\hat{e}^{(k_j)})$.
As the $k$-fold partitions in repeated $k$-fold CV are not dependent on both random and stratified partition, the $r$ total means $(\hat{e}^{(k_j)})$ can be considered as independent variables. So we have $\frac{1}{r^2}Var(\sum_{j=1}^{r}\hat{e}^{(k_j)}) = \frac{1}{r^2}\sum_{j=1}^{r}Var(\hat{e}^{(k_j)}) = \frac{1}{r^2}\sum_{j=1}^{r}Tvar = \frac{1}{r^2}\cdot r\cdot\frac{1}{n}p(1-p) = \frac{1}{nr}p(1-p)$. And thus $Rvar = \frac{1}{nr}p(1-p)$. □

Note that the above approximations hold under the condition that the numbers of correct and wrong predictions are both not less than five. If the classifier is not too bad or the data is not seriously imbalanced, this condition could be satisfied and variances can be approximated by Eqs. (7)–(9).

**Theorem 3.** *Let $e_0$ be the error (the average absolute deviation between predicted value and actual value) of a regression model on someone data set, and its variance with respect to different test samples is $\sigma^2$. Suppose the model is stable under the perturbations caused by data partition in /repeated k-fold CV [9]. If group means $\hat{e}^{(t)}$ (t=1,2,...,k) in /repeated k-fold CV are mutually independent [13], then we have the following approximations:*

$$Tvar = \frac{\sigma^2}{n}, \tag{10}$$

$$Gvar = \frac{k\sigma^2}{n}, \tag{11}$$

$$Rvar = \frac{\sigma^2}{nr}. \tag{12}$$

**Proof.**

(1) In many practical cases, if the number of random variables is larger than 30, the normal approximation will be satisfactory regardless of the shape of the population. If not, the central limit theorem will work if the distribution of the population is not severely nonnormal [27]. The group mean can be seen as a random variable with regard to different folds. We can conclude that if the fold number $k$ is over 30 or the distribution of group mean is not severely nonnormal, the distribution of total mean can be approximated by a normal distribution, i.e. $\hat{e}^{(k)} \sim N(e_0, \sigma^2/n)$, and $Tvar = \frac{\sigma^2}{n}$.

(2) The group means are identically distributed due to the stability of regression model. By the independence of group mean, $Tvar = Var(\hat{e}^{(k)}) = Var(\frac{1}{k}\sum_{t=1}^{k}\hat{e}^{(t)}) = \frac{k\cdot Var(\hat{e}^{(t)})}{k^2} = \frac{\sigma^2}{n}$. Thus $Gvar = Var(\hat{e}^{(t)}) = \frac{k\sigma^2}{n}$.

(3) As the $k$-fold partitions in repeated $k$-fold CV are not dependent on both random and stratified partition, the $r$ total means can be considered as independent variables. So we have
$Rvar = Var(\hat{e}^{(k,r)}) = Var(\frac{1}{r}\sum_{j=1}^{r}\hat{e}^{(k_j)}) = \frac{1}{r^2}\sum_{j=1}^{r}Var(\hat{e}^{(k_j)}) = \frac{1}{r^2}\cdot r\cdot\frac{1}{n}\sigma^2 = \frac{1}{nr}\sigma^2$. □

As a special case of $k$-fold CV, leave-one-out CV (LOOCV) has the lowest bias in estimating regression error and is used in model selection [28,29]. Its variances can also be approximated by Eqs. (10)–(12) on data set whose size is larger than 30. Based on this, interval estimation is available for error estimation.

### 3.3. Normalized variance

Obviously, $p(1-p)$ or $\sigma^2$ appear in all of the classification or regression equations, respectively. A new variance can be defined as follows.

**Definition 1.** The normalized variance ($Nvar$) of $k$-fold CV is defined as:

$$Nvar = \frac{Gvar}{k/n} = \frac{Tvar}{1/n} = \frac{Rvar}{1/nr} \tag{13}$$

or

$$Nvar = \begin{cases} p(1-p) & \text{in classification} \\ \sigma^2 & \text{in regression} \end{cases}$$

Theoretically speaking, $Nvar$ integrates the variances of classification and regression, and is only inversely proportional to the true accuracy of the model.

(1) In real classification problems, $Nvar$ can be estimated by $Gvar$, $Tvar$ or $Rvar$. Then, the accuracy of the classification model has the estimation $\hat{p} = \frac{1}{2} \pm \frac{1}{2}\sqrt{1-4Nvar}$. Note that the accuracy of most two-class classifiers is higher than that of a random model ($p > 0.5$), and thus $\hat{p} = \frac{1}{2} + \frac{1}{2}\sqrt{1-4Nvar}$.

Another common estimate of the accuracy is the average $\hat{p}' = 1 - \frac{1}{n}\sum_{i=1}^{n} e_i$. Ideally, therefore, $\frac{1}{2} + \frac{1}{2}\sqrt{1-4Nvar} \approx 1 - \frac{1}{n}\sum_{i=1}^{n} e_i$ or

$$Nvar \approx \hat{p}'(1-\hat{p}') = (1 - \frac{1}{n}\sum_{i=1}^{n} e_i)\cdot\frac{1}{n}\sum_{i=1}^{n} e_i. \tag{14}$$

(2) In real regression problems, $\sigma^2$ can be estimated by $Nvar$.

The sample variance $\frac{1}{n-1}\sum_{i=1}^{n} {e_i}^2$ is a general estimate of $\sigma^2$, so it should have

$$Nvar \approx \frac{1}{n-1}\sum_{i=1}^{n} {e_i}^2 \approx \frac{1}{n}\sum_{i=1}^{n} {e_i}^2 = MSE \tag{15}$$

when $n$ is sufficiently large. $MSE$ denotes mean square error of regression models. Eqs. (14) and (15) will be validated in our experiments (Section 4.3).

### 3.4. Explanation of some empirical results

Rodriguez [20] found that

(1) In all (classification) cases, the variance of the CV estimator decreases with the sample size ($n$).
(2) The variance of the estimator is lower for repeated $k$-fold CV than for the non-repeated version.

The above results can be proved by the proposed theorems.

(1) This is clear from Eq. (7) in Theorem 2.
(2) By Eqs. (7) and (9), we have $Rvar = \frac{1}{nr}p(1-p) < \frac{1}{n}p(1-p) = Tvar$ as the number of repetitions $r > 1$.

It is assumed that the repeated CV stabilizes the error estimation and, therefore, reduces the variance of the $k$-fold CV estimator, especially for small samples [22].

This can be proved by Theorem 2.

(1) For $Rvar = \frac{1}{nr}p(1-p) < \frac{1}{n}p(1-p) = Tvar$, repeated $k$-fold CV reduces the variance of the $k$-fold CV.
(2) By Eqs. (7) and (9), the decrement $Tvar - Rvar = \frac{1}{n}p(1-p) - \frac{1}{nr}p(1-p) = \frac{r-1}{nr}p(1-p)$. It is obvious that this decrement becomes notable when the sample size is small.

## 4. Experiments and analysis

In this section, we empirically check the above conclusions, including Theorems 2 and 3 and Eqs. (14) and (15). We present our experimental framework, empirical results, and analysis.

### 4.1. Experimental framework

To achieve broad coverage, 20 data sets were employed for learning five kinds of models [26,30]. The first ten binary-class or multi-class problems have 5–60 inputs and 270–4898 examples. The other data sets for regression have 6–68 inputs and 308–5875 examples (see Table 1). Each data set was used to train two NNs with 20 and 10 + 10 hidden nodes (NN20, NN10+10) and three SVMs with linear, polynomial, and Gaussian kernels (SVM_L, SVM_P, SVM_G). The CV variances were examined for $k = 2, 5, 10, 20$. The sample scales were 100%, 50%, and 25% of the whole dataset, and the number of repetitions was taken as either 5, 10, or 15. The samples were partitioned in random or stratified way. $Gvar$, $Tvar$, and $Rvar$ were obtained by computing the mean of ten sample variance estimations. More specifically, each $Tvar$ was estimated by ten CV errors, and ten $Tvar$ values were averaged as a result. It means CV was repeated 100 times to calculate a $Tvar$ value. Three kinds of $Nvar$ values were then obtained from the definition in Eq. (13). The corresponding $\hat{p}'(1-\hat{p}')$ and $MSE$ were also calculated by $e_i$ in Eqs. (14) and (15). Altogether, 2400 sets of test results were obtained for the comparison of variances and validation of the above conclusions.

**Table 1**
Dataset information.

| Learning task | No. | Dataset | Instances | Features | Classes |
|---|---|---|---|---|---|
| Classification | 1 | Heart | 270 | 13 | 2 |
| | 2 | Breast cancer | 683 | 10 | 2 |
| | 3 | Blood transfusion | 748 | 5 | 2 |
| | 4 | Vehicle | 846 | 18 | 4 |
| | 5 | Splice | 1000 | 60 | 2 |
| | 6 | Diabetes | 1151 | 20 | 2 |
| | 7 | Wine quality (red) | 1599 | 12 | 6 |
| | 8 | Segment | 2310 | 19 | 7 |
| | 9 | Abalone | 4177 | 8 | 29 |
| | 10 | Wine quality (white) | 4898 | 12 | 7 |
| Regression | 11 | Yacht Hydrodynamics | 308 | 7 | – |
| | 12 | Housing | 506 | 14 | – |
| | 13 | Energy efficiency | 768 | 8 | – |
| | 14 | Concrete | 1030 | 9 | – |
| | 15 | Geographical origin of music | 1059 | 68 | – |
| | 16 | MG | 1385 | 6 | – |
| | 17 | Airfoil self-noise | 1503 | 6 | – |
| | 18 | Space_ga | 3107 | 6 | – |
| | 19 | Skill craft master table | 3395 | 20 | – |
| | 20 | Parkinson's telemonitoring | 5875 | 26 | – |

**Table 2**
Average of frequencies on 5 models.

| Dateset | *Gk* | *Gn* | *Tn* | *Rr* | *Rn* |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 0.94 | 0.94 | 0.85 |
| 2 | 0.94 | 0.88 | 0.79 | 0.94 | 0.73 |
| 3 | 0.96 | 0.81 | 0.69 | 0.93 | 0.76 |
| 4 | 1 | 0.98 | 0.94 | 0.95 | 0.85 |
| 5 | 1 | 1 | 0.99 | 0.91 | 0.90 |
| 6 | 1 | 0.98 | 0.95 | 0.92 | 0.87 |
| 7 | 1 | 1 | 0.98 | 0.94 | 0.93 |
| 8 | 1 | 0.90 | 0.78 | 0.92 | 0.93 |
| 9 | 0.76 | 1 | 0.97 | 0.96 | 0.97 |
| 10 | 1 | 0.98 | 0.98 | 0.94 | 0.93 |
| 11 | 0.87 | 0.92 | 0.91 | 0.96 | 0.96 |
| 12 | 0.89 | 0.91 | 0.91 | 0.94 | 0.90 |
| 13 | 0.79 | 0.85 | 0.80 | 0.94 | 0.91 |
| 14 | 0.96 | 0.89 | 0.88 | 0.94 | 0.94 |
| 15 | 0.88 | 0.84 | 0.93 | 0.94 | 0.96 |
| 16 | 0.89 | 1 | 0.90 | 0.96 | 0.94 |
| 17 | 0.99 | 0.91 | 0.87 | 0.95 | 0.91 |
| 18 | 0.91 | 0.93 | 0.93 | 0.96 | 0.96 |
| 19 | 0.88 | 0.96 | 0.96 | 0.95 | 0.94 |
| 20 | 0.93 | 0.94 | 0.95 | 0.94 | 0.97 |
| Average | 0.93 | 0.93 | 0.90 | 0.94 | 0.91 |

### 4.2. Results and analysis for Theorems 2 and 3

As Theorems 2 and 3 provide six equations Eqs. (7)–(12) about CV variances and other variables, we test them by examining the order relationship of each theoretical result. To quantify the order relationships or order consistency of experimental results with our equations, some conditional frequencies are defined as follows.

$$Gk = \hat{P}(Gvar_i > Gvar_j | k_i > k_j, n_i = n_j)$$

$$Gn = \hat{P}(Gvar_i > Gvar_j | n_i < n_j, k_i = k_j)$$

$$Tn = \hat{P}(Tvar_i > Tvar_j | n_i < n_j, k_i = k_j)$$

$$Rr = \hat{P}(Rvar_i > Rvar_j | r_i < r_j, n_i = n_j, k_i = k_j)$$

$$Rn = \hat{P}(Rvar_i > Rvar_j | n_i < n_j, r_i = r_j, k_i = k_j)$$

By Eqs. (8) and (11), it is obvious that *Gvar* increases with *k*. Meanwhile, the definition of the first frequency shows that *Gk* measures the probability of the situation that *Gvar* increases with *k* when *n* is fixed. So *Gk* can quantify the order relationship between *Gvar* and *k* implied by Eqs. (8) and (11). Similarly, the other four frequencies correspond to the relationships between *Gvar* and *n* in Eqs. (8) and (11), *Tvar* and *n* in Eqs. (7) and (10), *Rvar* and *r* in Eqs. (9) and (12), *Rvar* and *n* in Eqs. (9) and (12), respectively. All of frequencies should be 1 if Eqs. (7)–(12) in the two theorems hold.

CVs have been completed under various settings including data sets (20), models (5), *n* (3), *k* (6), *r* (3) and the ways of partitions (2). Then frequencies can be counted by data sets and models, and they are plotted in Fig. 6 consisting of 5 sub-figures (5 kinds of frequencies). In each sub-figure, frequencies are denoted as a circle with different sized wedges on 20 data sets (20 circles). The radius of each wedge in the circle indicates the frequency of a model. The radius in the legend is 1. It can be seen that all *Rr* values (sub-figure (d)) are equal or close to 1. A few values of other four frequencies are obviously less than 1. To examine these values violating order relationships, Table 2 lists frequencies on each data set. Each value is calculated by averaging one kind of frequencies of 5 models, i.e., it represents the average radius of five wedges in a circle of Fig. 6.

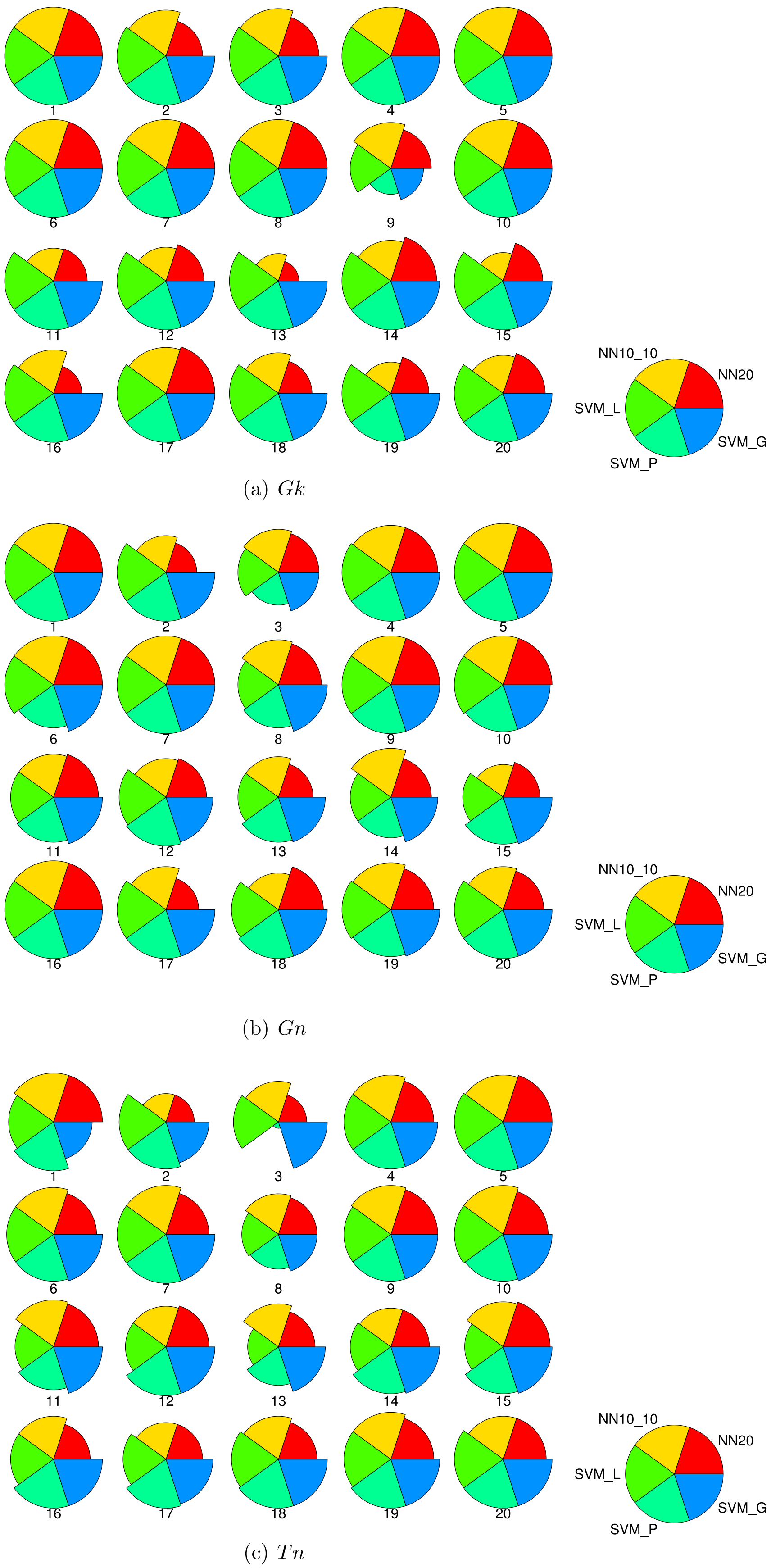
**Table 3**
Average of five frequencies.

| Dataset | NN20 | NN10+10 | SVM_L | SVM_P | SVM_G | Average |
|---|---|---|---|---|---|---|
| 1 | 0.94 | 0.96 | 0.96 | 0.98 | 0.91 | 0.95 |
| 2 | 0.71 | 0.76 | 0.96 | 0.93 | 0.92 | **0.86** |
| 3 | 0.74 | 0.83 | 0.93 | 0.70 | 0.96 | **0.83** |
| 4 | 0.90 | 0.93 | 0.96 | 0.96 | 0.97 | 0.95 |
| 5 | 0.94 | 0.96 | 0.98 | 0.96 | 0.97 | 0.96 |
| 6 | 0.91 | 0.94 | 0.96 | 0.93 | 0.97 | 0.94 |
| 7 | 0.91 | 0.98 | 0.99 | 0.98 | 0.98 | 0.97 |
| 8 | 0.88 | 0.93 | 0.88 | 0.88 | 0.95 | 0.90 |
| 9 | 0.92 | 0.99 | 0.95 | 0.88 | 0.91 | 0.93 |
| 10 | 0.94 | 0.96 | 0.99 | 0.97 | 0.97 | 0.96 |
| 11 | 0.87 | 0.89 | 0.92 | 0.94 | 0.99 | 0.92 |
| 12 | 0.85 | 0.80 | 0.92 | 1 | 0.98 | 0.91 |
| 13 | 0.74 | 0.81 | 0.85 | 0.92 | 0.97 | **0.86** |
| 14 | 0.89 | 0.89 | 0.90 | 0.94 | 0.98 | 0.92 |
| 15 | 0.87 | 0.82 | 0.91 | 0.97 | 0.99 | 0.91 |
| 16 | 0.84 | 0.92 | 0.96 | 0.98 | 1 | 0.94 |
| 17 | 0.84 | 0.89 | 0.94 | 0.98 | 0.98 | 0.93 |
| 18 | 0.84 | 0.87 | 0.98 | 1 | 0.99 | 0.94 |
| 19 | 0.87 | 0.88 | 0.99 | 0.96 | 0.99 | 0.94 |
| 20 | 0.88 | 0.89 | 0.98 | 0.98 | 0.99 | 0.95 |
| Average | **0.86** | **0.89** | 0.95 | 0.94 | 0.97 | 0.92 |

In Table 2, there are 76 values not less than 0.9 in all 100 values, 17 values between 0.8 and 0.9, and 7 values less than 0.8. The 76 values consist of 13 *Gk* values, 15 *Gn* values, 14 *Tn* values, 20 *Rr* values and 14 *Rn* values. All average frequencies (the last row) on 20 data sets are over 0.9.

There are 39 and 37 values greater than 0.9 in classification (the former 10 data sets) and regression (the latter 10 data sets), respectively. The averages are 0.93 and 0.92 in two kinds of problems. So the frequencies are unrelated to the learning task (classification or regression). However, different situations appear for several data sets and learning models. Frequencies are counted by data sets and learning models in Table 3, i.e., each value is calculated by averaging five kinds of frequencies of a model on a data set. Bold font highlights the average results less than 0.9.

In Table 3, the average values by data sets (the last column) in bold font indicate that inconsistent order relationships are more likely to occur in datasets 2, 3, and 13. This may be related to the quality of these datasets, such as the presence of outliers or too few samples (less than 800) for CV with large values of *k*. They may enhance the perturbation of $\hat{e}^{(t)}$ and lead to unstable performance of models. From the view of learning model, the results of

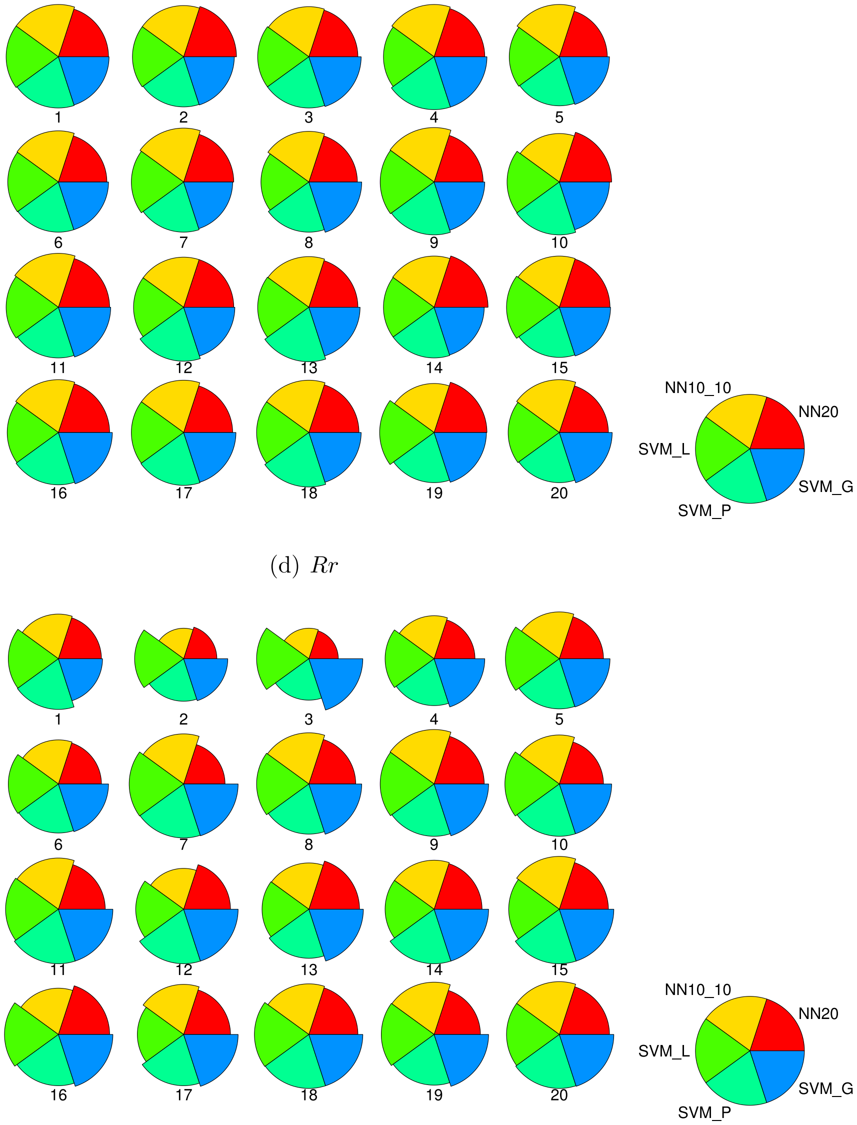


(a) $Gk$

(b) $Gn$

(c) $Tn$

**Fig. 6.** Star plot of each frequency for 5 models on 20 data sets.



(d) $Rr$

(e) $Rn$

**Fig. 6.** Continued

NNs are less than 0.9, while those of SVMs are over 0.9. In other words, NNs are more likely to have low frequencies than SVMs in our experiments. This may be because one of the most important NN parameters, the number of hidden nodes, has not been optimized, and the two NNs may not be suited for some datasets. Whereas the parameters in LIBSVM have been optimized prior to the validation [30]. The experiment results support that consistent order relationships are more likely to appear in well-specified models.

Three kinds of variances on classification dataset *Splice* and regression dataset *Parkinson's Telemonitoring* are plotted in Figs. 7 and 8 to show the relations in Theorems 2 and 3 intuitively. From these figures, it can be seen that *Gvar* increases with $k$ and decreases with $n$. In each model, *Tvar* decreases with $n$. *Rvar* decreases with $n$ and $r$. All order relationships are consistent with the theorems.

Generally speaking, although they are affected by small-scale data set or mismatching model, five average frequencies in Table 2 are over 0.9. It indicates that our equations in the proposed theorems usually hold in terms of the order in real supervised learning problems. Thus, it is reliable to determine or predict which variance is less by our theorems before applying $k$-fold CV.

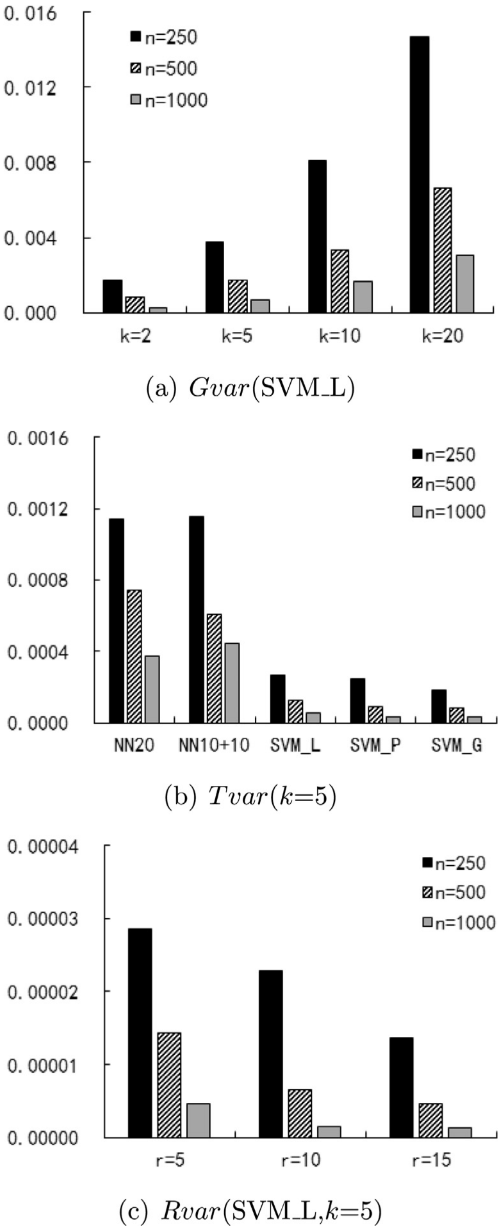
**Table 4**
Pearson correlation coefficient of *Nvar* and error.

| *Nvar* | $\hat{e}$(classification) | $\hat{p}'(1-\hat{p}')$ | $\hat{e}$(regression) | *MSE* |
|---|---|---|---|---|
| $Gvar \cdot n/k$ | 0.41 | 0.69 | 0.76 | 0.82 |
| $Tvar \cdot n$ | 0.14 | 0.36 | 0.32 | 0.37 |
| $Rvar \cdot n \cdot r$ | 0.12 | 0.33 | 0.33 | 0.37 |

### 4.3. Results and analysis for normalized variance

In Eq. (13), *Nvar* is defined in three ways, $Gvar \cdot n/k$, $Tvar \cdot n$, and $Rvar \cdot n \cdot r$. Eqs. (14) and (15) show the theoretical relationships between *Nvar* and $\hat{p}'(1-\hat{p}')$ or *MSE*. The empirical relations are examined by Pearson correlation coefficients of *Nvar* values and indicators about model error listed in Table 4.

From Table 4, the coefficients of $Gvar \cdot n/k$ are obviously larger than those of $Tvar \cdot n$ and $Rvar \cdot n \cdot r$. Coefficients of *Nvar* and

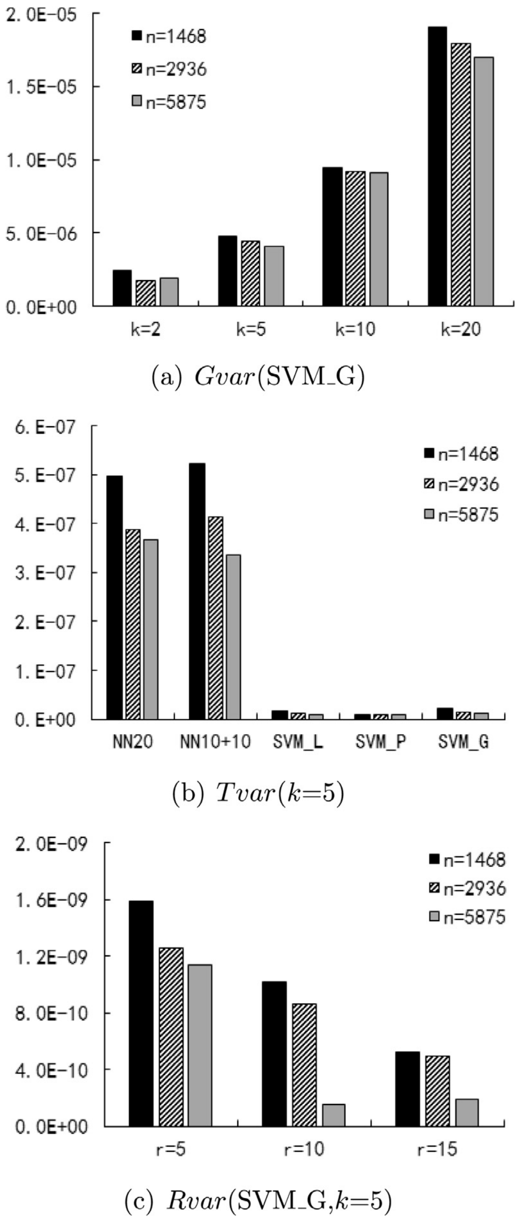


(a) $Gvar$(SVM_L)

(b) $Tvar(k{=}5)$

(c) $Rvar$(SVM_L,$k$=5)

**Fig. 7.** Classification variances on *Splice*.



(a) $Gvar$(SVM_G)

(b) $Tvar(k{=}5)$

(c) $Rvar$(SVM_G,$k$=5)

**Fig. 8.** Regression variances on *Parkinson's Telemonitoring*.

$\hat{p}'(1-\hat{p}')$ or *MSE* are larger than those of *Nvar* and $\hat{e}$ both for classification and regression. That is because the former relation is linear and the latter relation is quadratic in theory. More importantly, the coefficients of $Gvar \cdot n/k$ and $\hat{p}'(1-\hat{p}')$ or *MSE* is considerable. Thus generalization error of a model can be reflected by *Nvar* in the form of $Gvar \cdot n/k$ to some extent.

The definition of *Nvar* indicates that model error is the only factor which is relevant to *Nvar*. Theoretically, *Nvar* is independent of fold number $k$. Here, one-way analysis of variance (ANOVA) is employed to check whether *Nvar* ($Gvar \cdot n/k$) is sensitive to $k$ in real problems. And the results are listed in Table 5.

In Table 5, $p$-value$<$ 0.05 indicates that *Nvar* values with different $k$ values have significant difference. 49 $p$-values in 50 regression problems (11th–20th datasets) and 56 $p$-values in 60 SVM models (3rd–5th columns) are more than 0.05. Values less than 0.05 mainly appear on classification data sets with NNs.

The misclassification rates of each model on ten classification sets are plotted in Fig. 9. It can be seen that the misclassification rates are around 0.2 for SVMs. While they are about 0.45 for

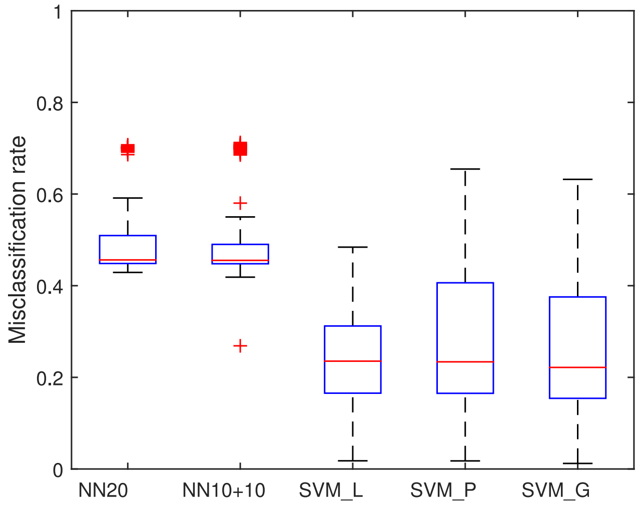


**Fig. 9.** Misclassification rate on different models.

**Table 5**
*p*-values of ANOVA.

| Dataset | NN20 | NN10+10 | SVM_L | SVM_P | SVM_G |
|---|---|---|---|---|---|
| 1 | 0.001[a] | 0.000[a] | 0.353 | 0.217 | 0.115 |
| 2 | 0.000[a] | 0.000[a] | 0.277 | 0.100 | 0.898 |
| 3 | 0.062 | 0.003[a] | 0.997 | 0.102 | 0.077 |
| 4 | 0.000[a] | 0.000[a] | 0.959 | 0.025[a] | 0.633 |
| 5 | 0.000[a] | 0.000[a] | 0.691 | 0.574 | 0.304 |
| 6 | 0.000[a] | 0.000[a] | 0.539 | 0.049[a] | 0.677 |
| 7 | 0.000[a] | 0.000[a] | 0.659 | 0.200 | 0.998 |
| 8 | 0.000[a] | 0.000[a] | 0.101 | 0.179 | 0.254 |
| 9 | 0.000[a] | 0.000[a] | 0.018[a] | 0.246 | 0.124 |
| 10 | 0.000[a] | 0.000[a] | 0.425 | 0.388 | 0.259 |
| 11 | 0.269 | 0.602 | 0.770 | 0.442 | 0.845 |
| 12 | 0.180 | 0.439 | 0.905 | 0.632 | 0.613 |
| 13 | 0.514 | 0.452 | 0.988 | 0.949 | 0.993 |
| 14 | 0.102 | 0.357 | 0.543 | 0.333 | 0.729 |
| 15 | 0.188 | 0.245 | 0.724 | 0.625 | 0.809 |
| 16 | 0.250 | 0.239 | 0.908 | 0.037[a] | 0.974 |
| 17 | 0.065 | 0.240 | 0.429 | 0.922 | 0.707 |
| 18 | 0.054 | 0.126 | 0.079 | 0.074 | 0.220 |
| 19 | 0.610 | 0.239 | 0.721 | 0.889 | 0.918 |
| 20 | 0.058 | 0.105 | 0.999 | 0.999 | 0.962 |

[a] $p$-value$<$ 0.05

NNs. The reason why $p$-value$<$ 0.05 mainly appears on classification data sets with NNs may be that NNs models are badly specified in classification experiments.

On one hand, *Nvar* is relevant to model error and independent of $k$ from its definition. On the other hand, *Nvar* ($Gvar \cdot n/k$) is significantly correlated with error indicator ($\hat{p}'(1-\hat{p}')$ or *MSE*), and *Nvar* values have no significant difference among different fold numbers when models are not badly specified. Therefore, *Nvar* is a useful error indicator in model evaluation by $k$-fold CV.

## 5. Conclusions

Classification and regression are two of the most important tasks in machine learning. Although there are kinds of good models, the parameters or model settings have a strong impact on performance when solving real problems. Models are usually selected or evaluated by the prediction error. Thus, error estimation has been promoted using variance analysis of $k$-fold CV, which can also provide a novel way for guiding model selection. The contributions of this paper are as follows: (1) When the numbers of samples and folds are both large enough, we proved that CV variance and its accuracy (*EAD*) have the quadratic relationship, allowing the accuracy to be improved quantitatively by reducing the variance. (2) The inherent relationships between CV variance and its key factors have been derived, thereby it is feasible and reliable to predict which variance is less before applying $k$-fold CV. Theoretical explanations have been given for some empirical evidence of Rodriguez and Kohavi from the respect of variance analysis. (3) The bias of CV error is generally related to $k$, while the proposed normalized variance has significant correlation with the error and is unrelated to $k$ so that it can serve as a stable error measurement in classification and regression. Moreover, our theorems support the fact that large data size and repetition are effective ways for $k$-fold CV to improve the accuracy of error estimation from the perspective of variance analysis.

Although the two ways are helpful for error estimation, we cannot ignore the issue of computational efficiency, especially for big data. Finding effective data partitions or sampling methods during the CV procedure might be a means of reducing the computational cost, and will therefore be a focus of our future work. Moreover, the robustness of the proposed approaches and how to promote the accuracy of error estimation when the distribution of the group mean is severely nonnormal are worth further exploration.

## Acknowledgments

The work described in this paper was partially supported by the National Natural Science Foundation of China (Nos. 61673249,61503229), Research Project Supported by Shanxi Scholarship Council of China (No. 2016-004), and Graduate Innovation Project of Shanxi Province (No. 215548901016).

## References

[1] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of k-fold cross–validation, J. Mach. Learn. Res. 5 (2004) 1089–1105.

[2] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.

[3] A. Zollanvari, U.M. Braga-Neto, E.R. Dougherty, On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers, Pattern Recognit. 42 (11) (2009) 2705–2723.

[4] P. Refaeilzadeh, L. Tang, H. Liu, “Cross-validation.” Encyclopedia of database systems, Springer US, 2009, pp. 532–538.

[5] J.D. Rodriguez, A. Perez, J.A. Lozano, A general framework for the statistical analysis of the sources of variance for classification error estimators, Pattern Recognit. 46 (3) (2013) 855–864.

[6] J. Fan, S. Guo, N. Hao, Variance estimation using refitted cross-validation in ultrahigh dimensional regression, J. R. Stat. Soc.(Stat. Methodol.) 74 (1) (2012) 37–65.

[7] H. Ishibuchi, Y. Nojima, Repeated double cross-validation for choosing a single solution in evolutionary multi-objective fuzzy classifier design, Knowl. Based Syst. 54 (2013) 22–31.

[8] B. Efron, R.J. Tibshirani, An introduction to the bootstrap, CRC press, 1994, pp. 45–55.

[9] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proc. 14th Int. Joint Conf. on AI, Morgan-Kaufmann, 1995, pp. 1137–1145.

[10] J. Kim, Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap, Comput. Stat. Data Anal. 53 (11) (2009) 3735–3745.

[11] D. Krstajic, L.J. Buturovic, D.E. Leahy, et al., Cross-validation pitfalls when selecting and assessing regression and classification models, J. Cheminform. 6 (1) (2014) 10–24.

[12] H. Huttunen, J. Tohka, Model selection for linear classifiers using bayesian error estimation, Pattern Recognit. 48 (11) (2015) 3739–3748.

[13] T.T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognit. 48 (9) (2015) 2839–2846.

[14] T.T. Wong, Parametric methods for comparing the performance of two classification algorithms evaluated by k-fold cross validation on multiple data sets, Pattern Recognit. 65 (2017) 97–107.

[15] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. 10 (7) (1998) 1895–1923.

[16] E. Alpaydin, Combined 5 × 2 cv f test for comparing supervised classification learning algorithms, Neural Comput. 11 (8) (1999) 1885–1892.

[17] C. Bergmeir, J.M. Benintez, On the use of cross-validation for time series predictor evaluation, Inf. Sci. 191 (2012) 192–213.

[18] M. Markatou, H. Tian, S. Biswas, et al., Analysis of variance of cross-validation estimators of the generalization error, J. Mach. Learn. Res. 6 (2005) 1127–1168.

[19] C. Nadeau, Y. Bengio, Inference for the generalization error, Mach. Learn. 52 (3) (2003) 239–281.

[20] J.D. Rodriguez, A. Perez, J.A. Lozano, Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2010) 569–575.

[21] J.G. Moreno-Torres, J.A. Sez, F. Herrera, Study on the impact of partition induced dataset shift on k-fold cross-validation, IEEE Trans. Neural Netw. Learn. Syst. 23 (8) (2012) 1304–1312.

[22] R. Kohavi, Wrappers for performance enhancement and oblivious decision graphs, Computer Science Department, Stanford University, 1995 Phd thesis.

[23] W. Yu, W. Ruibo, J. Huichen, et al., Blocked 3 × 2 cross-validated *t*-test for comparing supervised classification learning algorithms, Neural Comput. 26 (1) (2014) 208–235.

[24] A. Isaksson, M. Wallman, H. Goransson, M.G. Gustafsson, Cross-validation and bootstrapping are unreliable in small sample classification, Pattern Recognit. Lett. 29 (14) (2008) 1960–1965.

[25] M.H. DeGroot, M.J. Schervish, Probability and Statistics, fourth ed., Addison Wesley, 2011, pp. 337–341.

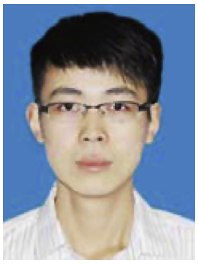
[26] M. Lichman, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, <http://archive.ics.uci.edu/ml/>, 2014-12-8.

[27] D.C. Montgomery, G.C. Runger, Applied statistics and probability for engineers, John Wiley and Sons, 2010, pp. 238–242.

[28] S. An, W. Liu, S. Venkatesh, Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression, Pattern Recognit. 40 (8) (2007) 2154–2162.

[29] Z. Shao, M.J. Er, Efficient leave-one-out cross-validation-based regularized extreme learning machine, Neurocomputing 194 (2016) 260–270.

[30] C.-C. Chang, C.-J. Lin, LIBSVM data: classification, regression, and multilabel, <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets>, 2015-11-26.



**Gaoxia Jiang**, is a lecturer in the School of Computer and Information Technology in Shanxi University. He received his M.S. degree from North China Electric Power University in 2012. His research interests include machine learning and data mining.

**Wenjian Wang**, received the B.S. degree in computer science from Shanxi Unibersity, China, in 1990, the M.S. degree in computer science from Hebei Polytechnic University, China, in 1993, and Ph.D. degree in applied mathematics from Xi'an Jiao Tong University, China, in 2004. She worked as a research assistant at the Department of Building and Construction, The City University of Hong Kong from May 2001 to May 2002. She has been with the Department of Computer Science at Shanxi University since 1993, where she was promoted as Associate Professor in 2000 and as Full Professor in 2004, and now serves as a Ph.D. supervisor in Computer Application Technology and System Engineering. She has published more than100 academic papers on machine learning, computational intelligence, and data mining. Her current research interests include machine learning theory, data mining etc.