IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON BIG DATA LEARNING AND DISCOVERY

Received December 23, 2018, accepted January 5, 2019, date of publication January 11, 2019, date of current version March 29, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2892062*

# On Estimating Model in Feature Selection With Cross-Validation

**CHUNXIA QI, JIANDONG DIAO, AND LIKE QIU**
Shandong Foreign Trade Vocational College, Qingdao 266100, China

Corresponding author: Like Qiu (qllike@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61379127, Grant 61379128, and Grant 61572448, in part by the Shandong Vocational Education Reform Project under Grant 2017121, and in part by the Project of Shandong Province Higher Educational Science and Technology Program under Grant J17RB149.

**ABSTRACT** Both wrapper and hybrid methods in feature selection need the intervention of learning algorithm to train parameters. The preset parameters and dataset are used to construct several sub-optimal models, from which the final model is selected. The question is how to evaluate the performance of these sub-optimal models? What are the effects of different evaluation methods of sub-optimal model on the result of feature selection? Aiming at the evaluation problem of predictive models in feature selection, we chose a hybrid feature selection algorithm, FDHSFFS, and conducted comparative experiments on four UCI datasets with large differences in feature dimension and sample size by using five different cross-validation (CV) methods. The experimental results show that in the process of feature selection, twofold CV and leave-one-out-CV are more suitable for the model evaluation of low-dimensional and small sample datasets, tenfold nested CV and tenfold CV are more suitable for the model evaluation of high-dimensional datasets; tenfold nested CV is close to the unbiased estimation, and different optimal models may choose the same approximate optimal feature subset.

**INDEX TERMS** Feature selection, cross-validation, nested cross-validation, wrapper, hybrid.

## I. INTRODUCTION

With the increase of data volume, the noise and redundancy in data are also increasing. Feature selection has become an important preprocessing step in data mining because of its good ability to remove noise and redundancy. It has been widely used in pattern recognition [1], data mining [2], machine learning [3], information retrieval [4] and recommendation [5], [6].

Model evaluation is an unavoidable topic in feature selection. Cross-validation(CV) is the most commonly used method for model evaluation in feature selection. Suppose that there are $m$ samples in the dataset used to build the model and they are usually divided into two parts, training set $m_{tr}$ and test set $m_{te} = m - m_{tr}$. The error produced in the training process is often called training error or CV error. The error produced in the test process is often called test error or generalization error, which refers to the error in new samples. $m_{tr}$ is used to select features and optimal model. The selection of the optimal model is usually based on the training error, that is, the model with the minimum training error is the optimal model. Once the final model is determined, $m_{te}$ is used to evaluate the performance of the model. Our goal is to obtain the model with the smallest generalization error.

Strictly speaking, we should use all the $m$ samples for model selection, not some of them. If only the training set is used for model selection, there are $\binom{m}{m_{tr}}$ different ways to divide the dataset. CV is used to calculate the average predictive power of all (partial) dataset partitioning methods, and then the model with the best average predictive power is selected.

Obviously, the computation load of this method is very huge. Especially for the wrapper methods, which use the prediction accuracy of the learning algorithms as the criterion of feature selection. Every time we select the feature, we call the learning algorithm to calculate the CV error. With the increase of feature dimension and the number of samples, the amount of calculation will increase dramatically. In addition, in the process of determining the optimal model, the CV error and test error are used to determine the optimal model simultaneously in fact. This makes the test set become one part of the training process and the true generalization error is not estimated [7].

Therefore, is it reasonable for CV to evaluate the performance of sub-optimal models for a certain task? How do different CV evaluation methods affect the selection of approximate optimal feature subset? Although there are a lot of research results [8], [9] about the model evaluation problem and they have been applied to practical tasks [10]–[13]. However, as there are few studies on the model evaluation of feature selection, these problems are still relatively vague, and further study is needed.

A method of nested CV($CV_{nest}$) was presented in literature [14]. 10-fold CV is used in the inner layer to determine the optimal model. LOOCV (Leave one out of cross validation, LOO) is used in the outer layer to estimate the generalization error. Simulation results on the datasets show that $CV_{nest}$ can provide a nearly unbiased generalization error estimation. However, a large number of studies [30], [31], [34]–[36] have proved that LOO usually produce high variance, and the model evaluation effect is not as good as that of 10-fold CV [23].

Therefore, in order to solve the problem of model evaluation in feature selection, this paper improves the $CV_{nest}$ method. 10-fold CV is used in the inner layer to determine the optimal model and used in the outer layer to estimate the generalization error. The improved $CV_{nest}$ method is named as $CV_{nest}(10, 10)$. In order to improve the computational efficiency, we embed $CV_{nest}(10, 10)$ into a hybrid feature selection method, FDHSFFS [15] algorithm. The reason why we choose FDHSFFS is that it is a popular hybrid method at present, and can improve the calculation performance while guaranteeing the prediction accuracy. The error estimation results of various CV methods are compared in the feature selection process of FDHSFFS. The polynomial fitting method is applied to construct the model. The comparative experiments are carried out on four UCI datasets with different feature dimension and sample number. The results show that 10 times of 2-fold CV and LOO are more suitable for low-dimensional data. In addition, $CV_{nest}(10, 10)$ and 10 times of 10-fold CV are more suitable for high-dimensional data, etc.

The rest of this paper is organized as follows. Related work is introduced in section II, the design of the feature selection process of FDHSFFS embedded with various CV methods is detailed in section III, the experiments and results analysis are presented in section IV, and the conclusion of this paper is given in section V.

## II. RELATED RESEARCH

When designing hybrid and wrapper feature selection methods, the problem of error estimation must be considered. The accuracy of error estimation directly affects the result of feature selection and the choice of optimal model. The commonly used error estimation methods include AIC(Akaike Information Criterion) [7], [8], $C_p$ [9], jackknife [10], hold-out, bootstrap [11]–[13] and CV.

The commonly used error estimation method in feature selection is $k$-fold CV. The value of $k$ is usually set as 10, which is called 10-fold CV. Other commonly used values of $k$ are 2,5 and so on. Since Mosier [24] first proposed the CV method in 1951, it has attracted wide attention from researchers, especially LOO. Researchers have carried out a lot of theoretical and experimental researches about it [25]–[28].

In literature [21], LOO, several variants of self-help method and some other methods were compared through five items sampling experiments, trying to find a suitable error estimation method for small sample datasets. Experimental results show that LOO obtains almost unbiased error estimation, but it is usually accompanied by unacceptably high variance, especially on small sample datasets. And the self-help method of 0.632 performs better. In literature [29], CV were used for decision tree pruning and 10-fold CV was selected. Experiments show that 10-fold CV can always select the right decision tree. In literature [30], $\epsilon 0$ self-help method was compared with LOO with the nearest neighbor method as the classifier on artificial dataset, and it claims that the confidence interval estimated by self-help method is less than that of LOO. On the basic of literature [30], literature [30] compared hierarchical CV with two self-help methods on the nearest neighbor classifier. The results show that compared with LOO, hierarchical 2-fold CV has relatively low variance and better performance.

In literature [32], a feature subset selection experiment for regression was conducted. In the experiment, LOO, $k$-fold CV, hierarchical $k$-fold CV, self-help deviation correction and local CV were compared on two artificial datasets with 60 and 160 samples respectively. The experimental results show that: (1) LOO has lower deviation and root mean square error, 2-fold CV and 5-fold CV have higher deviation and root mean square error only for model selection with many features; (2) 10-fold CV has significantly lower performance on small sample datasets; (3) 10-fold CV performs better than LOO when used for model selection.

In literature [33], experiments were conducted using real data to verify the effectiveness of CV on the pruning of decision tree. The results show that 10-fold CV can generate unbiased tree on datasets with at least 200 samples.

In literature [34], experiments were carried out on low-dimensional synthetic and real datasets. In the experiments, CV, bootstrap and resubstitution methods were compared and analyzed. The results show that among all these CV methods, 10 times of 10-fold CV has the best performance, but its computational complexity is too large and other CV methods have too high variance and a large number of outliers; bootstrap method, especially that for 0.632 estimation, has the best performance, but its computational cost is too high.

Literature [35] validated the performance of CV and repeated CV on several datasets. The experiment results show that the average results of 10 times of CV and 30 times of CV are not better than that of single CV, but their computational burden is larger than that of single CV.

CV is also widely used in feature selection. For example, in literature [36], 5-fold CV were applied to improve the generalization performance of SVM model and guide the removal of irrelevant and redundant features in brain-computer interface, good application results are obtained. In literature [37], CV was applied to guide the feature selection to improve the performance of target detection in pedestrian detection. Literature [38] proposed a feature selection algorithm for the product quality monitoring in the production process. It uses 10-fold CV to reduce the generalization error, guide the selection of features with strong correlation with product quality, and achieved good results. Literature [39] proposed a new method to reliably estimate the prediction accuracy and select the most predictive features in a high-dimensional survival prediction setting. To avoid overfitting while selecting features with high predictive power, the proposed approach estimates accuracy and performs feature selection using repeated nested CV with novel feature combination heuristics. This combination of aggregating CV runs by weighting results in sparser feature selection with more accurate estimation of predictive power. Literature [40] proposed a unified feature selection framework to reduce the dimension of image/video data, which can be applied to both supervised and semi-supervised scenarios. It used a 10-fold CV to optimize the parameters. Literature [41] proposed a novel semi-supervised feature selection framework by mining correlations among multiple tasks and apply it to different multimedia applications. It uses a 5-fold CV to learn the optimal parameters and achieve good results. Literature [42] proposed a Convex Sparse Principal Component Analysis (CSPCA) algorithm and applied it to feature selection task, a 5-fold CV is used and the experimental results showed that the CSPCA outperformed the other state-of the-art unsupervised feature selection.

In summary, although most researchers choose the CV method as the error estimation method in the feature selection process, there is few application and theoretical research of CV method in feature selection, and less related research in hybrid feature selection. So we choose the hybrid feature selection method FDHSFFS to verify the performance of various CV methods. This paper aims to provide a reference for the selection of error estimation method in feature selection process.

## III. DESIGN OF THE ALGORITHM

FDHSFFS is a kind of hybrid feature selection algorithm in which the filter method is used to select candidate features and the prediction accuracy of the learning algorithm to verify the performance of the candidate features. Therefore, the result of error estimation determines the selection of approximate optimal feature subset and optimal model.

In this section, the feature selection process of FDHSFFS is designed in which various of CV methods are embedded. Specifically, two feature selection processes of FDHSFFS are detailed in which repeated CV and $CV_{nest}(10, 10)$ are embedded respectively.

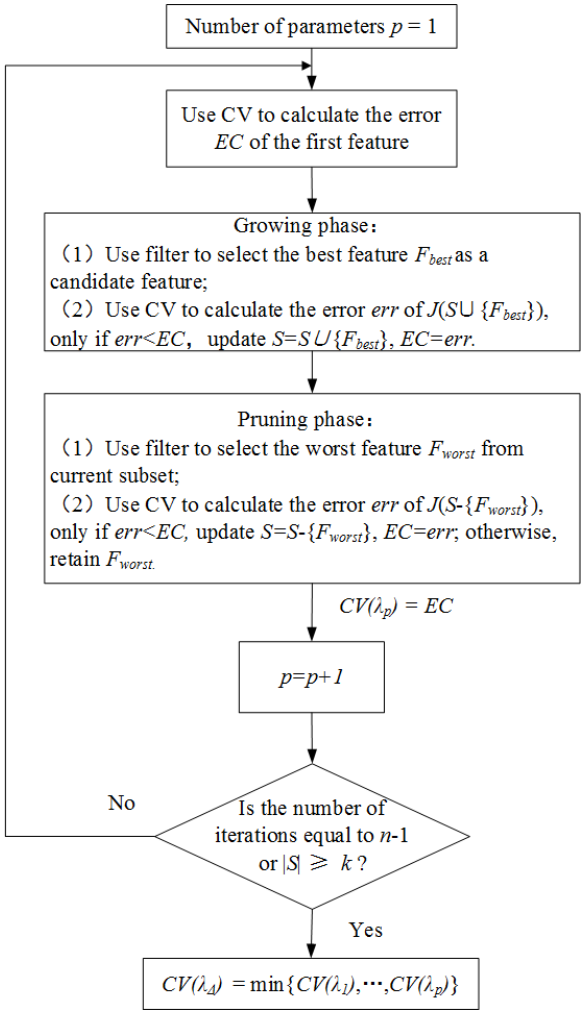


**FIGURE 1.** The model evaluation flow chart of FDHSFFS embedded with repeated CV.

### A. FEATURE SELECTION PROCESS EMBEDDED WITH REPEATED CV

The feature selection process of FDHSFFS is shown in Fig. 1 in which repeated CV is embedded. The parameters to be optimized in FDHSFFS are collectively denoted by $\lambda$ (possibly one or more parameters, they are all denoted by $\lambda$ uniformly). We train different models according to different values of $\lambda$, but it is not feasible to train models for all parameter configurations. Therefore, we can only select a range and a change step for parameter $\lambda$. For example, within the range of [0, 1], 0.1 is set as the step size, then there are 11 candidate parameter values to be evaluated. The final selected value is chosen from these candidate values. Assume that there are $p$ values for parameter $\lambda$, the $p$-th value is denoted by $\lambda_p$, $J(S)$ represents the performance of feature subset $S$, $k$ is the maximum feature dimension limited by FDHSFFS algorithm and $n$ is the dimension of original feature set.

It can be seen from Fig. 1 that FDHSFFS has two stopping conditions. One is if $|S|$ is greater than $k$, and the other is when the number of iterations equals to $n - 1$. Repeated CV

error is calculated for each addition or removal of a candidate feature. When all the $p$ parameters are tried, the test error set $\{CV(\lambda_1), \cdots, CV(\lambda_p)\}$ corresponding to the $p$ models is obtained. Then the minimum test error is denoted as $CV(\lambda_\Delta)$, $CV(\lambda_\Delta) = min\{CV(\lambda_1), \cdots, CV(\lambda_p)\}$, its corresponding parameter is denoted as $\lambda_\Delta$. The model determined by $\lambda_\Delta$ is the optimal model and the feature set given by this model is the approximate optimal feature subset.

It should be noted that the criterion of choosing the optimal model here is actually the test error of repeated CV. So the test set also participates in the process of model selection.

The computation of FDHSFFS algorithm is mainly focused on the computation of repeated CV. The more values of parameter $\lambda$, the higher the feature dimension, the more computation times of repeated CV, the greater the amount of computation is.

### *B. FEATURE SELECTION PROCESS EMBEDDED WITH $CV_{NEST}$(10, 10)*

When $CV_{nest}(10, 10)$ is used to select the model, the feature selection process of FDHSFFS is shown in Fig. 2 and Fig. 3.

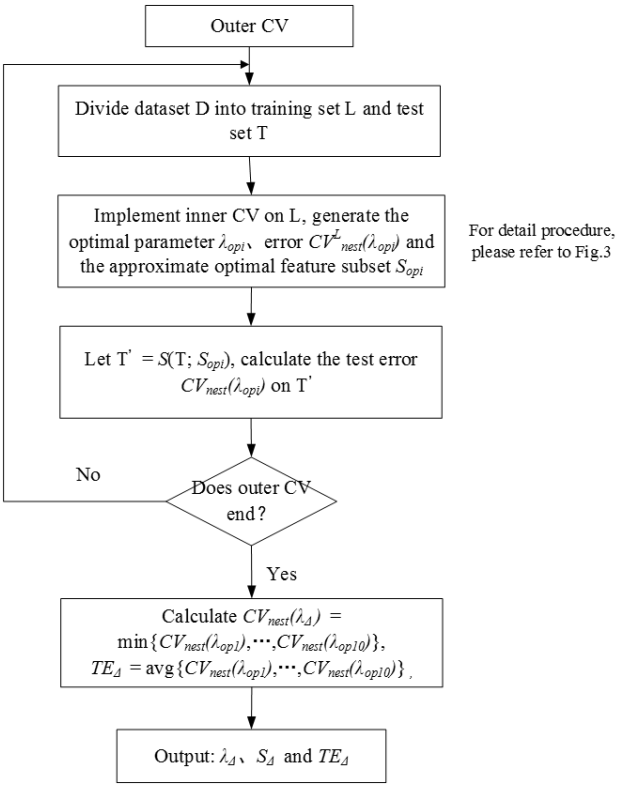


**FIGURE 2.** The outer model evaluation flow chart of FDHSFFS embedded with $CV_{nest}$(10, 10).

In Fig. 2, the model determined by $\lambda_\Delta$ is the optimal model, the feature subset $S_\Delta$ corresponding to $\lambda_\Delta$ is the approximate optimal feature subset, and the test error denoted by $TE_\Delta$, $TE_\Delta = avg\{CV_{nest}(\lambda_{op1}\}, \cdots, CV_{nest}(\lambda_{op10})\}$.

Comparing Figs. 1, 2, and 3, it can be seen that there are many differences in the evaluation process of models embedded with different CV methods.

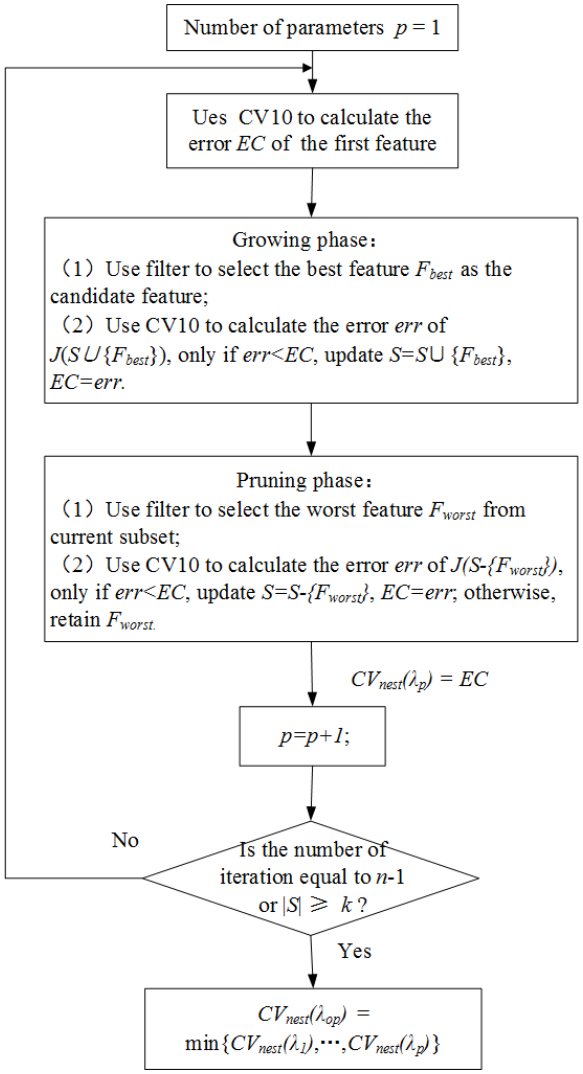


**FIGURE 3.** The inner model evaluation flow chart of FDHSFFS embedded with $CV_{nest}$(10, 10).

**(1)** The amount of calculation is different. All the repeated CV calculations in Fig. 1 are replaced by single CV calculation in Fig. 3, and the internal computation is reduced.

**(2)** The ways to determine the optimal parameters are different. In Fig. 1, $CV(\lambda_\Delta)$ is the minimum value of $p$ errors. While in Fig. 3, the inner layer of $CV_{nest}(10, 10)$ trains one model for each parameter, and produces an average CV error which denoted by $CV_{nest}(\lambda_i)(i = 1, \cdots, p)$. When all the models with different parameter settings are trained, the $CV_{nest}(\lambda_i)$ with the minimum value is chosen from the $p$ values and denoted by $CV_{nest}(\lambda_{op})$. The number of $CV_{nest}(\lambda_{op})$ is 10, which is produced by the 10 times of outer loops, and $CV_{nest}(\lambda_\Delta)$ is denoted by $CV_{nest}(\lambda_\Delta) = min_{i=1,\cdots,10} CV_{nest}(\lambda_{opi})$. The corresponding parameter $\lambda_\Delta$ is the optimal parameter and the model trained by this parameter is the optimal model.

**(3)** The ways to calculate the test error are different. In Fig. 1, we chose the minimum error from $p$ errors, it is the

test error and denoted by $CV(\lambda_\Delta)$. In Fig. 2, if the parameter $\lambda_\Delta$ is chosen, then the optimal model is determined, and the approximate optimal feature subset generated by the model is also determined. While the test error, $TE_\Delta = \frac{1}{10}\sum_{i=1}^{10} CV_{nest}(\lambda_{opi})$ is the average value of all models' errors.

## IV. SIMULATION EXPERIMENTS

In this section, we conduct a large number of experiments to compare the results of feature selection using different CV methods. We compare the estimated errors of $CV_{nest}(10, 10)$, repeated 10 times CV10 (r10CV10), repeated 10 times CV5 (r10CV5), repeated 10 times CV2 (r10CV2) and LOO in the feature selection algorithm of FDHSFFS. For r10CV$k$, only the test error $CV(\lambda_\Delta)$ is estimated. And for $CV_{nest}(10, 10)$, both the training error $CV_{nest}(\lambda_\Delta)$ and test error $TE_\Delta$ are estimated. We compare various CV methods in FDHSFFS feature selection process on four UCI datasets with significantly different feature dimensions and sample sizes. We compare the error estimates results, computational efficiency, the selected optimal model and approximate optimal feature subset of various CV methods. We chose polynomial fitting as the learning algorithm in the feature selection process.

The experiments run on a common desktop computer with Intel (R) Core (TM) i5-4690 CPU 3.5GHz, 8GB memory, Windows 7 operating system. And the simulation software is MATLAB 2016.

### A. DATASETS

The public datasets come from the UCI machine learning library [34] and the description information of them are shown in Table 1. Four datasets with different feature dimensions and sample sizes are selected elaborately, including dataset SP with low feature dimension and middle sample size, dataset GSAFM with high feature dimension and high sample size, dataset Breast Cancer Wisconsin Prognostic (BCW (P)) with low feature dimension and sample size and dataset UJIL with high dimension and sample size.

**TABLE 1. Description of datasets.**

| Datasets | Feature dimension | The number of samples |
|---|---|---|
| SP | 32 | 1044 |
| BCW(P) | 34 | 194 |
| GSAFM | 438 | 47 |
| UJIL | 528 | 6000 |

In order to speed up the calculation, 6000 samples are randomly selected from 19937 samples as the total sample space of UJIL.

### B. METHOD OF PERFORMANCE MEASUREMENT

Variance, bias and mean square error (MSE) are commonly used in regression tasks. While MSE is composed of variance and deviation and it can balance variance and deviation to a certain extent. Therefore, we use MSE as the performance measurement in the experiments.

The MSE is as follows

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\widehat{y}_i - y_i)^2. \tag{1}$$

### C. EXPERIMENT RESULTS AND ANALYSIS

To avoid over-fitting and under-fitting, regularization is done in the experiments and the cost function is as follows

$$J(\theta) = \sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \gamma\sum_{j=1}^{n}\theta_j^2. \tag{2}$$

Here, $\gamma$ is the regularization parameter, $y^{(i)}$ denotes the $i$-th real value, $h_\theta(x^{(i)})$ denotes the $i$-th prediction value.

The range of regularization parameter $\gamma$ is [0, 1]. In FDHSFFS algorithm, the dimension $k$ of the selected feature subset is limited. In the experiments, the dimension of the dataset is assumed to be $n$. In order to improve the calculation efficiency, the value of $k$ follows the following rule, $k = min\{|n * 0.2|, 10\}$, that is, the maximum value of $k$ does not exceed 10.

#### 1) RESULTS AND ANALYSIS ON DATASET SP

Fig. 4 shows a comparison of the estimated error distribution of r10CV10, r10CV5, r10CV2, LOO and $CV_{nest}(10, 10)$. The coordinate system of LOO is black and that of other methods is red. It can be seen that the estimated error distributions of the five CV methods are approximate. This is obvious especially in r10CV5 and r10CV2 as their distributions coincide completely. It is shown that the test errors of the five CV methods are approximately distributed on dataset SP, which is of low dimension (no more than 32) and medium sample size.

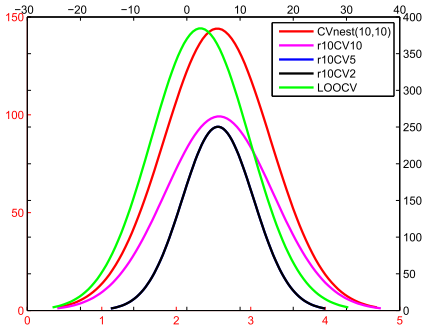


**FIGURE 4. Estimated error distributions of various CV methods on dataset SP.**

Fig. 5 shows the training error and test error distributions of $CV_{nest}(10, 10)$ in the model selection process. It can be seen that the test error is higher than its training error. The training error mainly focuses on the left side of the center point, that is, the training error is always small in most cases.

Table 2 shows the results of feature selection for various CV methods on dataset SP, the selected optimal parameters, the average training error, the average test error and the computational complexity.

**TABLE 2.** Experiment results of various CV methods on dataset SP.

| CV methods | The approximate optimal feature subsets | $\lambda_\Delta$ | Running time (s) | $CV_{nest}(\lambda_\Delta)$ | $TE_\Delta$ |
|---|---|---|---|---|---|
| r10CV10 | {32} | 1 | 242.59 | - | 2.5749 |
| r10CV5 | {32} | 0.01 | 231.17 | - | 2.5606 |
| r10CV2 | {32} | 0.6 | 209.77 | - | 2.5505 |
| Loo | {32} | 1 | 43.15 | - | 2.5508 |
| $CV_{nest}(10, 10)$ | {32} | 0.6 | 215.50 | 1.5522 | 2.5515 |

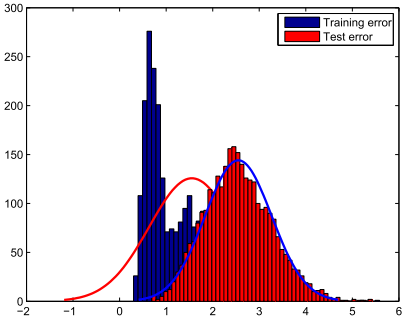


**FIGURE 5.** Distributions of the training error and test error of $CV_{nest}(10, 10)$ on dataset SP.

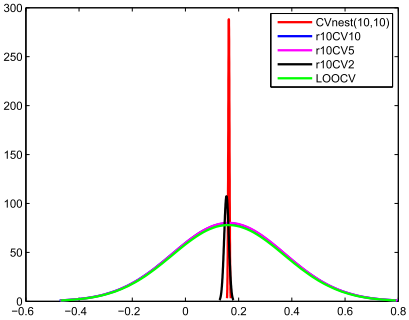


**FIGURE 6.** Estimated error distributions of various CV methods on dataset BCW(P).

As only the test set of $CV_{nest}(10, 10)$ does not participate in the optimal model selection, so two kinds of errors are estimated for $CV_{nest}(10, 10)$, average training error $CV_{nest}(\lambda_\Delta)$ and average test error $TE_\Delta$. For other CV methods, only the test error $TE_\Delta$ is estimated. This estimation method is used in other datasets.

The following conclusions can be obtained from Table 2.

***1) Approximate optimal feature subset:*** Although the optimal models determined by various CV methods are different, the same approximate optimal feature subsets are selected eventually. This shows that different optimal models may determine the same approximate optimal feature subset in the process of feature selection. This is because the criterion of selecting features in FDHSFFS is to select the feature set that can reduce the prediction error. Although the estimation errors of different optimal models may be different, the same approximate optimal feature subset may be produced.

***2) Error:*** The errors of the several CV methods are approximate and the differences among them are less than 0.03%. The test error of r10CV2 is the smallest, followed by that of LOO,$CV_{nest}(10, 10)$, r10CV5 and r10CV10.

***3) Amount of calculation:*** As LOO does not need to perform CV repeatedly, so the computational effort of LOO is minimal, followed by that of r10CV2, and other CV methods consume approximate time.

***4) $CV_{nest}(10, 10)$ can realize approximate unbiased estimation:*** The test error estimated by $CV_{nest}(10, 10)$ is higher than its training error, but the difference between them is less than 1%.

### 2) RESULTS AND ANALYSIS ON DATASET BCW(P)

Fig. 6 shows the comparison of the error distribution of r10CV10, r10CV5, r10CV2, LOO and $CV_{nest}(10, 10)$. It can be seen that the error distributions of r10CV10, r10CV5 and LOO are almost coincident, while the error distributions of r10CV2 and $CV_{nest}(10, 10)$ are approximate and more concentrated. The test error distributions of the five CV methods are approximate.

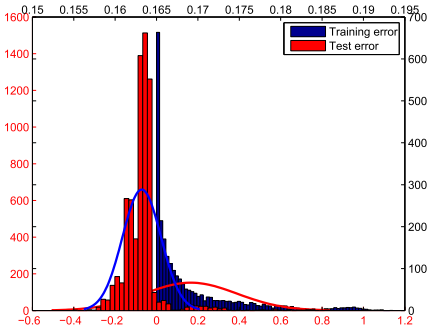


**FIGURE 7.** Distributions of the training error and test error of $CV_{nest}(10, 10)$ on dataset BCW(P).

Fig. 7 shows the distributions of training error and test error of $CV_{nest}(10, 10)$ in the process of model selection. The coordinate system of the distribution of test error is black, and that of the training error distribution is red. It can be seen that the test error of $CV_{nest}(10, 10)$ is slightly higher than its training error; the training error is mainly concentrated on the left side of the center point, that is, most of the estimated values are small; and most of the estimated values of prediction error are concentrated near the mean (center point).

Table 3 shows the results of feature selection, the selected optimal parameters, the average training error, and the running time of FDHSFFS for various CV methods on dataset BCW (P).

The following conclusions can be obtained from Table 3.

***1) Approximate optimal feature subset:*** Similar to the result of feature selection on dataset SP, all CV methods

**TABLE 3.** Experiment results of various CV methods on dataset BCW(P).

| CV methods | The approximate optimal feature subsets | $\lambda_\Delta$ | Running time (s) | $CV_{nest}(\lambda_\Delta)$ | $TE_\Delta$ |
|---|---|---|---|---|---|
| r10CV10 | {25,34} | 0.1 | 43.87 | - | 0.1616 |
| r10CV5 | {25,34} | 1 | 43.78 | - | 0.1625 |
| r10CV2 | {25,34} | 1 | 43.29 | - | 0.1543 |
| Loo | {25,34} | 1 | 5.14 | - | 0.1610 |
| $CV_{nest}(10, 10)$ | {25,34} | 1 | 40.02 | 0.1653 | 0.1672 |

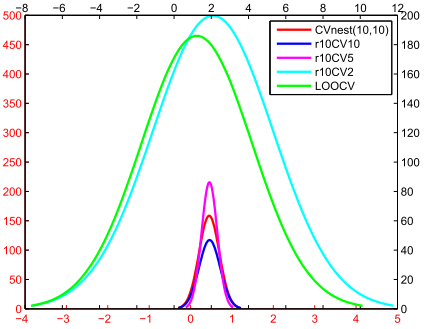


**FIGURE 8.** Estimated error distributions of various CV methods on dataset GSAFM.

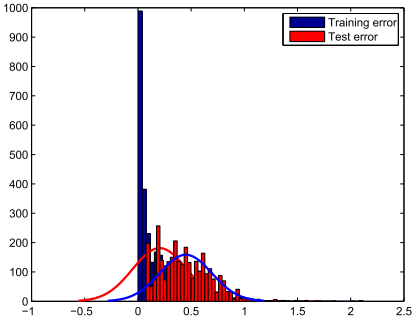


**FIGURE 9.** Distributions of the training error and test error of $CV_{nest}(10, 10)$ on dataset GSAFM.

choose the same approximate optimal feature subset. It is further verified that the same approximate optimal feature subset may be generated by different optimal models.

***2) Error:*** Similar to the results on dataset SP, the test error of r10CV2 is the lowest, followed by that of LOO, r10CV10 and r10CV5, and the test error of $CV_{nest}(10, 10)$ is the highest.

***3) Amount of calculation:*** LOO takes the least time to determine the approximate optimal feature subset and optimal model, followed by $CV_{nest}(10, 10)$. And the amount of calculation of other CV methods is approximate.

***4)*** $CV_{nest}(10, 10)$ ***can realize approximate unbiased estimation:*** The test error of $CV_{nest}(10, 10)$ is slightly higher than its training error, and the difference is less than 0.005%, this closes to unbiased estimation.

### 3) RESULTS AND ANALYSIS ON DATASET GSAFM

Fig. 8 shows the comparison of the error distributions of r10CV10, r10CV5, r10CV2, LOO and $CV_{nest}(10, 10)$. The coordinate system of the error estimation distribution of LOO is black, and that of other CV methods is red. It can be seen that the errors of r10CV10, r10CV5 and $CV_{nest}(10, 10)$ are more concentrated.

Fig. 9 shows the distributions of the training error and test error of $CV_{nest}(10, 10)$ in the process of model selection. It can be seen that its test error is slightly higher than its training error. The training error is mainly concentrated on the left side of the center point, that is, most of the estimated values are small.

Table 4 shows the results of feature selection, the selected optimal parameters, the average training error, and the amount of calculation of FDHSFFS for various CV methods on dataset GSAFM.

The following conclusions can be obtained from Table 4.

***1) Approximate optimal feature subset:*** The feature selection results in Table 4 further verify that different models may determine the same approximate optimal feature subsets in the feature selection process.

***2) Error:*** The test error of r10CV10 is the lowest, followed by that of $CV_{nest}(10, 10)$, r10CV5 and r10CV2, and that of LOO method is significantly higher than other CV methods.

***3) Amount of calculation:*** With the rapid growth of feature dimension (from 32 in SP to 438 in GASFM), the amount of calculation of $CV_{nest}(10, 10)$ is obviously higher than that of other methods, and its computational efficiency drops sharply. LOO still takes the least time to determine the approximate optimal feature subset and the optimal model, followed by r10CV2.

***4)*** $CV_{nest}(10, 10)$ ***can realize approximate unbiased estimation:*** The test error of $CV_{nest}(10, 10)$ is slightly higher than its training error and the difference is only 0.25%, this is close to unbiased estimation.

### 4) EXPERIMENT RESULTS AND ANALYSIS ON DATASET UJIL

Fig. 10 shows the estimated error distributions of r10CV10, r10CV5, r10CV2, LOO and $CV_{nest}(10, 10)$. The coordinate of LOO is black and that of other CV methods is red. It can be seen that the errors of various CV methods vary greatly, and the error distribution of $CV_{nest}(10, 10)$ is the most concentrated, followed by that of r10CV2.

Fig. 11 shows the distributions of the training error and test error of $CV_{nest}(10, 10)$ in the model selection process. It can be seen that the test error of $CV_{nest}(10, 10)$ is obviously higher than its training error, and the distributions of training error and test error are more uniform and they mainly concentrate in the vicinity of the mean value.

**TABLE 4.** Experiment results of various CV methods on dataset GSAFM.

| CV methods | The approximate optimal feature subsets | $\lambda_\Delta$ | Running time (s) | $CV_{nest}(\lambda_\Delta)$ | $TE_\Delta$ |
|---|---|---|---|---|---|
| r10CV10 | {6,5} | 0.06 | 135.59 | - | 0.4459 |
| r10CV5 | {6,5} | 0.006 | 135.08 | - | 0.4560 |
| r10CV2 | {6,5} | 0.006 | 126.46 | - | 0.5419 |
| Loo | {6} | 1 | 36.62 | - | 1.2222 |
| $CV_{nest}(10, 10)$ | {6,5} | 0.06 | 385.35 | 0.2039 | 0.4500 |

**TABLE 5.** Experiment results of various CV methods on dataset UJIL.

| CV methods | The approximate optimal feature subsets | $\lambda_\Delta$ | Running time (s) | $CV_{nest}(\lambda_\Delta)$ | $TE_\Delta$ |
|---|---|---|---|---|---|
| r10CV10 | {523,121,156,162,167,13,51,40,268,33} | 0.003 | 24656.61 | - | 856.13 |
| r10CV5 | {523,121, 156,162, 167,13,51,40,268} | 0.003 | 18284.74 | - | 863.30 |
| r10CV2 | {523,156,155,162,120, 51,47, 268,33,40} | 1 | 14263.94 | - | 867.78 |
| Loo | {3,523, 156,155, 14,42,41,51,52,40} | 0.006 | 1663.84 | - | 1141.30 |
| $CV_{nest}(10, 10)$ | {523,156,524,526,14,162,155,120,51,39} | 0.1 | 17933.02 | 828.61 | 831.58 |

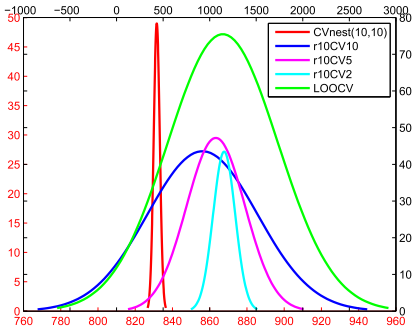


**FIGURE 10.** Estimated error distributions of various CV methods on dataset UJIL.

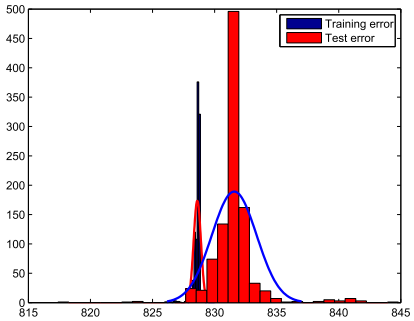


**FIGURE 11.** Distributions of the training error and test error of $CV_{nest}(10, 10)$ on dataset UJIL.

Table 5 shows the results of feature selection, the selected optimal parameters, the average training error, and the amount of calculation for determining the optimal model for various CV methods on dataset UJIL.

The following conclusions can be obtained from Table 5.

***1) Approximate optimal feature subset:*** The approximate optimal feature subsets determined by various CV methods are different, but the similarity of the selected features is very high. Especially for r10CV10 and r10CV5, their selected approximate optimal feature subsets are almost the same.

***2) Error:*** On the high-dimensional (greater than 438) datasets GSAFM and UJIL, the test errors of LOO are obviously higher than that of other CV methods, mainly because its variance is too high, and the errors of other CV methods are similar.

***3) Amount of calculation:*** The amount of calculation of r10CV10 is the largest, followed by that of r10CV5, $CV_{nest}(10, 10)$ and r10CV2, and that of LOO is still the smallest.

***4) The errors of the*** $CV_{nest}(10, 10)$***:*** The test error of $CV_{nest}(10, 10)$ is higher than its training error and the difference is 2.97%.

## V. CONCLUSION

Aiming at the error estimation in the feature selection of supervised regression task, $CV_{nest}(10, 10)$, r10CV10, r10CV5, r10CV2 and LOO are selected to estimate the errors in the feature selection process of FDHSFFS, and comparative experiments are carried on four UCI datasets with large differences in feature dimension and sample size. The following conclusions are obtained from the results.

(1) On low-dimensional(less than 32) datasets, the test errors of r10CV2 and LOO are less than that of other CV methods, and their amount of calculation of feature selection is also less than that of $CV_{nest}(10, 10)$, r10CV10 and r10CV5. So it is recommended to apply r10CV2 or LOO to estimate the errors for feature selection on low-dimensional (less than 32) datasets with small sample size.

(2) On high-dimensional (more than 438) datasets, the test errors of r10CV2 and LOO are significantly higher than that of other CV methods. And the estimated errors of $CV_{nest}(10, 10)$ and r10CV10 are lower. However, the computational efficiency of $CV_{nest}(10, 10)$ and r10CV10 decreases significantly with the increase of feature dimension. For practical tasks, high accuracy and low computational complexity always need to be weighed. But with the development of computer hardware, high-performance computing has become popular. So, it may be considered to adopt $CV_{nest}(10, 10)$ or r10CV10 to estimate errors for high-dimensional datasets with large sample size.

(3) Different optimal models may select the same approximate optimal feature subset.

(4) $CV_{nest}(10, 10)$ can realize approximate unbiased estimation.

The experimental results of this paper provide a reference for the selection of model evaluation method in the process of feature selection.

## ACKNOWLEDGMENT

Like Qiu was with the Information Management Department, Shandong Foreign Trade Vocational College, Qingdao 266100, China.

## REFERENCES

[1] W. Liu, X. Chang, L. Chen, and Y. Yang, "Semi-supervised Bayesian attribute learning for person re-identification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, AAAI Press, 2018, pp. 7162–7169.
[2] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Trans. Nanobiosci.*, vol. 9, no. 1, pp. 31–37, Mar. 2010.
[3] L. Yu and H. Liu "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. 20th Int. Conf. Int. Conf. Mach. Learn.* Bellevue, WA, USA: AAAI Press, 2003, pp. 856–863.
[4] Z. Cheng and J. Shen, "On very large scale test collection for landmark image search benchmarking," *Signal Process.*, vol. 124, pp. 13–26, Jul. 2016.
[5] Z. Cheng and J. Shen, "On effective location-aware music recommendation," *ACM Trans. Inf. Syst.*, vol. 34, no. 2, pp. 1–32, 2016.
[6] Z. Cheng, Y. Ding, L. Zhu, and M. Kankanhalli. (2018). "Aspect-aware latent factor model: Rating prediction with ratings and reviews." [Online]. Available: https://arxiv.org/abs/1802.07938
[7] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 10, pp. 6567–6572, 2002.
[8] B. Gandek *et al.*, "Cross-validation of item selection and scoring for the SF-12 health survey in nine countries: Results from the IQOLA Project," *J. Clin. Epidemiol.*, vol. 51, no. 11, pp. 1171–1178, 1998.
[9] R. J. Hijmans, "Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model," *Ecology*, vol. 93, no. 3, pp. 679–688, 2012.
[10] L. B. Kaplan, G. J. Szybillo, and J. Jacoby, "Components of perceived risk in product purchase: A cross-validation," *J. Appl. Psychol.*, vol. 59, no. 3, pp. 287–291, 1974.
[11] G. Zhang, M. Y. Hu, B. E. Patuwo, and D. C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis," *Eur. J. Oper. Res.*, vol. 116, no. 1, pp. 16–32, 1999.
[12] N. Nguyen, P. Milanfar, and G. Golub, "Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1299–1308, Sep. 2001.
[13] M. Szász *et al.*, "Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients," *Oncotarget*, vol. 7, no. 31, pp. 49322–49333, 2016.
[14] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *Bioinformatics*, vol. 7, p. 91, Feb. 2006.
[15] J. Q. Gan, B. A. S. Hasan, and C. S. L. Tsui, "A filter-dominating hybrid sequential forward floating search method for feature subset selection in high-dimensional space," *Int. J. Mach. Learn. Cybern.*, vol. 5, no. 3, pp. 413–423, 2014.
[16] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.
[17] R. Shibata, "An optimal selection of regression variables," *Biometrika*, vol. 68, no. 1, pp. 45–54, 1981.
[18] C. L. Mallows, "Some comments on Cp," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.
[19] J. Shao and C. F. J. Wu, "A general theory for jackknife variance estimation," *Ann. Statist.*, vol. 17, no. 3, pp. 1176–1197, 1989.
[20] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.
[21] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *J. Amer. Stat. Assoc.*, vol. 78, no. 382, pp. 316–331, 1983.
[22] B. Efron and R. Tibshirani, "Improvements on cross-validation: The 632+ bootstrap method," *J. Amer. Stat. Assoc.*, vol. 92, no. 438, pp. 548–560, 1997.
[23] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.* Burlington, MA, USA: Morgan Kaufmann Publishers, 2001, pp. 1137–1143.
[24] C. I. Mosier, "Problems and designs of cross-validation 1," *Educ. Psychol. Meas.*, vol. 11, no. 1, pp. 5–11, 1951.
[25] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *Technometrics*, vol. 16, no. 1, pp. 125–127, 1974.
[26] S. Geisser, "The predictive sample reuse method with applications," *J. Amer. Stat. Assoc.*, vol. 70, no. 350, pp. 320–328, 1975.
[27] G. Wahba and S. Wold, "A completely automatic french curve: fitting spline functions by cross validation," *Commun. Statist.*, vol. 4, no. 1, pp. 1–17, 2007.
[28] K.-C. Li, "Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: Discrete index set," *Ann. Statist.*, vol. 15, no. 3, pp. 958–975, 1987.
[29] L. I. Breiman *et al.*, "Classification and regression trees (CART)," Tech. Rep., 1984, p. 358, vol. 40, no. 3.
[30] A. K. Jain, R. C. Dubes, and C.-C. Chen, "Bootstrap techniques for error estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 628–633, Sep. 1987.
[31] S. M. Weiss, "Small sample error rate estimation for k-NN classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 285–289, Mar. 1991.
[32] L. Breiman and P. Spector, "Submodel selection and evaluation in regression. The X-random case," *Int. Stat. Rev.*, vol. 60, no. 3, pp. 291–319, 1992.
[33] S. M. Weiss and N. Indurkhya, "Decision tree pruning: Biased or optimal?" in *Proc. Nat. Conf. Artif. Intell.*, Seattle, WA, USA, Jul./Aug. 1994, pp. 626–632.
[34] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.
[35] G. Vanwinckelen and H. Blockeel, "On estimating model accuracy with repeated cross-validation," *PLoS ONE*, vol. 7, no. 2, pp. 995–1004, 2012.
[36] K. Tanaka, T. Kurita, F. Meyer, L. Berthouze, and T. Kawabe, "Stepwise feature selection by cross validation for eeg-based brain computer interface," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2006, pp. 4672–4677.
[37] A. Sorsa and K. Leiviskä, "Feature selection from Barkhausen noise data using genetic algorithms with cross-validation," in *Adaptive and Natural Computing Algorithms*. Berlin, Germany: Springer, 2009, pp. 213–222.
[38] C. Shao *et al.*, "Feature selection for manufacturing process monitoring using cross-validation," *J. Manuf. Syst.*, vol. 32, no. 4, pp. 550–555, Oct. 2013.
[39] M. Laimighofer, J. Krumsiek, F. Buettner, and F. J. Theis, "Unbiased prediction and feature selection in high-dimensional survival regression," *J. Comput. Biol.*, vol. 23, no. 4, p. 279, 2016.
[40] W. Liu, C. Gao, X. Chang, and Q. Wu, "Unified discriminating feature analysis for visual category recognition," *J. Vis. Commun. Image Represent.*, vol. 40, pp. 772–778, Oct. 2016.
[41] X. Chang and Y. Yang, "Semisupervised feature analysis by mining correlations among multiple tasks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2294–2305, Oct. 2017.
[42] X. Chang, F. Nie, Y. Yang, C. Zhang, and H. Huang, "Convex sparse PCA for unsupervised feature learning," *ACM Trans. Knowl. Discovery Data*, vol. 11, no. 1, pp. 1–16, 2016.
[43] *Machine Learning Repository*. Accessed: Nov. 2014. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html

**CHUNXIA QI** was born in Laiwu, China, in 1979. She received the B.S. degree in computer science and technology from the Wuhan University of Technology, Wuhan, China, in 2003.

She is currently a Professional Leader with the Information Management Department, Shandong Foreign Trade Vocational College. Her main research interests include machine learning, computer science, and e-commerce.

**JIANDONG DIAO** was born in Qingdao, China, in 1964. He received the B.S. degree in computer science and technology from Qufu Normal University, Qufu, China, in 2003, and the M.S. and Ph.D. degrees from the Ocean University of China, Qingdao, China, in 2010.

He is currently the President of the Shandong Foreign Trade Vocational College. His main research interests include e-commerce, information economics, and big data.

**LIKE QIU** was born in Jimo, China, in 1979. She received the B.S. degree in computer science and technology from the Ocean University of China, Qingdao, China, in 2006, and the M.S. and Ph.D. degrees from the Ocean University of China, in 2003 and 2017, respectively.

She is currently a Lecturer with the Shandong Foreign Trade Vocational College. Her main research interests include machine learning, sensor networks, and big data.