# Fault Diagnosis of Sewage Treatment Equipment Based on Feature Selection

Mingzhu Lou

# Fault diagnosis of sewage treatment equipment based on feature selection

MingZhu Lou

School of Information Engineering, Nanchang Institute of Technology, Jiangxi 330099, P.R. China

**Abstract** There are many factors that affect the operation state in the wastewater treatment process. Generally, the probability of failure is much less than the probability of normal operation. Fault diagnosis of wastewater treatment is a high-dimensional unbalanced data classification. In this study, we propose a feature selection-based method to improve the classification performance of wastewater treatment fault diagnosis. Two filter-based feature selection methods and one wrapper-based feature selection method were used for experiments. Three classifiers of C4.5, Naive Bayes, and RBF-SVM were used to evaluate the proposed method. Experimental results show that the proposed method can significantly improve the overall classification accuracy and AUC value on the wastewater treatment fault diagnosis dataset.

**Keywords** Fault diagnosis· Wastewater treatment · Imbalanced classification · Feature selection

## 1 Introduction

Wastewater treatment plants are a key infrastructure to build ecological civilization and improve the quality of the water environment. The wastewater treatment process is extremely complex, and there are many influencing factors, which will cause problems such as failure, normal and stable operation, and environmental pollution in the treatment process [1]. Therefore, fault diagnosis and corresponding management of wastewater treatment plants are of great importance [2].

In recent years, many scholars have researched wastewater treatment fault diagnosis and achieved some results [3~6]. In the process of sewage treatment, the data collected by sensors have the characteristics of high-dimensional and unbalanced, that is, the samples of normal data are much more than the samples of fault data [7]. The distribution of samples in different feature space is different. There are certain features that are beneficial to the classification of small categories. The main idea of our study is to select features with significant distinguishing power to improve the classification performance of unbalanced sewage treatment faults.

Feature selection methods can be divided into three categories of filter, wrapper, and embedded [8-10]. Filter methods filter out irrelevant features independent of the subsequent learning process. The filter feature selection method is universal, straightforward in principle, and fast in operation [11,12]. Wrapper methods determine the optimal

feature subset according to the evaluation result of the feature subset by using a classifier [13,14]. In the process of feature selection, the wrapper method requires classifier training and testing of candidate subsets, so the algorithm complexity is high. Embedded feature selection is the integration of the feature selection process and the classifier training process [15,16].

In this study, two filter-based feature selection methods and one wrapper-based feature selection method were used for wastewater treatment fault diagnosis. Three classifiers of C4.5 [17], Naive Bayes[18], and RBF-SVM [19] were used to evaluate the proposed method. Experimental results demonstrate that the proposed method can significantly improve the overall classification accuracy and AUC value on the wastewater treatment fault diagnosis dataset.

The remainder of the paper is organized as follows. In Section 2, we provide a brief review of existing work on imbalanced classification problems. In Section 3, we describe two types of filter feature selection algorithms and a wrapper feature selection algorithm used in this paper. Section 4 presents the experimental results and analysis of real sewage equipment processing data. Finally, section 5 concludes the paper.

## 2 Related work

Wastewater treatment fault diagnosis belongs to an imbalanced classification. Considerable work has been done to deal with the problem of unbalanced classification. At the preprocessing, this work is mainly included sampling-based methods and feature selection-based methods. In this paper, we focus on the application of the feature selection method in fault diagnosis of wastewater treatment equipment. Feature selection can effectively remove redundant features and irrelevant features in the dataset, reduce the impact of irrelevant data on the classifier, make the final generated classifier more concise and easier to understand, and effectively improve the performance of the classifier. Feature selection can be effective for some imbalanced classification since the distribution of samples in different feature space is different and some features are beneficial to the classification of small categories.

The filter method requires a criterion to evaluate correlations between features and categories. The filter method assumes that features that are more relevant to the category contribute more to the classification, so these features are preferentially selected. The wrapper approach uses the classification performance of the learning algorithm as the evaluation criteria for feature subsets. In the process of subset evaluation, the data corresponding to the feature subset to be evaluated is used as the training set training classifier, and then the cross-validation method is used to evaluate the performance of the feature subset.

In addition to feature evaluation, a key problem of feature selection technology is how to search from feature subset space. The common search strategies include global optimal search, random search, and heuristic search. Global optimal search is to determine the global optimal subset by enumerating all feature combinations. Time complexity is exponential in terms of data dimensionality for optimal search algorithms. Due to its extremely high time complexity, global optimal search is rarely used.

Heuristic search is divided into deterministic heuristic and nondeterministic heuristic algorithms. Deterministic heuristic search mainly includes sequence forward selection, sequence backward selection, and two-way selection. Sequence forward selection starts with an empty feature set, evaluates each feature individually to find the best feature, and places the feature into the feature set. The search then tries each of the remaining features to find the best single feature and places it into the feature set again. This process continues until no improvement is achieved when adding a new feature. Sequence backward selection is the opposite of the sequence forward selection. Sequence backward selection starting with the original feature set, remove the feature at a time that results in the most improvement in the evaluation index. Bidirectional selection combines sequence forward selection and sequence backward selection. In this study, we focus on a forward search rather than a backward search. The main reason for this choice is that forward selection is much more efficient than backward deletion.

## 3 Feature selection methods

For filter methods, feature selection and classification algorithms do not interfere with each other. The filter method needs a relevance measure to assess the correlation between the features and categories. In this paper, we focus on two widely used feature correlation assessment criteria: information gain and ReliefF.

### 3.1 Information gain

Information gain is one of the most widely used feature evaluation methods based on information entropy theory. Information entropy can measure the diversity of variables. The higher the value of a variable, the greater its uncertainty. Information entropy is descriptive information uncertainty, and is defined as:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

Where $D$ denotes a variable, $p_i$ represents the probability of the $ith$ event in variable $D$. $Info(D)$ is the information entropy of variable $D$. Conditional Entropy is defined as follows.

$$Info_A(D) = -\sum_{j=1}^{y} \frac{|D_j|}{|D|} \times Info(D_j) \tag{2}$$

$Info_A(D)$ is the conditional entropy of variable $D$ under a given variable $A$. Variable $D$ is needed to be divided to multiple categories according to the value of variable $A$. On the basis above, information gain $Gain(A)$ is defined as:

$$Gain(A) = Info(D) - Info_A(D) \tag{3}$$

Information gain $Gain(A)$ measures the decrease in the uncertainty of variable $D$ under given a variable $A$.

In a classification problem, *D* can be regarded as the class labels of all samples and *A* can be regarded as a feature. According to Eq (3), the significance of feature *A* can is obtained.

### 3.2 ReliefF

ReliefF is a nearest-neighbor-based feature evaluation algorithm. ReliefF algorithm is a feature weighting algorithm that assigns different weights to features according to the correlation of each feature and class. The features with a weight less than a certain threshold will be removed. The flow of ReliefF is as follows:

| |
|---|
| Input: Training dataset *D*, Sample sampling times $m$, Feature weight threshold, Number of nearest neighbor samples *k*<br>Output: The weight of each feature W(i) |
| 1 All feature scores set to 0, T is the empty set |
| 2 for $i$=1 to m do |
| 3 Randomly select a sample *R* from *D* |
| 4 Find the k nearest neighbor samples $H_j(j=1,2,...k)$ of *R* from the sample set of the same class of *R*, and find the *k* nearest neighbor samples $M_j(C)$ from each sample set of different classes; |
| 5 for *A=1* to *N*, // Calculate the weight of each feature, where *N is the number of features* |
| 6 $W(A)=W(A)-\sum_{j=1}^{k} diff(A,R,H_j)/(mk)+\sum_{C\neq Class(R)}\left[\frac{p(C)}{1-p(Class(R))}\sum_{j=1}^{k} diff(A,R,M_j(C))\right]/(mk)$ |

$diff(A,R_1,R_2)$ represents the distance between sample and sample according to feature A. $M_j(C)$ represents the jth nearest neighbor sample in class C.

According to a feature, if the distance between a sample and its nearest neighbor of a different category is greater than the distance between the sample and its nearest neighbor of the same category, the feature has a strong ability to identify the sample. ReliefF method outputs ranking scores according to the weight of each feature.

### 3.3 Wrapper evaluation

In the wrapper approach, the importance of a subset of features is evaluated using an inductive algorithm. In this study, we use a sequence forward selection search strategy for the wrapper method named WrapperEval. WrapperEval starts with the empty set of features and searches forward for the optimal feature subset from the original dataset by greedy hill-climbing augmented. The selection of the classification algorithm is not fixed. Three classical classification algorithms: C4.5, naïve Bayes, and RBF-SVM were used to evaluate feature subsets. Five-fold cross-validation was used to evaluate the accuracy of the learning scheme on a candidate feature subset. In addition, the termination condition we set *searchTermination*=5 means that the program terminates when optimal features are added five consecutive times without any improvement in classification performance. Chart flow of WrapperEval is shown in Fig. 1.

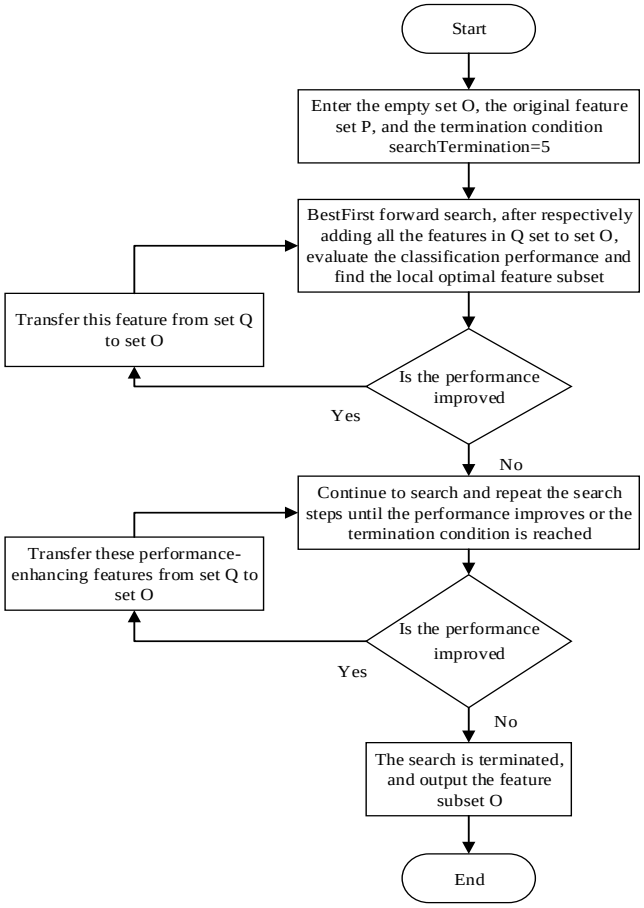


**Fig. 1.** Flow of WrapperEval

## 4. Experimental results and analysis

In this section, we evaluate the effectiveness of three feature selection algorithms for fault diagnosis of sewage treatment equipment. We first present the experimental framework, including the benchmark dataset, classification algorithms, and assessment metrics. The results and discussions are presented subsequently.

### 4.1 Water treatment plant dataset

We used the real dataset Water Treatment Plant from the UCI machine learning library [20] for experiments. The Water Treatment Plant dataset comes from the daily measures of sensors in an urban wastewater treatment plant. The objective is to classify the operational state of the plant to predict faults through the state variables of the plant at each of the stages of the treatment process. The dataset contains 527 samples, each

corresponding to one day of operational monitoring. Each sample has 38 features, including flow rate, pH value, conductivity, etc.

According to the operating status of the wastewater treatment process, all samples are divided into 13 classes. The operating states corresponding to each class are shown in Table 1. We are focused on studying the imbalanced two-classification problem. Therefore, we combined the categories of similar states to obtain 6 imbalanced datasets, namely, water1, water2, water3, water4, water5, and water6, respectively. Table 2 lists in detail the number of samples in the 6 datasets, the proportion of minority samples, and the imbalance ratio.

**Table 1.** The status of Water Treatment Plant

| ClassIndex | Status (Class) | ClassIndex | Status (Class) |
|---|---|---|---|
| 1 | Normal situation1 | 8 | Storm-1 |
| 2 | Secondary settler problems-1 | 9 | Normal situation with low influent |
| 3 | Secondary settler problems-2 | 10 | Storm-2 |
| 4 | Secondary settler problems-3 | 11 | Normal situation2 |
| 5 | Normal situation with performance over the mean | 12 | Storm-3 |
| 6 | Solids overload-1 | 13 | Solids overload-2 |

**Table 2**. Description of the datasets

| Dataset | Minority Class | Majority Class | Attributes | Instances | Minority | Majority | %Minority | Imbalance Ratio |
|---|---|---|---|---|---|---|---|---|
| Water1 | All other classes | Class of'1,5,9,11' | 38 | 527 | 14 | 513 | 2.66% | 0.03:0.97 |
| Water2 | Class of '5' | Class of'1,11' | 38 | 513 | 116 | 397 | 22.61% | 0.230.77 |
| Water3 | Class of'9' | Class of'1,11' | 38 | 513 | 69 | 444 | 13.45% | 0.13:0.87 |
| Water4 | Class of'5' | Class of'1' | 38 | 391 | 116 | 275 | 29.67% | 0.30:0.70 |
| Water5 | Class of'9' | Class of'5' | 38 | 185 | 69 | 116 | 37.30% | 0.37:0.63 |
| Water6 | Class of'5,9' | Class of'1,11' | 38 | 513 | 185 | 328 | 36.06% | 0.36:0.64 |

### 4.2 Classification algorithms

Three classical classifiers of C4.5, naïve Bayes, and RBF-SVM were used to evaluate the performance of feature selection algorithms.

C4.5 is a classical decision tree algorithm. The C4.5 classifier can generate prunedorunpruned decision trees, because unpruned decision trees may lead to overfitting of the model, so the C4.5 classification algorithm in this article utilizes pruned decision trees.

The naive Bayes classification algorithm is a probabilistic classification algorithm based on the Bayes theorem and features independence hypothesis. Equation (4) is the Bayesian formula, which is also the posterior probability, which represents the probability that the sample is classified into this class.

$$P(c_i \mid x) = \frac{P(x \mid c_i)P(c_i)}{P(x)} \tag{4}$$

$$h(x) = argmaxP(c_i \mid x) \tag{5}$$

Equation (5) indicates that the sample is regarded as the class with the largest posterior probability.

SVM is a learning machine that minimizes structural risks based on statistical learning theory. Because it can establish nonlinear decision boundaries, it has high accuracy. In our experiments, we choose the radial basis function kernel as the kernel function of the support vector machine. Compared with other kernel functions, the radial basis function kernel has higher performance and lower computational cost.

### 4.3 Assessment metric

For a two-class classification problem, in which the outcomes are labeled as either positive or negative. Given a testing dataset comprising P positive and N negative samples, the task of any classification model is to assign a class label to each sample. If the outcome of a prediction is positive and the actual value is also positive, then it is called a true positive. However, if the actual value is negative, then it is regarded as a false positive. Conversely, a true negative occurs when both the prediction outcome and the actual value are negative, and a false negative when the prediction outcome is negative while the actual value is positive. The confusion matrix is shown in Table 3.

**Table 3.** confusion matrix

| Class | Predict positive | Predict negative |
|---|---|---|
| Actual positive | TP | FN |
| Actual negative | FP | TN |

$$Accuracy = \frac{TP + TN}{P + N} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$TPR = Recall = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{TN + FP} \tag{9}$$

$$F - measure = \frac{(\beta^2 + 1)Precision * Recall}{\beta^2 (Precision + Recall)} \tag{10}$$

The accuracy represents the proportion of samples that were correctly classified. For an imbalanced dataset, the cost of misclassification of a minority class sample is higher, therefore, higher overall accuracy cannot represent better classification performance. AUC (Area Under Curve) refers to the area under the ROC characteristic curve and the horizontal and vertical axis. AUC is the area under the Receiver Operating Characteristics graph that is plotted on a two-dimensional graph, with pairs of true positive rate TPR over false positives rate FPR. AUC is a reliable performance measure; it has been widely used to evaluate classifier performance regardless of the severity of the class imbalance. Therefore, we take the AUC value as the most important evaluation index for the fault diagnosis performance of wastewater treatment. In addition, we also examine other performance indexes such as the overall classification accuracy (Accuracy).

In the experiment, a 10-fold cross-validation technique was used to measure the classification performance of the algorithm. In 10-fold cross-validation, the dataset is

equally divided into 10 subsets, one subset is taken as the test set, and the remaining nine subsets are used as the training set. Each subset is selected as the test dataset of the other 9 training subsets, and the average of the 10 test results is calculated as the final result.

### 4.3 Results and analysis

We compared four methods of InfoGain, ReliefF, WrapperEval, and Original (means no feature selection) in our experiments.

Filter methods (InfoGain and ReliefF) output the feature subsets according to the ranking score of the feature. In the experiment, we did not set a fixed threshold value but obtained 18 feature subsets within the interval with step size 2 as the unit [2]. That is, the top-2 in the ranker sequence features as the first feature subset, the top-4 features as the second feature subset, … and the top-36 features as the last feature subset. Then, three classification algorithms are adopted to evaluate the performance of all these feature subsets. The feature subset with the best performance as the optimal feature subset is then compared with other algorithms. The number of features of the resulting optimal feature subset is shown in parentheses following the accuracy. The results of the best performance of each evaluation index are shown in bold.

Tables 4-6 show the classification performance results of the two Filter type feature selection algorithms under the three classification algorithms.

As shown in Table 4 that for the C4.5 classification algorithm, the InfoGain algorithm has improved the overall classification accuracy on 5 datasets including Water2, Water3, Water4, Water5, and Water6. Among them, the InfoGain algorithm is better than the ReliefF algorithm on Water4 and Water6 datasets. AUC value is improved on all datasets. And InfoGain algorithm is better than the ReliefF algorithm of the Water1 and Water4 datasets. The ReliefF algorithm has improved the overall classification accuracy on four datasets of Water1, Water3, Water4, and Water6. Among them, it is better than the InfoGain algorithm on the three datasets of Water1, Water3, and Water5. AUC has been enhanced on all datasets. Among them, Water3, Water5, Water6, and the other three datasets are best if InfoGain algorithm. It is shown that the classification performance of the ReliefF algorithm based on the C4.5 classification algorithm is better than the compared algorithm.

One can see from Table 5 that for the NaiveBayes classification algorithm, the InfoGain algorithm has improved the overall classification accuracy on 4 datasets including Water2, Water3, Water4, and Water6. Among them, the InfoGain algorithm is preferable to the ReliefF algorithm on the Water4 dataset. AUC value is improved on all datasets. The ReliefF algorithm has improved the overall classification accuracy on four datasets of Water2, Water3, Water5, and Water6, and is better than the InfoGain algorithm. AUC value has been improved on all datasets. In the data of Water3, Water5, and Water6, the ReliefF algorithm is better than InfoGain algorithm. It is shown that the classification performance of the ReliefF algorithm under the NaiveBayesclassification algorithm is better than the comparison algorithm.

It can be seen from Table 6 that for the RBF-SVM classifier, the InfoGain algorithm has improved the overall classification accuracy on the Water2, Water4, and Water5 datasets. InfoGain algorithm is preferable to the ReliefF algorithm on the Water5

dataset.AUC value has been improved on the datasets of Water2, Water4, Water5, and Water6. Among which the datasets of Water5 and Water6, the InfoGain algorithm is better than the ReliefF algorithm. The ReliefF algorithm has improved the overall classification accuracy and AUC value on two datasets such as Water2 and Water4. However, the ReliefF algorithm is not significantly better than the InfoGain algorithm on all datasets, indicating that the InfoGain algorithm is better than the comparison algorithm under the RBF-SVM classification algorithm.

Based on the C4.5 and NaiveBayes classification algorithms, the ReliefF algorithm has better performance, and under the RBF-SVM classification algorithm, the InfoGain algorithm has better performance.

Tables 7-9 list the experimental results of the Original and the WrapperEval algorithm for three classification algorithms. One can see that the AUC values of all datasets under the three classification algorithms have improved for the WrapperEval algorithm. For the Water1, Water3, Water4, and Water6 datasets, the experimental results of the WrapperEval on the accuracy index for the three classification algorithms have improved to different degrees compared with the Original. Among the 18 groups of TP Rate comparisons, WrapperEval has 14 groups increased; the increase of TP Rate indicates that the correct classification of minority classes is improved. The experimental results show that the WrapperEval algorithm has significantly improved the classification performance.

**Table 4.** Comparison of classification performance between filter methods and Original based on C4.5 classification algorithm

| Algorithm | | Water1 | Water2 | Water3 | Water4 | Water5 | Water6 |
|---|---|---|---|---|---|---|---|
| Orginal | Accuracy | 0.9810 | 0.8460 | 0.8928 | 0.8363 | 0.7892 | 0.7817 |
| | AUC | 0.869 | 0.785 | 0.788 | 0.800 | 0.826 | 0.774 |
| InfoGain | Accuracy | 0.9829(10) | 0.8324(2) | 0.9006(2) | **0.8670**(4) | 0.8270(2) | **0.7992**(2) |
| | AUC | **0.915** | **0.851** | 0.874 | **0.870** | 0.866 | 0.8435 |
| ReliefF | Accuracy | **0.9848**(10) | 0.8324(2) | **0.9123**(4) | 0.8465(12) | **0.8595**(10) | 0.7778(4) |
| | AUC | 0.870 | **0.851** | **0.887** | 0.862 | **0.878** | **0.885** |

**Table 5.** Comparison of classification performance between filter methods and Original based on NaiveBayes classification algorithm

| Algorithm | | Water1 | Water2 | Water3 | Water4 | Water5 | Water6 |
|---|---|---|---|---|---|---|---|
| Orginal | Accuracy | 0.9658 | 0.8713 | 0.8713 | 0.8977 | 0.8595 | 0.8129 |
| | AUC | 0.920 | 0.893 | 0.923 | 0.934 | 0.921 | 0.8975 |
| InfoGain | Accuracy | 0.9658(12) | **0.8772**(10) | 0.8928(4) | **0.9028**(10) | 0.8486(36) | 0.8382(16) |
| | AUC | 0.990 | 0.921 | 0.943 | 0.958 | 0.922 | 0.923 |
| ReliefF | Accuracy | 0.9564(12) | **0.8772**(10) | **0.8967**(10) | 0.8926(8) | **0.8757**(10) | **0.8460**(12) |
| | AUC | **0.992** | **0.927** | **0.947** | **0.962** | **0.930** | **0.9375** |

**Table 6.** Comparison of classification performance between filter methods and Original based on RBF-SVM classification algorithm

| Algorithm | | Water1 | Water2 | Water3 | Water4 | Water5 | Water6 |
|---|---|---|---|---|---|---|---|
| Orginal | Accuracy | 0.9734 | 0.7739 | 0.8655 | 0.7033 | 0.6270 | 0.6394 |
| | AUC | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| InfoGain | Accuracy | 0.9734(4) | **0.7797**(2) | 0.8655(4) | **0.7877**(2) | **0.6378**(2) | 0.6257(2) |
| | AUC | 0.5 | **0.601** | 0.5 | **0.697** | **0.514** | **0.5045** |
| ReliefF | Accuracy | 0.9734(2) | **0.7797**(2) | 0.8655(4) | **0.7877**(2) | 0.6270(6) | 0.6394(4) |
| | AUC | 0.5 | **0.601** | 0.5 | **0.697** | 0.5 | 0.5 |

**Table 7.** Comparison of classification performance between WrapperEval and Original based on C4.5 classification algorithm

| Dataset | Methods | Accuracy | TP Rate | FP Rate | Precision | F-Measure | AUC |
|---|---|---|---|---|---|---|---|
| Water1 | Original | 0.9810 | 0.571 | **0.008** | 0.667 | 0.615 | 0.869 |
| | WrapperEval | **0.9829** | **0.643** | **0.008** | **0.692** | **0.667** | **0.908** |
| Water2 | Original | 0.8460 | 0.629 | 0.091 | 0.67 | 0.649 | 0.785 |
| | WrapperEval | **0.8616** | **0.647** | **0.076** | **0.714** | **0.679** | **0.853** |
| Water3 | Original | 0.8928 | **0.638** | 0.068 | 0.595 | 0.615 | 0.788 |
| | WrapperEval | **0.9181** | 0.623 | **0.036** | **0.729** | **0.672** | **0.895** |
| Water4 | Original | 0.8363 | 0.681 | 0.098 | 0.745 | 0.712 | 0.800 |
| | WrapperEval | **0.8824** | **0.750** | **0.062** | **0.837** | **0.791** | **0.912** |
| Water5 | Original | 0.7892 | 0.681 | **0.147** | 0.734 | 0.707 | 0.826 |
| | WrapperEval | **0.8432** | **0.826** | **0.147** | **0.770** | **0.797** | **0.852** |
| Water6 | Original | 0.7817 | **0.619** | 0.089 | 0.602 | 0.6105 | 0.774 |
| | WrapperEval | **0.7934** | 0.609 | **0.071** | **0.6475** | **0.6275** | **0.8335** |

**Table 8.** Comparison of classification performance between WrapperEval and Original based on NaiveBayes classification algorithm

| Dataset | Methods | Accuracy | TP Rate | FP Rate | Precision | F-Measure | AUC |
|---|---|---|---|---|---|---|---|
| Water1 | Original | 0.9658 | 0.857 | 0.031 | 0.429 | 0.571 | 0.92 |
| | WrapperEval | **0.9848** | **0.929** | **0.014** | **0.65** | **0.765** | **0.995** |
| Water2 | Original | 0.8713 | **0.776** | 0.101 | 0.692 | **0.732** | 0.893 |
| | WrapperEval | 0.8713 | 0.69 | **0.076** | **0.727** | 0.708 | **0.937** |
| Water3 | Original | 0.8713 | 0.841 | 0.124 | 0.513 | 0.637 | 0.923 |
| | WrapperEval | **0.9220** | **0.899** | **0.074** | **0.653** | **0.756** | **0.968** |
| Water4 | Original | 0.8977 | 0.862 | 0.087 | 0.806 | 0.833 | 0.934 |
| | WrapperEval | **0.9182** | **0.871** | **0.062** | **0.856** | **0.863** | **0.974** |
| Water5 | Original | **0.8595** | 0.913 | 0.172 | 0.759 | **0.829** | 0.921 |
| | WrapperEval | 0.8486 | **0.928** | **0.198** | **0.736** | 0.821 | **0.955** |
| Water6 | Original | 0.8129 | 0.761 | 0.089 | 0.6435 | 0.6855 | 0.8975 |
| | WrapperEval | **0.8674** | **0.794** | **0.0535** | **0.7555** | **0.7695** | **0.9505** |

**Table 9.** Comparison of classification performance between WrapperEval and Original based on RBF-SVM classification algorithm

| Dataset | Methods | Accuracy | TP Rate | FP Rate | Precision | F-Measure | AUC |
|---|---|---|---|---|---|---|---|
| Water1 | Original | 0.9734 | 0 | **0** | 0 | 0 | 0.5 |
| | WrapperEval | **0.9772** | **0.143** | **0** | **1** | **0.25** | **0.571** |
| Water2 | Original | 0.7739 | 0 | **0** | 0 | 0 | 0.5 |
| | WrapperEval | **0.8421** | **0.56** | 0.076 | **0.684** | **0.616** | **0.742** |
| Water3 | Original | 0.8655 | 0 | **0** | 0 | 0 | 0.5 |
| | WrapperEval | **0.9025** | **0.362** | 0.014 | **0.806** | **0.5** | **0.674** |
| Water4 | Original | 0.7033 | 0 | **0** | 0 | 0 | 0.5 |
| | WrapperEval | **0.8696** | **0.716** | 0.065 | **0.822** | **0.765** | **0.825** |
| Water5 | Original | 0.6270 | 0 | **0** | 0 | 0 | 0.5 |
| | WrapperEval | **0.7730** | **0.667** | 0.164 | **0.708** | **0.687** | **0.751** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Water6 | Original | 0.6394 | 0 | 0 | 0 | 0 | 0.5 |
| | WrapperEval | **0.7602** | **0.345** | 0.0535 | **0.336** | **0.3405** | **0.6455** |

Tables 10-12 list the experimental results of the InfoGain, ReliefF, and WrapperEval for three classification algorithms respectively.

For the C4.5 classification algorithm, the WrapperEval is better than the two filter feature selection algorithms in the accuracy and AUC value on the datasets of Water2, Water3, and Water4. For the Water5 dataset, the ReliefF algorithm obtains the best result. For the naïve Bayes classifier, the AUC value of the WrapperEval is the highest for all datasets; and accuracy is the highest on the four datasets of Water1, Water3, Water4, and Water6. Therefore, in the performance comparison of NaiveBayes and RBF-SVM classification algorithms, the WrapperEval is better than other comparison algorithms.

To further illustrate the classification performance of the WrapperEval, Figs. 2-4 provide an overall intuitive comparison of the Accuracy value; Figs. 5 -7 provide an overall intuitive comparison of the AUC value. As can be seen from the figures, the three feature selection algorithms have significantly improved compared to the Original, and the WrapperEval algorithm performs the best.

**Table 10.** Comparison of classification performance of InfoGain, ReliefF, and WrapperEval based on C4.5 classification algorithm

| Algorithm | | Water1 | Water2 | Water3 | Water4 | Water5 | Water6 |
|---|---|---|---|---|---|---|---|
| InfoGain | Accuracy | 0.9829(10) | 0.8324(2) | 0.9006(2) | 0.8670(4) | 0.8270(20) | **0.7992(6)** |
| | AUC | **0.915** | 0.851 | 0.874 | 0.870 | 0.866 | 0.8435 |
| ReliefF | Accuracy | **0.9848(10)** | 0.8324(2) | 0.9123(4) | 0.8465(12) | **0.8595(10)** | 0.7778(4) |
| | AUC | 0.870 | 0.851 | 0.887 | 0.862 | **0.878** | **0.885** |
| WrapperEval | Accuracy | 0.9829(8) | **0.8616(15)** | **0.9181(6)** | **0.8824(5)** | 0.8432(10) | 0.7934(5) |
| | AUC | 0.908 | **0.853** | **0.895** | **0.912** | 0.852 | 0.8335 |

**Table 11.** Comparison of classification performance of InfoGain, ReliefF, and WrapperEval based on NaiveBayes classification algorithm

| Algorithm | | Water1 | Water2 | Water3 | Water4 | Water5 | Water6 |
|---|---|---|---|---|---|---|---|
| InfoGain | Accuracy | 0.9658(12) | **0.8772(10)** | 0.8928(4) | 0.9028(10) | 0.8486(36) | 0.8382(16) |
| | AUC | 0.990 | 0.921 | 0.943 | 0.958 | 0.922 | 0.923 |
| ReliefF | Accuracy | 0.9564(12) | **0.8772(10)** | 0.8967(10) | 0.8926(8) | **0.8757(10)** | 0.8460(12) |
| | AUC | 0.992 | 0.927 | 0.947 | 0.962 | 0.930 | 0.9375 |
| WrapperEval | Accuracy | **0.9848(7)** | 0.8713(12) | **0.9220(11)** | **0.9182(15)** | 0.8486(18) | **0.8674(14)** |
| | AUC | **0.995** | **0.937** | **0.968** | **0.974** | **0.955** | **0.9505** |

**Table 12.** Comparison of classification performance of InfoGain, ReliefF, and WrapperEval based on RBF-SVM classification algorithm

| Algorithm | | Water1 | Water2 | Water3 | Water4 | Water5 | Water6 |
|---|---|---|---|---|---|---|---|
| InfoGain | Accuracy | 0.9734(4) | 0.7797(2) | 0.8655(4) | 0.7877(2) | 0.6378(2) | 0.6257(2) |
| | AUC | 0.5 | 0.601 | 0.5 | 0.697 | 0.514 | 0.5045 |
| ReliefF | Accuracy | 0.9734(2) | 0.7797(2) | 0.8655(4) | 0.7877(2) | 0.6270(6) | 0.6394(4) |
| | AUC | 0.5 | 0.601 | 0.5 | 0.697 | 0.5 | 0.5 |
| WrapperEval | Accuracy | **0.9772(1)** | **0.8421(8)** | **0.9025(5)** | **0.8696(7)** | **0.7730(7)** | **0.7602(9)** |
| | AUC | **0.571** | **0.742** | **0.674** | **0.825** | **0.751** | **0.6455** |

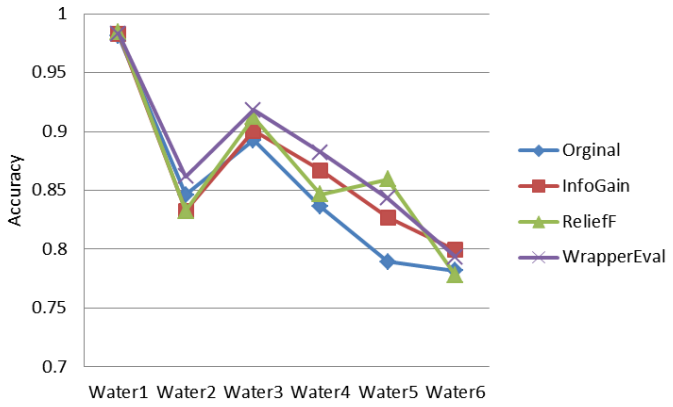


**Fig. 2.** Comparison of accuracy based on C4.5

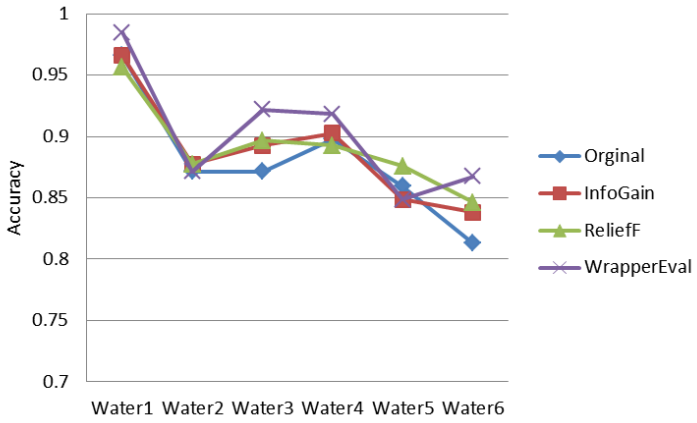


**Fig. 3.** Comparison of accuracy based on NaivBayes

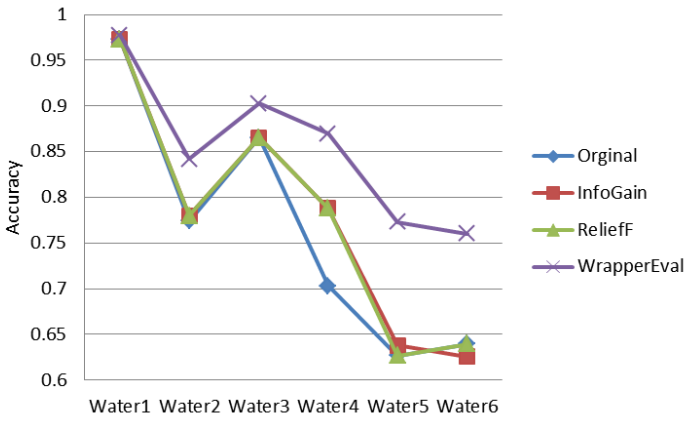


**Fig. 4.** Comparison of accuracy based on RBF-SVM

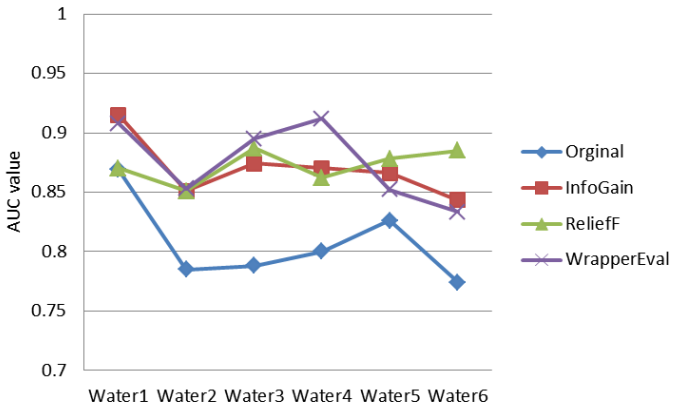


**Fig. 5.** Comparison of AUC based on C4.5

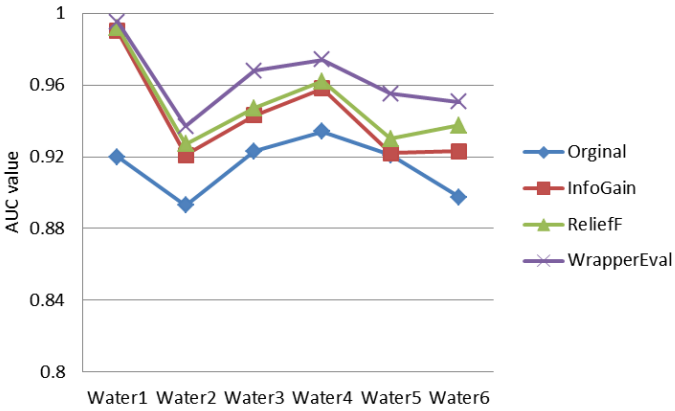


**Fig. 6.** Comparison of AUC based on NaivBayes

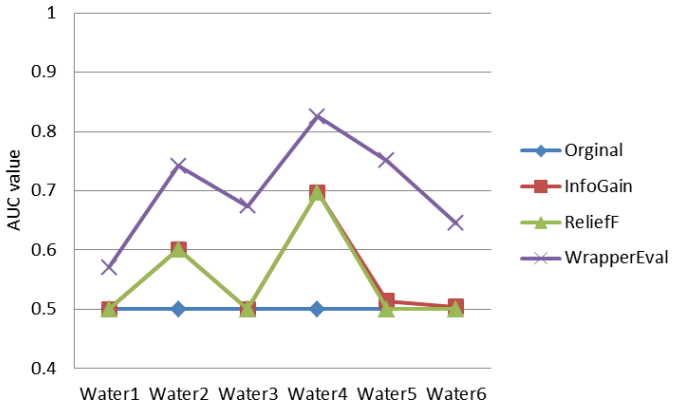


**Fig. 7.** Comparison of AUC based on RBF-SVM

## 5. Conclusion

Wastewater treatment fault diagnosis belongs to high-dimensional imbalanced data classification problem. In this study, we adopt feature selection algorithms to reduce the dimensionality of the data and improve the fault diagnosis of wastewater treatment equipment. Two filter-based feature selection methods and one wrapper-based feature selection method were used for experiments. Experimental results demonstrate that, compared with the original feature set classification data, the three feature selection algorithms have a significant improvement in the overall classification accuracy and AUC value. In addition, our experiments show that WrapperEval is better than InfoGain and ReliefF. The conclusions obtained have definite application value in the fault diagnosis of wastewater treatment equipment. Future work is to examine the distribution of samples while considering the feature attributes of the dataset and to change the spatial distribution of samples by sampling algorithm to further improve the classification performance of fault diagnosis of wastewater treatment equipment.

## Acknowledgments

Research on this work was partially supported by the funds from Jiangxi Education Department (No. GJJ211919).

## References

1. Speece, Richard E . Anaerobic biotechnology for industrial wastewater treatment[J]. Environmental Science & Technology, 17(9):416A (1983).
2. PR Shrestha., S Shrestha.: Troubleshooting for improved bio P at Lundåkraverket wastewater treatment plant, Landskrona, Sweden[J]. (2008).
3. Villegas, Fuente, SainzPalmero.: Fault diagnosis in a wastewater treatment plant using dynamic Independent Component Analysis[C]// Control & Automation. IEEE, (2010).
4. XU Yuge, DENG Wenkai, CHEN Liding.: Online fault diagnosis in wastewater treatment process by kernel-based weighted extreme learning machine[J]. Ciesc Journal, (2016).
5. Tao E P, Shen W H, Liu TL., et al.: Fault diagnosis based on PCA for sensors of laboratorial wastewater treatment process[J]. Chemometrics and Intelligent Laboratory Systems, 128:49–55 (2013).
6. Ribeiro R,. Pinheiro C C., Arriaga T., et al.: Model Based Fault Diagnosis for Performance Control of a Decentralized Wastewater Treatment Plant[J]. Computer Aided Chemical Engineering, 33:691-696 (2014).
7. He H., Garcia EA.: Learning from Imbalanced Data[J]. IEEE Transactions on Knowledge and Data Engineering. 21(9):1263-1284 (2009).
8. Yin L., Ge Y., Xiao K., et al.: Feature selection for high-dimensional imbalanced data [J]. Neurocomputing. 105(3): 3-11 (2013).
9. Shang C., Li M., Feng S., et al.: Feature selection via maximizing global information gain for text classification[J]. Knowledge-Based Systems, 54(Complete):298-309 (2013).
10. Guyon I , Elisseeff, André.: An Introduction to Variable and Feature Selection[J]. Journal of Machine Learning Research, 3(6):1157-1182 (2013).

11. Li Y., Liang X., Lin J., et al.: Train axle bearing fault detection using a feature selection scheme based multi-scale morphological filter[J]. Mechanical Systems & Signal Processing, 101(FEB.15):435-448 (2018).
12. Hancer E., Xue B., Zhang M.: Differential evolution for filter feature selection based on information theory and feature ranking[J]. Knowledge-Based Systems, 140(Jan.15):103-119 (2018).
13. Ma L., Li M., Gao Y., et al.: A Novel Wrapper Approach for Feature Selection in Object-Based Image Classification Using Polygon-Based Cross-Validation[J]. IEEE Geoscience & Remote Sensing Letters, 14(3):409-413 (2017).
14. Tran CT., Zhang M., Andreae P., et al:. Improving performance for classification with incomplete data using wrapper-based feature selection[J]. Evolutionary Intelligence, 9(3):1-14 (2016).
15. Ma S., Song X, Huang J.: Supervised group Lasso with applications to microarray data analysis[J]. bmc bioinformatics, 8(1):1-17 (2017).
16. Mistry K., Zhang L., Neoh SC., et al.: A Micro-GA Embedded PSO Feature Selection Approach to Intelligent Facial Emotion Recognition[J]. IEEE Transactions on Cybernetics, 47(6):1496-1509 (2017).
17. Steven L., Salzberg.: Book Review. C4.5: Programs for Machine Learning by J. Ross Quinlan[J]. Machine Learning, 16(3):235-240 (1994).
18. Hang Guo., Lizhu Zhou.: A Framework for Titled Document Categorization with Modified Multinomial Naivebayes Classifier[C]// Advanced Data Mining and Applications, Third International Conference, ADMA 2007, Harbin, China, August 6-8, Proceedings. Springer-Verlag, (2007).
19. Zhang J ., Ji R., Yuan X., et al.: Recognition of Pest Damage for Cotton Leaf Based on RBF-SVM Algorithm[J]. Nongye Jixie Xuebao/Transactions of the Chinese Society of Agricultural Machinery, 42(8):178-183 (2011).
20. UC Irvine Machine Learning Repository, http：//archive.ics.uci.edu/ml/2009.