Research Article

A Correlation-Based Feature Selection Algorithm for Operating Data of Nuclear Power Plants

Yuxuan He¹, **Hongxing Yu²**, **Ren Yu³**, **Jian Song¹**, **Haibo Lian¹**, **Jiangyang He¹**,
and **Jiangtao Yuan¹**

¹Navy Submarine Academy, Qingdao, Shandong 266000, China

²Nuclear Power Institute of China, Chengdu, Sichuan 610041, China

³Naval University of Engineering, Wuhan, Hubei 430033, China

Correspondence should be addressed to Jian Song; jian_s@hotmail.com

Received 4 April 2021; Revised 11 August 2021; Accepted 12 August 2021; Published 29 August 2021

Academic Editor: Keith E. Holbert

Copyright © 2021 Yuxuan He et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nuclear power plant operating data are characterized by a large variety, strong coupling, and low data value density. When using machine learning techniques for fault diagnosis and other related research, feature selection enables dimensionality reduction while maintaining the physical meaning of the original features, thus improving the computational efficiency and generalization ability of the learning model. In this paper, a correlation-based feature selection algorithm is developed to implement feature selection of nuclear power plant operating data. The proposed algorithm is verified by experiments and compared with traditional correlation-based feature selection algorithms. The experiments and comparison results show that the proposed algorithm is effective in realizing the dimensionality reduction of nuclear power plant operating data.

1. Introduction

During the real-time operation of a nuclear power plant, the parameters such as temperature and pressure are monitored constantly and recorded. When studies based on machine learning are carried out for fault diagnosis or anomaly detection using nuclear power plant operating data, dimensionality reduction is required to avoid the calculation delay caused by the excessive data volume and the interference of weakly related parameters on the prediction accuracy [1–5]. Feature selection is interpretable because it retains the physical meaning of the original feature, which makes it more advantageous in related research [6, 7]. Currently, the feature selection method for nuclear power plant operating data mainly uses expert knowledge or simple data preprocessing. Santhosh et al. obtained a new feature set by artificially selecting features according to professional knowledge [8–10]. Wang et al. comprehensively considered the reliability requirements of the nuclear power system and the coupling relationship between the parameters and expert experience feedback to determine the strong correlation

parameters [11]. Wang et al. used statistical methods to remove low variance features in time-series data [12]. The work of Peng et al. was based on relevance analysis and used the Pearson coefficient to evaluate correlation to delete features that have little impact on classification [13]. Na et al. proposed a method, which combines correlation analysis with a genetic algorithm to realize automatic input selection [14, 15]. However, faced with specific classification problems, some important features may be mistakenly removed or the new feature set may contain a large number of redundant features, which makes it impossible to achieve satisfactory dimensionality reduction effects with guaranteed model performance. Since nuclear power plant operating data have many types and strong coupling characteristics, common feature selection algorithms sometimes cannot accurately identify redundant features. In this context, this paper proposes a correlation-based feature selection algorithm for operating data of nuclear power plants (NPP-FS).

The main contributions of the proposed method can be summarized as follows:

- (i) We consider the characteristics of nuclear power plant operating data and propose that the maximal information coefficient should be used to determine the strength of the correlation between features and class. The results are used as the basis for feature ranking to delete irrelevant features.
- (ii) We propose an improved approximate Markov blanket concept, which can avoid the risk of excessively removing redundant features when making redundant judgments on nuclear power plant operating data.
- (iii) We conduct a series of experiments on the data generated by the simulation based on different backgrounds. The effects of the proposed method are investigated, and the feature selection performance is compared with traditional feature selection methods. The results show that the proposed method outperforms the conventional method in the dimensionality reduction effect.

The remainder of this article is arranged as follows. Section 2 reviews classical correlation-based feature selection algorithms and points out the problems in applying these algorithms to nuclear power plant operating data. Sections 3 and 4 introduce the details of the proposed feature selection algorithm. Section 5 evaluates the validity of the proposed theory and method through experiments. Conclusions are given in Section 6, which summarizes the advantages and disadvantages of the proposed approach as well as some future topics.

2. Related Work

The feature selection algorithm based on correlation is a common method to obtain the input features of model learning, and it has received widespread attention in many fields such as text classification [16–19], bioinformatics [20], and genetic analysis [21]. The correlation theory regards the feature-class correlation as feature relevance and the feature-feature correlation as redundancy [22]. Then, an original feature set is considered to be composed of strongly relevant features, weakly relevant but nonredundant features, redundant features, and irrelevant features. It is expected that a feature subset containing only nonredundant features and strongly relevant features can be obtained through feature selection algorithms. The most representative of these algorithms are correlation-based feature selection (CFS) [23], minimal-redundancy-maximal-relevance criterion (mRMR) [24], and fast correlation-based filter (FCBF) [25, 26]. CFS uses a forward search strategy to estimate the performance of a subset of features rather than a single feature. It uses Merit_s as an evaluation function to examine the effect of adding a certain feature to the Merit_s value. The mRMR is similar to the CFS algorithm in that it uses a mutual information-based evaluation function to estimate the correlation and redundancy while also taking into account the performance of the feature subset as a whole. FCBF uses symmetric uncertainty as to the correlation measure and decouples the relevance analysis and redundancy analysis to sequentially perform

feature ranking and feature search to gradually remove irrelevant and redundant features.

Unlike the publicly available datasets used to validate such feature selection algorithms, nuclear power plant operating data have the following characteristics. (1) Various data sources: in addition to variables such as temperature, pressure, water level, and flow from the loop, process parameters such as neutron detection and radiation monitoring are also involved. (2) Strong coupling between variables: for example, the reactor physical and thermal-hydraulic parameters influence each other. (3) Low data value density: fault or abnormality data are low, and the occurrence of a certain fault or abnormality is often reflected only by individual variables.

The characteristics possessed by nuclear power plant operating data can amplify the shortcomings of the above feature selection algorithm. CFS and mRMR sometimes have poor recognition performance for strongly coupled features to take into account the overall performance of feature subsets. The symmetric uncertainty used in FCBF is weak in identifying nonlinear relationships, which can cause the results of feature ranking to fail. These problems pose challenges for the application of correlation-based feature selection algorithms.

To overcome the problems of existing algorithms based on correlation and meet the needs of nuclear power plant operating data feature selection, we develop a novel algorithm that can provide a collection of features with strong correlation and low redundancy.

3. Correlation-Based Measures

Choosing an appropriate measure of correlation is crucial for a correlation-based feature selection algorithm. The mutual information is first introduced in this section. We then focus on two mutual information-based correlation measures adopted by the NPP-FS algorithm. Maximal information coefficient is applied to measure nuclear power plant operating data. Symmetric uncertainty is used to improve the approximate Markov blanket.

3.1. Mutual Information. There are many measures based on the information-theoretical concept of entropy among nonlinear correlation measures. It solves the measurement of the uncertainty of random variables. The entropy of variable X can be defined as

$$H(X) = - \sum_i P(x_i) \log P(x_i), \quad (1)$$

where $P(x_i)$ represents the prior probabilities for all values of X .

Conditional entropy is the conditional probability distribution of random variable X when random variable Y occurs alone. It is defined as

$$H(X|Y) = - \sum_j P(y_j) \sum_i P(x_i|y_j) \log P(x_i|y_j), \quad (2)$$

where $P(x_i|y_j)$ represents the posterior probabilities of X given the values of Y .

Mutual information can be expressed by equations (1) and (2), given by [27]

$$IG(X|Y) = H(X) - H(X|Y). \quad (3)$$

We can easily know that $IG(X|Y) = IG(Y|X)$ by derivation, and it can be expressed as $I(X, Y)$. Mutual information is not normalized, and it tends to have more values. It is generally not directly used for correlation measurement.

3.2. Maximal Information Coefficient. The maximal information coefficient is a nonparametric statistical method proposed by Reshef et al. that can measure the correlation between two variables [28]. Under the condition of sufficient sample data and information, it can identify any type of functional relationship (including the superposition of noise-free functions), and it is more equitable than mutual information. The calculation method of the maximal information coefficient (MIC) is as follows.

Given a data set D consisting of n -ordered pairs of variables, and a scatter plot consisting of all pairs of variables in this set, the X -axis and Y -axis are divided into a grid according to the number of x , y , and the grid G is obtained. Under the condition that x and y remain unchanged, grid G has multiple division methods. If $I(D|G)$ represents the mutual information of the dataset D divided by G , then the highest mutual information is expressed as follows:

$$I^*(D, x, y) = \max I(D|G). \quad (4)$$

Normalize the highest mutual information obtained under the condition of different x and y to form the characteristic matrix $M(D)_{x,y}$, and it can be expressed by the following equation:

$$M(D)_{x,y} = \frac{I^*(D, x, y)}{\log \min(x, y)}. \quad (5)$$

On the dataset D , the MIC can be calculated as follows:

$$\text{MIC}(D) = \max_{xy < B(n)} \{M(D)_{x,y}\}, \quad (6)$$

where $xy \leq B(n) = n^\alpha$, n is the number of samples, and α is an adjustable parameter. The value of α will affect the density of the grid so that it will affect the judgment of the correlation. According to Reshef's suggestion, α is 0.6 in the experiment involved in this article.

The MIC is symmetric due to the symmetry of mutual information and its score is in the range $[0, 1]$. When there is a noise-free relationship between two variables, the MIC tends to 1, and when the two variables are statistically independent, the MIC tends to 0.

3.3. Symmetric Uncertainty. Symmetric uncertainty (SU) is a correlation measure based on normalized mutual information proposed by Kvalseth [29] and is often used in feature selection algorithms based on correlation. The calculation method of SU is shown in the following equation [30]:

$$\text{SU}(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)}, \quad (7)$$

where $\text{SU}(X, Y)$ is the SU of random variables X and Y , $I(X, Y)$ is the mutual information between X and Y , and $H(\cdot)$ represents the information entropy of the random variable.

SU can compensate for the bias that mutual information tends to have more value features and at the same time normalizes its values to the range $[0, 1]$. $\text{SU}(X, Y) = 0$ indicates X and Y are independent, and $\text{SU}(X, Y) = 1$ means that one variable can be completely predicted by another variable. In addition, SU is also symmetric to a pair of variables because of the symmetry of mutual information.

4. Method

4.1. Improved Approximate Markov Blanket. Markov blanket is a method proposed by Koller to eliminate redundant features, and its definition is as follows [31].

M_j does not contain f_j , and we say that M_j is a *Markov blanket* for f_j iff $P(F - M_j - \{f_j\}, C|f_j, M_j) = P(F - M_j - \{f_j\}, C|M_j)$.

The meaning of the above formula is that M_j contains all the information of f_j about C and $F - M_j - \{f_j\}$, that is, when there is a feature subset M_j , feature f_j does not contribute to classification. The redundancy feature based on the Markov blanket is defined as follows.

A feature f_j is redundant iff f_j is weakly relevant and there is a Markov blanket in F .

Because obtaining a feature's Markov blanket is an NP-hard problem, the approximate Markov blanket method is usually used to eliminate redundant features in applications. A commonly used approximate Markov blanket is defined as follows.

f_j forms an approximate *Markov blanket* for f_i iff $J(f_j, C) \geq J(f_i, C)$ and $J(f_j, f_i) \geq J(f_i, C)$, and $J(\cdot)$ is optional correlation measure.

This approximation method takes the feature-class correlation and the feature-feature correlation as the solution condition of the Markov blanket at the same time, and the time complexity is $O(N \log N)$. This can avoid enumerating all combinations of features under non-approximate conditions. However, the use of this approximation method in practical applications may misjudge strongly relevant features as redundant features and eliminate them, which will cause the subset obtained by feature selection to lose part of the information needed for classification [32]. Therefore, this paper proposes an improved approximate Markov blanket method, which is defined as follows.

f_j forms an approximate *Markov blanket* for f_i iff $\text{MIC}(f_j, C) \geq \text{MIC}(f_i, C)$, $\text{MIC}(f_j, f_i) \geq \text{MIC}(f_i, C)$, and $\text{SU}(f_j, f_i) > \gamma$.

In the formula, the value of γ is in the range $[0, 1]$, which is used to constrain the Markov blanket to contain the information amount of the eliminated feature. The larger the value of γ , the higher the feature-feature correlation of the eliminated features and the stronger the redundancy. This

improvement can be explained by Figure 1 as follows. When eliminating redundant features, the feature-feature correlation consists of two parts: class-related (yellow parts) and class-independent (red parts). We impose a constraint through SU so that the deleted feature has high class-related rather than class-independent. The reason for not using the MIC as a constraint is its strong ability to mine the correlation between variables, and it may introduce more class-independent correlations into the results.

Through this improvement, it is possible to ensure that the deleted redundant features have high feature-feature correlation, thereby avoiding misjudgment of strongly relevant features and obtaining a feature set with better performance.

4.2. Algorithm. Based on the selected correlation measures and the improved approximate Markov blanket presented before, we develop a two-stage algorithm, named NPP-FS. Given a dataset with d features and a class C , the final output is a selected subset after two stages. The detailed process of the algorithm is shown in Figure 2.

The first stage uses the original dataset as input for relevance analysis (lines 5–11). First, the MIC is used to calculate the correlation of each feature, and then the feature correlation ranking method is used to eliminate some features. In this process, set a threshold σ with a value of $[0, 1]$, add all the features with a feature-class correlation higher than σ to an empty set S_{list}' , and finally arrange the features in S_{list}' in descending order of the feature-class correlation to obtain the relevance analysis subset. The feature of the first stage is that only the correlation of a single feature is considered, and the time complexity is low, but the redundant features in the dataset cannot be eliminated. The features in the relevance analysis subset should be strongly correlated or weakly correlated, and the features whose feature-class correlation is less than σ are regarded as irrelevant features and are eliminated in relevance analysis. Therefore, when σ takes different values, a relevance analysis subset with different dimensions can be obtained, which also determines the minimum feature-class correlation of the features in the final subset that can be selected by the NPP-FS algorithm.

The second stage uses a subset of relevance analysis as input for redundancy analysis (lines 12–22). Based on the feature correlation ranking results, the redundancy of features is evaluated through MIC and SU, and then the improved approximate Markov blanket is used to eliminate redundant features. The process of redundancy analysis includes a nested loop. The inner loop applies an improved approximate Markov blanket to the feature f_j and all features in the current set whose ranking is lower than f_j under the condition of a given γ threshold. Only one feature is judged at a time. If the judgment condition is met, f_i is regarded as a redundant feature and eliminated from the current set. After completing the inner loop, the outer loop uses the next feature of f_j in the current set as the new f_j . The outer loop ends when there is no new f_j . The feature set at this time is the final subset obtained by the NPP-FS

algorithm. The feature of the second stage is that the relevance and redundancy of features are comprehensively considered, and finally a feature selection subset with higher relevance and lower redundancy can be obtained.

Combining the first stage and second stage, it can be seen that under the condition of a given relevance analysis subset, the result of redundancy analysis is only affected by the value of γ . When the σ value is fixed, the dimension of the final feature selection subset can be further controlled by selecting different γ values. Therefore, in practical applications, the values of σ and γ can be adjusted at the same time to generate feature selection subsets composed of different features and having lower dimensions to adapt to different datasets and classification problems.

5. Empirical Study

5.1. Data Sources. Due to irresistible reasons, it is very difficult to obtain fault data in nuclear power plants, so the amount of fault data that can be collected is limited [33]. To verify the effectiveness of the NPP-FS algorithm and apply it to the classification of steady conditions and typical accident conditions, 12 conditions were simulated by nuclear power plant simulation software [34]. It generated a total of 10800 sample data, and each sample point includes 90 parameters (not considering time series). The dataset needed for the experiment can be constructed from the generated sample data. Table 1 lists the detailed information of the sample data.

5.2. Correlation Measure Verification. This experiment is mainly used to verify the validity of the MIC as a correlation measure of nuclear power plant operating data. The experimental program is shown in Figure 3.

First, we construct the experimental dataset shown in Table 2 according to the original data in Table 1. To reflect the advantage of the MIC to evaluate the correlation, the SU is used for comparative experiments and combined with professional knowledge to evaluate the use of the MIC as a correlation measure of nuclear power plant operating data.

The MIC is calculated by the minepy module [35]. When the features are continuous data, the calculation of SU is discretized by the histogram method [36]. The experimental results are retained to 6 decimal places.

Figure 4 shows the calculation results of feature-class correlation, and the meaning of the abbreviations can be found in [34]. In experiments 1–5, the average value of the MIC is higher than the average value of the SU. In experiment 1, the difference between the MIC and the SU is the smallest, which is 0.236. In experiment 2, the difference between the two is the largest, which is 0.437. This shows that in the nuclear power plant operating data, the MIC is stronger than the SU as a whole to measure the correlation. But when the calculated value is less than 0.2, the difference between the calculation results of the two correlation measures is small, indicating that the performance of the two indicators is relatively close at this time. In addition, the situation where the two measures are 0 is the same,

FIGURE 1: Venn diagram showing the relationships among feature i , feature j , and class C .

```

1 Input:  $S = \{f_1, \dots, f_d, C\}$  //a training data set
2  $\sigma$  //a threshold of stage-one
3  $\gamma$  //a threshold of stage-two
4 Output:  $S_{best}$  //a selected subset
5 begin
6   for  $i=1$  to  $d$  do begin
7     calculate  $MIC_{i,c}$  for  $f_i$ ;
8     if ( $MIC_{i,c} > \sigma$ )
9       append  $f_i$  to  $S'_{list}$ ;
10    end
11    order  $S'_{list}$  in descending  $MIC_{i,c}$  value;
12     $f_j = \text{getFirstElement}(S'_{list})$ ;
13    do begin
14       $f_i = \text{getNextElement}(S'_{list}, f_j)$ ;
15      if ( $f_i \neq \text{NULL}$ )
16        do begin
17          if ( $MIC_{i,j} \geq MIC_{i,c}$ ) and ( $SU_{i,j} > \gamma$ )
18            remove  $f_i$  from  $S'_{list}$ ;
19             $f_i = \text{getNextElement}(S'_{list}, f_j)$ ;
20          end until ( $f_i == \text{NULL}$ );
21           $f_j = \text{getNextElement}(S'_{list}, f_j)$ ;
22          end until ( $f_j == \text{NULL}$ );
23       $S_{best} = S'_{list}$ ;
24    end;

```

FIGURE 2: NPP-FS algorithm.

indicating that the two measures have the same effect in identifying features that are completely unrelated to classification. The above analysis shows that the results of calculating the operating data of nuclear power plant by the two measures are different, and the order of the features according to the number of the correlation measure is also inconsistent.

The mechanism of different operating conditions in experiments 1–5 is clear. Combined with professional knowledge, it can be found that the use of SU to evaluate feature-class correlation will assign lower values to some features that are more important for classification, such as TAVG, PSGA, and PSGB in experiment 1, P in experiment 2, WFWA and WFWB in experiment 3, WRCA and WRCB in experiment 4, and LVPZ in experiment 5.

The reason for this phenomenon is illustrated in Figure 5, which shows the TAVG and TBLD data contained in the class label “75% power” in experiment 1. It can be found from Figure 5 that the parameters of the feature TAVG fluctuate around the setting value, while the parameters of the TBLD have a fixed value. Combining

Figure 4, it can be seen that when the operating parameters of a feature fluctuate in the data under the same class label, the SU cannot accurately identify its correlation, but this type of feature is widely present in the operating data of nuclear power plants.

Based on the above comparative analysis, when the MIC is used to evaluate the correlation, its ability to mine the correlation is overall stronger, and the relative value of different features is also in good agreement with the professional knowledge.

5.3. Sensitivity Analysis Experiment. The NPP-FS algorithm contains two adjustable parameters, σ and γ . σ is used to delete irrelevant features, and γ is used as a constraint to approximate the Markov blanket to avoid eliminating strongly relevant features. This experiment mainly explores the influence of the different values of these two parameters on the results of the NPP-FS algorithm. The experimental program is shown in Figure 6.

The experiment uses the method of establishing a classifier model to evaluate the feature selection results of the NPP-FS algorithm when different parameters are selected. First, we construct the experimental dataset shown in Table 3 according to the simulated data in Table 1. In data preprocessing, the NPP-FS algorithm is first used for feature selection, and the grid search method shown in Figure 6 is used to obtain the feature selection results when different parameter values are taken. After the experimental dataset obtains the feature selection subset by the NPP-FS algorithm, to avoid the inconsistency of the feature value range from affecting the classifier, the data are normalized. Since the maximum and minimum values of each feature in the dataset are known, the following formula is used for normalization:

$$X = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \quad (8)$$

After data preprocessing is completed, three classifiers of logistic regression (LR) [37], support vector machine (linear kernel, SVM) [38], and k nearest neighbors (KNN) [39] are used for evaluating the selected subset of features. The evaluation of the classifier uses the average of 10-fold cross-validation accuracy. Finally, through the performance of the classifier on different feature selection subsets, the results of the NPP-FS algorithm in different parameter values are evaluated.

Figure 7 shows how the accuracy of the classifier varies with σ and γ . The results show that the accuracy of the nuclear power plant operating data on the three classifiers is different. This is caused by the different calculation principles of the classifier. On the same experimental dataset, the results of LR and SVM are relatively consistent, and the results of the two and KNN are different.

5.3.1. The Results of LR and SVM. The classification accuracy obtained by the sensitivity analysis experiment using LR and SVM shows an obvious partition phenomenon. In each graph, there is a transition zone with a relatively large

TABLE 1: Original data.

Number	Condition	Initial condition	Criteria	Sample size
1	50% power	—	—	900
2	75% power	—	—	900
3	100% power	—	—	900
4	Loss of coolant accident (hot leg)	100% power	2/10/100 cm ²	300/300/300
5	Loss of coolant accident (cold leg)	100% power	2/10/100 cm ²	300/300/300
6	Steam generator A tube rupture	100% power	0.5/1/2 (*100%)	300/300/300
7	Steam generator B tube rupture	100% power	0.5/1/2 (*100%)	300/300/300
8	Steam line break inside containment	100% power	25/50/75 cm ²	300/300/300
9	Steam line break outside containment	100% power	25/50/75 cm ²	300/300/300
10	Loss of flow (locked rotor)	50%/75%/100% power	—	300/300/300
11	Turbine trip	50%/75%/100% power	—	300/300/300
12	Loss of feedwater accident	50%/75%/100% power	—	300/300/300
Total				10800

FIGURE 3: Program of correlation measure verification experiment.

TABLE 2: Correlation measure verification experiment data.

Number	Classes	Data number in Table 1	Class label
1	3	{1} {2} {3}	50% power/75% power/100% power
2	2	{1, 2, 3} {4, 5}	Normal condition/loss of coolant accident
3	2	{1, 2, 3} {12}	Normal condition/loss of feedwater
4	2	{1, 2, 3} {10}	Normal condition/loss of flow (locked rotor)
5	2	{1, 2, 3} {6, 7}	Normal condition/steam generator tube rupture

accuracy gradient. According to the theory of the NPP-FS algorithm, in the relevance analysis, irrelevant, weakly relevant, and strongly relevant features will be eliminated in turn as σ increases. In the redundancy analysis, with the increase of γ , there is a tendency to keep the strongly relevant and weakly relevant features in the feature subset in turn. Therefore, the shape of the transition zone is the result of the combined effect of σ and γ . When the parameter values in the vicinity of the transition zone change, the strongly relevant features will be eliminated or remain in the feature subset, resulting in a relatively large gradient of accuracy in the transition zone.

According to the NPP-FS algorithm shown in Figure 8, the dimensionality of the subset generated by σ and γ changes can be further found, that is, the results of LR and SVM have a strong correlation with the change of the feature selection subset dimension with the parameters. σ with higher accuracy should be taken in the σ interval smaller

than the transition zone, but when σ is less than 0.2, the change of the feature selection subset dimension has no obvious effect on the accuracy. The value of γ to obtain results with higher accuracy is usually greater than 0.4. However, when $\sigma \in [0.1, 0.3]$ and $\gamma \in [0, 0.4]$, affected by the approximate Markov blanket, the dimension of the subset of experiment 4 first increases and then decreases with the increase of the value of γ . This also shows that the improved approximate Markov blanket can effectively alleviate the problem of incorrectly eliminating nonredundant features.

5.3.2. The Results of KNN. According to the results shown in Figure 7, KNN has better adaptability to the classification of operating conditions of nuclear power plants and can achieve higher accuracy in a larger parameter range. In binary classification, the accuracy of experiments 1 and 4 is above 99% in the full parameter experiment range; when

(a)

(b)

FIGURE 4: Continued.

FIGURE 4: Results of correlation measure verification experiment. (a)–(e) represent experiments 1 to 5, and the dotted line represents the average value.

FIGURE 5: Data with the class label of “75% power” in experiment 1.

$\sigma > 0.1$, the accuracy of experiment 2 is above 94%; experiment 3 has an accuracy higher than 96% in the full parameter experiment range. In combination with Figure 8, it can be found that in binary classification, the dimensionality of the feature selection subset has little effect on the accuracy of KNN, and only a few features are needed to obtain a higher classification accuracy. The results of experiments 5 and 6 show that in the multiclassification problem, the accuracy of KNN is also strongly related to the change of the feature selection subset dimension with parameters, but the range of parameters with higher accuracy is larger than that of LR and SVM.

5.4. Comparative Experiment and Analysis. To reflect the advantages of the NPP-FS, the same dataset and experimental program as mentioned in Section 5.3 are used to compare the NPP-FS algorithm with the feature selection algorithms CFS, mRMR, and FCBF. Using the data in Table 3 as input, we run the original feature set (Fullset) and all four feature selection algorithms, NPP-FS, CFS, mRMR, FCBF, respectively, in the three classifiers of LR, SVM, and KNN. The normalization method adopts equation (8), and the way to evaluate models is the average of 10-fold cross-validation. We use the sklearn library to build these models with default parameters, which can avoid the uncertain

FIGURE 6: Program of sensitivity analysis experiment.

TABLE 3: Sensitivity analysis experiment data.

Number	Classes	Samples	Data number in Table 1
1	2	4500	{1, 2, 3} {4, 5}
2	2	4500	{1, 2, 3} {6, 7}
3	2	1800	{4} {5}
4	2	3600	{4, 5} {6, 7}
5	6	5400	{4}{5}{6}{7}{8}{9}
6	12	10800	{1}{2}{3}{4}{5}{6}{7}{8}{9}{10}{11}{12}

FIGURE 7: Continued.

FIGURE 7: Results of the sensitivity analysis experiment. (a)–(f) represent experiments 1 to 6 with (i) LR, (ii) SVM, and (iii) KNN.

effects caused by parameter tuning. Tables 4–6, respectively, show the classification accuracy and the number of selected features (round half to integer) in the three classifiers of LR, SVM, and KNN in the above five cases.

It can be seen from Table 4 that the classification accuracy of NPP-FS on the LR model is lower than that of mRMR in experiment 1 and CFS in experiment 3, but the performance of NPP-FS is better than other algorithms in other experiments. The average accuracy of NPP-FS in six experiments is 2.42% higher than that of the CFS algorithm, 4.59% higher than that of the mRMR algorithm, and 17.9% higher than that of the FCBF algorithm. Although the average number of feature subsets generated by NPP-FS is larger than that of FCBF, NPP-FS has an obvious advantage in terms of classification accuracy. The results shown in Table 5 on the SVM classifier are relatively similar to those

on the LR classifier. The classification accuracy of NPP-FS is slightly lower than that of mRMR in experiment 1 and experiment 5 and slightly lower than that of CFS in experiment 3, but in all other experiments, NPP-FS outperformed the other algorithms. The mean of the six experiments' accuracy is 1.65% higher than that of CFS, 4.66% higher than that of mRMR, and 17.47% higher than that of FCBF. Similarly, although the NPP-FS algorithm produces a larger average number of feature subsets than FCBF in SVM, it has a remarkable advantage in terms of classification accuracy. As shown in Table 6, the results of NPP-FS, CFS, mRMR, and FCBF on the KNN classifier are better than those on LR and SVM classifiers overall. The accuracy of NPP-FS is slightly lower than that of FCBF in experiment 2, but in all other experiments, NPP-FS is superior to the rest of the algorithms. The average accuracy of

FIGURE 8: The reciprocal of the subset dimension changes with the parameter value. (a)–(f) represent experiments 1 to 6.

TABLE 4: Comparison of feature selection algorithms under LR classifier.

Number	Fullset (%)	NPP-FS (%)	CFS (%)	mRMR (%)	FCBF (%)
1	98.62 ± 0.55 (90)	96.84 ± 0.89 (2)	96.42 ± 0.86 (2)	98.62 ± 0.67 (2)	96.42 ± 0.62 (2)
2	100.00 ± 0.00 (90)	100.00 ± 0.00 (6)	97.49 ± 0.59 (6)	100.00 ± 0.00 (3)	92.47 ± 0.89 (2)
3	98.89 ± 0.93 (90)	96.94 ± 1.52 (3)	98.22 ± 0.10 (4)	83.17 ± 1.74 (20)	52.72 ± 2.49 (2)
4	99.30 ± 0.53 (90)	99.53 ± 0.34 (2)	98.92 ± 0.53 (7)	98.92 ± 0.50 (2)	98.92 ± 0.47 (3)
5	97.57 ± 0.87 (90)	96.44 ± 0.62 (17)	86.52 ± 1.00 (11)	96.29 ± 0.74 (20)	67.41 ± 1.44 (4)
6	98.34 ± 0.45 (90)	96.24 ± 0.65 (18)	93.94 ± 0.67 (17)	81.48 ± 1.01 (30)	70.67 ± 0.45 (6)
Average	98.79 ± 0.56 (90)	97.67 ± 0.67 (8)	95.25 ± 0.63 (8)	93.08 ± 0.78 (13)	79.77 ± 1.06 (3)

TABLE 5: Comparison of feature selection algorithms under SVM classifier.

Number	Fullset (%)	NPP-FS (%)	CFS (%)	mRMR (%)	FCBF (%)
1	100.00 ± 0.00 (90)	98.42 ± 0.90 (2)	96.42 ± 0.86 (2)	98.62 ± 0.47 (2)	96.42 ± 0.62 (2)
2	100.00 ± 0.00 (90)	100.00 ± 0.00 (6)	98.11 ± 0.60 (6)	100.00 ± 0.00 (3)	95.29 ± 0.92 (2)
3	98.67 ± 1.00 (90)	97.22 ± 1.52 (3)	98.17 ± 0.07 (4)	83.61 ± 2.29 (20)	52.78 ± 2.43 (2)
4	100.00 ± 0.00 (90)	99.47 ± 0.22 (2)	98.92 ± 0.53 (7)	98.92 ± 0.50 (2)	98.92 ± 0.47 (3)
5	99.43 ± 0.28 (90)	96.44 ± 0.82 (12)	91.19 ± 0.98 (11)	97.07 ± 0.98 (15)	70.33 ± 1.45 (4)
6	99.54 ± 0.19 (90)	97.33 ± 0.37 (21)	96.19 ± 0.55 (17)	82.72 ± 1.20 (30)	70.35 ± 0.97 (6)
Average	99.61 ± 0.25 (90)	98.15 ± 0.64 (8)	96.50 ± 0.60 (8)	93.49 ± 0.91 (12)	80.68 ± 1.14 (3)

NPP-FS in six experiments is 0.32% higher than that of the CFS algorithm, 2% higher than that of the mRMR algorithm, and 11.3% higher than that of the FCBF algorithm. The average number of NPP-FS feature subsets is the only one greater than the smallest FCBF among the four.

According to Tables 4–6, we can conclude that the overall performance of the NPP-FS algorithm is better than that of CFS, mRMR, and FCBF. This is made possible by the fact that NPP-FS obtains a subset of features that perform

well on each experimental dataset. The remaining three feature selection algorithms suffer, to varying degrees, from the problem of making the model performance sharply decline, such as CFS in experiment 5, mRMR in experiment 3 and experiment 6, and FCBF in experiment 3, experiment 5, and experiment 6. The reasons for this phenomenon are explained below. The strong coupling between nuclear power plant operating data features is often complex, and at the same time, deterministic mathematical relationships and features

TABLE 6: Comparison of feature selection algorithms under KNN classifier.

Number	Fullset (%)	NPP-FS (%)	CFS (%)	mRMR (%)	FCBF (%)
1	100.00 $\pm$ 0.00 (90)	99.78 $\pm$ 0.01 (1)	98.60 $\pm$ 0.61 (2)	98.64 $\pm$ 0.48 (2)	96.42 $\pm$ 0.62 (2)
2	99.93 $\pm$ 0.10 (90)	99.96 $\pm$ 0.07 (1)	99.96 $\pm$ 0.13 (3)	100.00 $\pm$ 0.00 (3)	99.98 $\pm$ 0.07 (2)
3	99.44 $\pm$ 0.43 (90)	99.83 $\pm$ 0.12 (6)	99.50 $\pm$ 0.39 (5)	95.89 $\pm$ 1.37 (4)	81.94 $\pm$ 2.27 (2)
4	100.00 $\pm$ 0.00 (90)	99.94 $\pm$ 0.01 (1)	99.94 $\pm$ 0.11 (3)	99.89 $\pm$ 0.18 (6)	99.94 $\pm$ 0.11 (3)
5	99.54 $\pm$ 0.35 (90)	99.52 $\pm$ 0.19 (6)	99.41 $\pm$ 0.26 (6)	94.52 $\pm$ 1.03 (7)	78.44 $\pm$ 0.74 (4)
6	99.75 $\pm$ 0.12 (90)	99.66 $\pm$ 0.23 (11)	99.33 $\pm$ 0.22 (9)	97.71 $\pm$ 0.35 (11)	74.13 $\pm$ 0.80 (6)
Average	99.78 $\pm$ 0.17 (90)	99.78 $\pm$ 0.11 (4)	99.46 $\pm$ 0.29 (5)	97.78 $\pm$ 0.57 (6)	88.48 $\pm$ 0.77 (3)

with such relationships are difficult to identify whether they are redundant features by general feature selection algorithms. When these features contain information necessary for classification, the feature selection algorithm incorrectly removes them as redundant features. If these features are not relevant to the classification, the results are not influenced by this factor. NPP-FS uses the MIC to evaluate the correlation, which can effectively identify the complex nonlinear relationships in the nuclear power plant operating data. Meanwhile, the misjudgment of redundant features is mitigated by an approximate Markov blanket improved for nuclear power plant operating data. Therefore, the overall performance of NPP-FS is superior in the experimental dataset.

6. Conclusion

In this study, we propose a novel correlation-based feature selection algorithm for solving the problem of poor performance in identifying redundant features when the existing correlation-based feature selection algorithm analyzes nuclear power plant operating data. We propose the MIC as a correlation measure for the characteristics of nuclear power plant operating data and conduct correlation analysis experiments on the MIC and SU by combining professional knowledge. The results of this section show that the MIC is more applicable to the nuclear power plant operating data. In addition, we purposefully improved an approximate Markov blanket to enable the task of eliminating redundant features of nuclear power plant operating data under a given constraint. Then, we demonstrate the validity of the proposed theory and method through parametric sensitivity analysis experiments. Finally, we compare the performance of the proposed algorithm with several typical correlation-based feature selection algorithms. The results show that the proposed algorithm performs better than the conventional algorithm.

Our proposed approach provides a general dimensionality reduction method for the application of machine learning techniques on nuclear power plant operating data. It can generate a feature-selected subset with good classification performance from labeled data under appropriate given parameters, resulting in a remarkable reduction of feature dimensionality, thus improving the computational efficiency and generalization ability of the model. However, there are two points to note in the application of this algorithm. (1) The MIC is less efficient to calculate compared to the traditional correlation measure. We can improve the efficiency of this algorithm by using SU to preprocess irrelevant features based on the characteristics that the MIC and SU respond consistently

to irrelevant features. (2) The sensitivity of the two parameters of the algorithm is poorly generalized across different datasets, and the parameters still need to be selected according to the specific dataset and classifier. When applying this algorithm, it is recommended to sample the data with class labels into a small sample dataset and then perform feature selection.

There are several topics for further focus in the future. First, it would be interesting to explore the performance of additional correlation measures on nuclear power plant operating data and methods for evaluating that performance. Second, the proposed approach can be combined with association rules and causal analysis in the field of nuclear power plants for studies such as anomaly detection.

Acronyms

NPP-FS:	The proposed feature selection algorithm
CFS:	Correlation-based feature selection
mRMR:	Minimal-redundancy-maximal-relevance criterion
FCBF:	Fast correlation-based filter
MIC:	Maximal information coefficient
SU:	Symmetric uncertainty
TAVG:	Average reactor coolant system temperature
PSGA:	Steam generator A pressure
PSGB:	Steam generator B pressure
P:	Reactor coolant system pressure
WFWA:	Steam generator A feedwater flow
WFWB:	Steam generator B feedwater flow
WRCA:	Reactor coolant loop A flow
WRCB:	Reactor coolant loop B flow
LVPZ:	Pressurizer level
TBLD:	Turbine load power
LR:	Logistic regression
SVM:	Support vector machine
KNN:	k nearest neighbors
Fullset:	Original feature set

Terms

Feature-class correlation:	Measure between feature and class
Feature-feature correlation:	Measure between features
Correlation measure:	A quantitative evaluation standard
Markov blanket:	A method to define redundant features

NP-hard:	A complexity class of decision problems
Strongly relevant:	Contains necessary information
Weakly relevant but non-redundant:	Contains part of necessary information
Redundant:	The information contained exists
Irrelevant:	Does not contain any necessary information

Nomenclature

$H(\cdot)$:	Information entropy
$IG(\cdot)/I(\cdot)$:	Mutual information
f :	A single feature
C :	Collection of class labels
F :	A feature set
$J(\cdot)$:	Optional correlation measure
{Data number}:	Dataset with a class label
σ :	A parameter of the algorithm
γ :	A parameter of the algorithm.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

- [1] D. Chang, M. Liu, and Y. Lee, "Accident diagnosis of a PWR fuel pin during unprotected loss of flow accident with support vector machine," *Nuclear Engineering and Design*, vol. 352, Article ID 110184, 2019.
- [2] M. Jianping and J. Jiang, "Semisupervised classification for fault diagnosis in nuclear power plants," *Nuclear Engineering and Technology*, vol. 47, no. 2, pp. 176–186, 2015.
- [3] T. Tambouratzis, J. Giannatsis, A. Kyriazis, and P. Siotropos, "Applying the computational intelligence paradigm to nuclear power plant operation," *International Journal of Energy Optimization and Engineering*, vol. 9, no. 1, pp. 27–109, 2019.
- [4] H. Wang, M.-j. Peng, J. Zheng, Y.-k. Liu, and B. R. Upadhyaya, "A hybrid fault diagnosis methodology with support vector machine and improved particle swarm optimization for nuclear power plants," *ISA Transactions*, vol. 95, pp. 358–371, 2019.
- [5] L. Pincirolì, P. Baraldi, A. Shokry et al., "A semi-supervised method for the characterization of degradation of nuclear power plants steam generators," *Progress in Nuclear Energy*, vol. 131, Article ID 103580, 2021.
- [6] B. Senliol, G. Gulgezen, L. Yu, and Z. Cataltepe, "Fast correlation based filter (FCBF) with a different search strategy," in *Proceedings of the 2008 23rd International Symposium on Computer and Information Sciences*, Istanbul, Turkey, 2008.
- [7] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Computing & Applications*, vol. 24, no. 1, pp. 175–186, 2014.
- [8] T. V. Santhosh, M. Kumar, I. Thangamani et al., "A diagnostic system for identifying accident conditions in a nuclear reactor," *Nuclear Engineering and Design*, vol. 241, no. 1, pp. 177–184, 2011.
- [9] T. V. Santosh, A. Srivastava, V. V. S. Sanyasi Rao, A. K. Ghosh, and H. S. Kushwaha, "Diagnostic system for identification of accident scenarios in nuclear power plants using artificial neural networks," *Reliability Engineering & System Safety*, vol. 94, no. 3, pp. 759–762, 2009.
- [10] J. Yang and J. Kim, "An accident diagnosis algorithm using long short-term memory," *Nuclear Engineering and Technology*, vol. 50, no. 4, pp. 582–588, 2018.
- [11] W. H. Wang, L. Yu, X. L. Wang, J. L. Hao, and J. T. Zheng, "Design of anomaly detection algorithm for nuclear power system based on included angle cosine," *Atomic Energy Science and Technology*, vol. 55, pp. 98–103, 2021.
- [12] S. L. Wang, Q. Cai, X. Q. Zhang, and Y. Q. Chen, "Study on diagnostic technique for nuclear power plant primary coolant circuit system break characteristics based on multi-classification SVM," *Atomic Energy Science and Technology*, vol. 48, pp. 462–468, 2014.
- [13] B.-S. Peng, H. Xia, Y.-K. Liu, B. Yang, D. Guo, and S.-M. Zhu, "Research on intelligent fault diagnosis method for nuclear power plant based on correlation analysis and deep belief network," *Progress in Nuclear Energy*, vol. 108, pp. 419–427, 2018.
- [14] M. G. Na, S. H. Shin, D. W. Jung, S. P. Kim, J. H. Jeong, and B. C. Lee, "Estimation of break location and size for loss of coolant accidents using neural networks," *Nuclear Engineering and Design*, vol. 232, no. 3, pp. 289–300, 2004.
- [15] M. G. Man Gyun Na, Y. R. Young Rok Sim, K. H. Sun Mi Lee et al., "Sensor monitoring using a fuzzy neural network with an automatic structure constructor," *IEEE Transactions on Nuclear Science*, vol. 50, no. 2, pp. 241–250, 2003.
- [16] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "Hybrid clustering analysis using improved krill herd algorithm," *Applied Intelligence*, vol. 48, no. 11, pp. 4047–4071, 2018.
- [17] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "A combination of objective functions and hybrid Krill herd algorithm for text document clustering analysis," *Engineering Applications of Artificial Intelligence*, vol. 73, pp. 111–125, 2018.
- [18] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "A new feature selection method to improve the document clustering using particle swarm optimization algorithm," *Journal of Computational Science*, vol. 25, pp. 456–466, 2018.
- [19] L. M. Abualigah, *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering*, Springer, Berlin, Germany, 2019.
- [20] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [21] S.-L. Wang, X.-L. Li, and J. Fang, "Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification," *BMC Bioinformatics*, vol. 13, no. 1, p. 178, 2012.
- [22] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," *Knowledge and Information Systems*, vol. 53, no. 3, pp. 551–577, 2017.
- [23] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 359–366, San Francisco, CA, USA, 2000.

- [24] H. C. Hanchuan Peng, F. H. Fuhui Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [25] L. Yu and H. Liu, "Feature selection for high-dimensional data: a fast correlation-based filter solution," in *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 856–863, Washington, DC, USA, 2003.
- [26] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 17, no. 3, pp. 300–304, 2004.
- [27] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
- [28] D. N. Reshef, Y. A. Reshef, H. K. Finucane et al., "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.
- [29] T. O. Kvalseth, "Entropy and correlation: some comments," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 17, no. 3, pp. 517–519, 1987.
- [30] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK, 1988.
- [31] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proceedings of the International Conference on Machine Learning*, pp. 284–292, Pittsburgh, PA, USA, 1996.
- [32] Z. Hua, J. Zhou, Y. Hua, and W. Zhang, "Strong approximate Markov blanket and its application on filter-based feature selection," *Applied Soft Computing Journal*, vol. 87, Article ID 105957, 2020.
- [33] Y. S. Chen, M. Lin, R. Yu, and T. S. Wang, "Research on simulation and state prediction of nuclear power system based on LSTM neural network," *Science and Technology of Nuclear Installations*, vol. 2021, Article ID 8839867, 11 pages, 2021.
- [34] International Atomic Energy Agency, *PCTTRAN Generic Pressurized Water Reactor Simulator Exercise Handbook*, IAEA, Vienna, Austria, 2019.
- [35] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello, "Minerva and minepy: a C engine for the MINE suite and its R, python and MATLAB wrappers," *Bioinformatics*, vol. 29, no. 3, pp. 407–408, 2013.
- [36] B. Liu and X. P. He, *Feature Selection of High-Dimensional Data: Theory and Algorithm*, Science Press, Beijing, China, 2016.
- [37] Z. H. Zhou, *Machine Learning*, Tsinghua University Press, Beijing, China, 2016.
- [38] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [39] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.