Two-Level Fault Diagnosis of SF6 Electrical Equipment Based on Big Data Analysis

Hongxia Miao 1, Heng Zhang 1,*, Minghua Chen 1, Bensheng Qi 1 and Jiyong Li 2

1 College of Internet of Things Engineering, HoHai University, Changzhou 213022, China; 19951482@hhu.edu.cn (H.M.); 1562510113@hhu.edu.cn (M.C.); qibensheng@hhu.edu.cn (B.Q.)
2 Electrical Engineering institute, Guangxi University, Nos.100, East University Road, Nanning 530004, China; jiyong_li@gxu.edu.cn
* Correspondence: hhuzhangheng@hhu.edu.cn

Received: 9 November 2018; Accepted: 21 December 2018; Published: 29 December 2018

Abstract: With the increase of the operating time of sulphur hexafluoride (SF6) electrical equipment, the different degrees of discharge may occur inside the equipment. It makes the insulation performance of the equipment decline and will cause serious damage to the equipment. Therefore, it is of practical significance to diagnose fault and assess state for SF6 electrical equipment. In recent years, the frequency of monitoring data acquisition for SF6 electrical equipment has been continuously improved and the scope of collection has been continuously expanded, which makes massive data accumulated in the substation database. In order to quickly process massive SF6 electrical equipment condition monitoring data, we built a two-level fault diagnosis model for SF6 electrical equipment on the Hadoop platform. And we use the MapReduce framework to achieve the parallelization of the fault diagnosis algorithm, which further improves the speed of fault diagnosis for SF6 electrical equipment.

Keywords: SF6 electrical equipment; Hadoop; fault diagnosis; parallelism

1. Introduction

SF6 electrical equipment refers to electrical equipment that uses sulphur hexafluoride (SF6) as insulation or arc extinguishing. SF6 electrical equipment has the advantages of small size, small maintenance, long life and good insulation. With the extensive popularization of intelligent substation, there are more and more power equipment [1]. The traditional oil-filled equipment is being replaced gradually by SF6 electrical equipment which have unique advantages and account for an increasing proportion of new equipment. With the increasing number of SF6 electrical equipment, the reliability requirements have also increased [2]. The internal discharge of SF6 electrical equipment leads to the decomposition of internal SF6 gas molecules. The chemical properties of its derivatives are very active. The derivatives can corrode the equipment, which can easily cause a decline in the insulation performance and cause serious damage to the safe and stable operation of the equipment [3,4].

In the literature [5], the phenomenon that the decision tree model or the radial basis function (RBF) neural network model is not used when the diagnosis is not high is firstly described. Then the RBF neural network and the decision tree model are used to diagnose the SF6 electrical equipment. This method requires two diagnostics for all data, which will take a lot of time.

The continuous creation of data has posed new research challenges due to its complexity, diversity and volume. Consequently, Big Data has increasingly become a fully recognized scientific field [6]. Now big data analytics are increasingly being used to solve real-world problems in life [7]. And the
cloud platform provides an open architecture for big data analytics, which can improve the utilization of data resources [8].

The arrival of the era of big data is accompanied by massive data, which makes the screening of valuable information a core step in the widespread application of big data. Hadoop is an open source distributed computing platform, its distributed file system HDFS and distributed computing framework MapReduce solved the problem of data storage and programming [9]. Hadoop’s ability to store and process data in bits is trusted. The backup mechanism is used in HDFS to maintain multiple copies of data. The task monitoring mechanism is used in MapReduce. Hadoop is able to dynamically move data between nodes and ensure the dynamic balance of each node, so processing is very fast [10].

MapReduce is a programming model and related implementation for processing and generating large data sets that are suitable for a wide variety of practical tasks [11]. MapReduce advantages over parallel databases include storage-system independence and fine-grain fault tolerance for large jobs [12]. At the same time, MapReduce can realize parallel processing of data and greatly improve the processing efficiency of monitoring data of SF6 electrical equipment [13].

In order to solve the above problems, the two-level fault diagnosis model of SF6 electrical equipment based on the content of SF6 gas derivative as the input for the monitoring data of SF6 electrical equipment was proposed in this paper. The first level fault diagnosis model is first used to determine whether the data is fault data. If yes, it enters the second level fault diagnosis model to identify the specific fault category. If not, the judgment is no longer performed, thereby improving the diagnostic efficiency. In order to quickly realize the fault diagnosis of SF6 electrical equipment for the massive SF6 electrical equipment condition monitoring data, this paper implements the fault diagnosis of SF6 electrical equipment on Hadoop platform and realizes the parallelization of SF6 electrical equipment fault diagnosis algorithm.

2. The Two-Level Fault Diagnosis Model for SF6 Electrical Equipment

2.1. Data Acquisition of SF6 Electrical Equipment

The two-level fault diagnosis model diagnoses the equipment by the content of SF6 gas derivatives. Therefore, it is necessary to select SF6 gas derivatives as the characteristic attributes first. When partial discharge occurs in SF6 electrical equipment, SF6 gas decomposes to produce a variety of derivatives. The formation mechanism of the derivatives is very complex, and mainly includes two processes. First, the SF6 gas decomposes to generate low fluorides, and then these low fluorides react with impurities, electrode materials, and insulating materials to generate other more stable derivatives. Derivatives derived from different discharge energies are also different. According to the magnitude of the released energy, discharge can be divided into three forms: arc, spark and corona discharge. Some references [14–16] mention that there is a large amount of SOF2 and a small amount of SO2F2 generated under arc discharge. The main derivative of spark discharge is also SOF2, in which the content of the derivative content is: \( \text{SOF2} > \text{SOF4} > \text{SO2F2} > \text{SO2} > \text{S2F10}/\text{S2OF10} \) [17]. During corona discharge, SOF2 is still the most abundant derivative, and the contents of SO2F2, S2OF10, and S2F10 are higher than those of the other two discharges. Therefore, SOF2, SOF4, SO2F2, SO2, S2OF10, and S2F10 were selected as the characteristic attribute of the algorithm [18]. And the types of faults in SF6 electrical equipment can be divided into three main types: arc, spark and corona discharge.

Infrared spectroscopy is a method for the quantitative and qualitative analysis of various infrared light absorbing compounds. The composition is analyzed by the difference in the absorption of infrared radiation by the substance. One of the great advantages of infrared spectroscopy is that virtually any sample may be studied in any state [19]. Infrared spectroscopy can be used to analyze the composition of SF6 gas in electrical equipment to determine the state of the equipment [20]. Infrared spectroscopy can be used to detect large quantities of material at a lower cost [21]. Therefore, we can install an infrared spectrum analyzer for gas composition analysis and data acquisition in SF6 electrical equipment.
2.2. Construction of two-level fault diagnosis model

The first level model is constructed using a random forest algorithm. The random forest model is used to filter out the fault data. Because it is only used to determine whether the equipment is faulty, the depth of the decision tree constructed in the forest is low, which can reduce the diagnosis time. The second level model is built by the neural network algorithm. The input was fault data filtered by a random forest model. The second-level model can not only diagnose the known types of faults, but also identify the new fault types through communicating with experts. And by updating the structure and weight of the neural network model, the model is continuously improved. Figure 1 is a block diagram of the SF6 electrical equipment fault diagnosis system.

The fault diagnosis process of SF6 electrical equipment is shown in Figure 2, and the main steps are follows.

1. Reading monitoring data, extract feature components and category components, then normalizing the feature components.
2. Putting the normalized data into the first-level random forest model. If the data is diagnosed as normal data, then output the result directly. Otherwise, go to step 3.
3. Putting the data that diagnosed as fault data in step 2 into the second-level neural network model. If the fault type of the equipment has been trained, then output the diagnosis result directly. Otherwise, go to step 4.
4. The fault type that cannot be correctly identified in step 3 is submitted to the experts for diagnosis to determine the type. Then the neural network is retrained to update the structure and weights of network so that it can be continuously improved.

![Figure 1](image1.png)

Figure 1. System block diagram for fault diagnosis of SF6 electrical equipment.

![Figure 2](image2.png)

Figure 2. Flow chart of two-level fault diagnosis algorithm.
The specific algorithm can refer to our paper: The two-level fault diagnosis model of SF6 electrical equipment [18].

Figure 3 is the architecture diagram of the secondary fault diagnosis system for SF6 electrical equipment.

![Architecture Diagram of the Two-Level Fault Diagnosis System for SF6 Electrical Equipment](image)

Figure 3. Architecture of the two-level fault diagnosis system for SF6 electrical equipment.

3. Monitoring Data Preprocessing

The fault diagnosis of SF6 electrical equipment is one of the methods to verify the operation status of the equipment, but the detection model of the equipment is mostly based on an ideal data set [22]. In the actual collection process, data redundancy, data loss, and data inconsistency will inevitably
occur due to issues such as collection equipment, external environmental disturbances, and human’s misoperation. This will eventually affect later data mining. Therefore, before the fault diagnosis of SF6 electrical equipment, this section fills in the missing values in the monitoring data to implement data pre-processing, and then implements the algorithm.

There are currently few data preprocessing studies on SF6 gas derivative monitoring data. In this section, the SF6 gas derivative content data is used as the processing object. Through the analysis of the data, it is found that there may be three cases of the missing monitoring data. A single approach cannot be fully applied to all situations. Therefore, the data missing values are classified, and different processing methods are adopted for each case.

There are three possible missing values for SF6 gas derivative content monitoring data. These three conditions are described as follows.

(1) The missing value is in the same range as the monitoring value recorded before and after it. There are two cases. Case 1: Monitoring data including missing values over a period of time is within normal limits. Case 2: Due to equipment failure, the monitored data including missing values over a period of time is in an abnormal range.

(2) The missing value is not in the same range as the monitoring value recorded before and after it. There are also two cases. Case 1: The device just fails and the SF6 gas derivative content rises from the normal range to the abnormal range. Case 2: The equipment was repaired and the content of SF6 gas derivatives dropped from the abnormal range to the normal range.

(3) The monitoring values are lost over a period of time.

For the above three cases, the process is as follows.

(1) The weighted interpolation method is used to deal with the case where the first type of data is missing, that is, the weighting factor is introduced for the mean interpolation method.

The basic mean interpolation method takes the arithmetic mean of the n numbers before and after the missing value as a substitute value, as shown in Equation (1).

\[ H_t = \frac{1}{2n} \left( \sum_{i=1}^{n} H_{t-i} + \sum_{i=1}^{n} H_{t+i} \right) \] (1)

The weighting factor is introduced for the mean interpolation method, as shown in Equation (2), wherein the record closer to the missing value time has a larger weight.

\[ H_t = \sum_{i=1}^{n} w_{t-i} H_{t-i} + \sum_{i=1}^{n} w_{t+i} H_{t+i} \] (2)

where \( w \) is the weight, the closer to the missing value, the greater the weight of the record.

There are six kinds of SF6 gas derivatives selected in this paper, and SOF2 is used as an example for pretreatment. In order to verify the validity of the algorithm, some SOF2 historical data was read from the database and created some missing cases. In the experiment, four monitoring values were chosen to estimate missing values. When SF6 electrical equipment is in normal condition, its SOF2 content does not exceed 10 \( \mu L/L \).

In this section, multiple sets of experiments were performed for each type of data loss. For reasons of space, only a part of the content was selected as the result.

Experiment 1: The historical data of SOF2 are within the normal range. The experimental data is shown in Table 1. The numerical values of the serial numbers 3, 4, and 5 are removed in turn, and the interpolation results of SOF2 are shown in Table 2.
Table 1. The historical data of SOF2.

<table>
<thead>
<tr>
<th>Id</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.4</td>
</tr>
<tr>
<td>2</td>
<td>6.8</td>
</tr>
<tr>
<td>3</td>
<td>6.5</td>
</tr>
<tr>
<td>4</td>
<td>6.9</td>
</tr>
<tr>
<td>5</td>
<td>6.7</td>
</tr>
<tr>
<td>6</td>
<td>5.9</td>
</tr>
<tr>
<td>7</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Table 2. The interpolation results of SOF2.

<table>
<thead>
<tr>
<th>Id</th>
<th>MIM</th>
<th>WIM</th>
<th>The Error of MIM</th>
<th>The Error of WIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>6.45</td>
<td>6.53</td>
<td>0.0077</td>
<td>0.0046</td>
</tr>
<tr>
<td>4</td>
<td>6.475</td>
<td>6.50</td>
<td>0.0616</td>
<td>0.0580</td>
</tr>
<tr>
<td>5</td>
<td>6.10</td>
<td>6.16</td>
<td>0.0896</td>
<td>0.0806</td>
</tr>
</tbody>
</table>

Note: MIM means mean interpolation method, WIM means weighted interpolation method.

Using the weighted interpolation method to calculate the missing values as follow:

\[ H_3 = w_1 H_1 + w_2 H_2 + w_3 H_4 + w_4 H_5 = 6.53 \]

where \( w_1 = w_4 = 0.2, w_2 = w_3 = 0.3 \).

According to Table 2, the average error of the mean interpolation method is 5.3%, and the average error of the weighted interpolation method is 4.8%. In this case, the weighted interpolation method works well.

Experiment 2: The historical data of SOF2 are within the abnormal range. The experimental data are shown in Table 3. The numerical values of the Nos. 3, 4, and 5 are removed in turn, and the interpolation results of SOF2 are shown in Table 4.

Table 3. The historical data of SOF.

<table>
<thead>
<tr>
<th>Id</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43.2</td>
</tr>
<tr>
<td>2</td>
<td>42.1</td>
</tr>
<tr>
<td>3</td>
<td>45.3</td>
</tr>
<tr>
<td>4</td>
<td>41.9</td>
</tr>
<tr>
<td>5</td>
<td>44.8</td>
</tr>
<tr>
<td>6</td>
<td>42.8</td>
</tr>
<tr>
<td>7</td>
<td>51.1</td>
</tr>
</tbody>
</table>

Table 4. The interpolation results of SOF2.

<table>
<thead>
<tr>
<th>Id</th>
<th>MIM</th>
<th>WIM</th>
<th>The Error of MIM</th>
<th>The Error of WIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>42.50</td>
<td>42.40</td>
<td>0.0618</td>
<td>0.0640</td>
</tr>
<tr>
<td>4</td>
<td>42.65</td>
<td>42.93</td>
<td>0.0179</td>
<td>0.0246</td>
</tr>
<tr>
<td>5</td>
<td>44.675</td>
<td>43.97</td>
<td>0.0438</td>
<td>0.0273</td>
</tr>
</tbody>
</table>

According to Table 4, the average error of the mean interpolation method is 4.12%, and the average error of the weighted interpolation method is 3.86%. The weighted interpolation method works well.

(2) Reading the SOF2 monitoring data from the database for a period of time. The missing value is not in the same range as the monitoring value recorded before and after it.

Experiment 3: When the content of the previous monitoring data of SOF2 missing value is within 10 \( \mu L/L \), and the content of the latter monitoring data is greater than 10 \( \mu L/L \), it indicates that the content of SOF2 fluctuates from the normal range to the abnormal range. The experimental data is shown in Table 5. If the experiment is continued using the weighted interpolation method and the mean interpolation method, the numerical values of the serial numbers 3 and 4 are sequentially removed, and the experimental results are shown in Table 6.
Table 5. The historical data of SOF2.

<table>
<thead>
<tr>
<th>Id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>8.8</td>
<td>8.3</td>
<td>8.6</td>
<td>12.9</td>
<td>16.2</td>
<td>15.3</td>
<td>18.2</td>
</tr>
</tbody>
</table>

Table 6. The interpolation results of SOF2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Id</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIM</td>
<td>11.55</td>
<td>12.10</td>
<td></td>
</tr>
<tr>
<td>WIM (weight: 0.2, 0.3)</td>
<td>11.36</td>
<td>12.16</td>
<td></td>
</tr>
<tr>
<td>WIM (weight: 0.125, 0.375)</td>
<td>11.07</td>
<td>12.25</td>
<td></td>
</tr>
<tr>
<td>WIM (weight: 0.125, 0.375)</td>
<td>0.3430</td>
<td>0.0620</td>
<td></td>
</tr>
<tr>
<td>The error of WIM (weight: 0.2, 0.3)</td>
<td>0.3209</td>
<td>0.0574</td>
<td></td>
</tr>
<tr>
<td>The error of WIM (weight: 0.125, 0.375)</td>
<td>0.2872</td>
<td>0.0539</td>
<td></td>
</tr>
</tbody>
</table>

It can be seen from Table 6 that although the error of the interpolation can be reduced by adjusting the weight, it is necessary to find a suitable weight. For sequence number 3, the method of linear interpolation is not suitable.

(3) When the SOF2 monitoring data is read from the database for a period of time, there is a continuous lack of data, and the linear interpolation method cannot be used at this time.

Considering case 2 and case 3, this paper uses the gray correlation degree to interpolate the data. The SF6 gas derivative content data complementing method based on gray correlation is to perform gray scale processing on other component data except the missing value attribute, and to find the closest set of data with the missing value tuple by calculating the correlation degree to make up the complement treatment. The main process is shown in Figure 4.

Figure 4. The data interpolation process based on method of grey correlation.

The specific steps are follows:

Step 1: Determining the main sequence and subsequence.

A tuple containing missing values is taken as the main sequence. A tuple, in the historical record, with at least one attribute value that is not within the normal range of the safety work is taken as a subsequence. Mark them as $X_1 \sim X_m$ and $m = 5000$.

Step 2: The matrix formed by the main sequence and subsequences is standardized to obtain a normalized matrix $X$.

Step 3: Calculate correlation coefficient.

The formula for calculating the correlation coefficient $rel(i)$ of $X_i$ for $X_0$ is:

$$rel(i) = \frac{\min AbsValue(i) + def^* \max AbsValue(i)}{absValue(i) + defC^* \max AbsValue(i)}$$

$$absValue(i) = abs(X_i - X_0)$$
where \( \text{absValue}(i) \) is the absolute difference sequence of \( X_i \) and \( X_0 \), \( \text{minAbsValue}(i) \) and \( \text{maxAbsValue}(i) \) are the minimum and maximum of \( \text{absValue}(i) \), \( \text{def}C \in (0, 1) \) is the resolution coefficient and \( \text{def}C = 0.5 \).

**Step 4:** Calculate relevance.

The formula for calculating relevance \( p(i) \) of \( X_i \) for \( X_0 \) is:

\[
p(i) = \text{rel}(i) \times w
\]

where \( w \) is the weight of each SF6 gas derivate. There are six kinds of derivatives, so the \( w \) is set to 1/6.

**Step 5:** Interpolated missing values.

Sorting the relevance obtained above, and extracting the first 10 sets of tuples those relevance are greater than 0.9 are taken to interpolate the missing values. If less than 10 sets, all the tuples satisfying the condition will be taken. The formula for interpolating missing values \( X_{0j} \) is:

\[
X_{0j} = \frac{1}{n} \sum_{i=1}^{n} X_{ij}
\]

where \( j \) is the column of the missing value in the tuple, \( n \) is the number of tuples taken.

Experiment 3 was re-examined using the gray correlation degree. The experimental results are shown in Table 7.

<table>
<thead>
<tr>
<th>Id</th>
<th>The Top 10 Values of Relevance</th>
<th>Interpolated Value</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>8.2, 8.5, 9.1, 8.3, 8.6, 9.5, 7.9, 8.2, 8.5</td>
<td>8.51</td>
<td>0.01</td>
</tr>
<tr>
<td>4</td>
<td>10.8, 11.5, 13.4, 12.3, 14.2, 15.6, 11.8, 12.9, 17.2, 13.5</td>
<td>13.32</td>
<td>0.033</td>
</tr>
</tbody>
</table>

From Table 7, it can be seen that the interpolated values for Nos. 3 and 4 are 8.51 and 13.32, the errors are 0.01 and 0.033 respectively. Therefore, the use of gray correlation can effectively solve the problem of missing data mutations.

The method of grey correlation is to find the closest tuples in the previous record to interpolate missing data. Therefore, when this method is used, there is no case of continuous lack of data.

After pre-processing the monitoring data of the SF6 electrical equipment, a two-level fault diagnosis model of the SF6 electrical equipment can be constructed. First, we train the model on the Hadoop platform. Second, we implemented parallelization of diagnostic algorithms on the Hadoop platform.

### 4. Implementation of the Two-Level Fault Diagnosis of SF6 Electrical Equipment Based on Hadoop

Hadoop is an open source distributed computing platform. Its distributed file system (HDFS) and distributed computing framework (MapReduce) solve the data storage and programming problems, respectively. Therefore, we implemented the parallelization of the above fault diagnosis algorithm on the Hadoop platform, which improved the processing speed of massive monitoring data and accelerated the diagnostic rate of SF6 electrical equipment.

#### 4.1. Implementation of Random Forest Algorithm Based on MapReduce

After preprocessing the SF6 electrical equipment monitoring data, a two-level fault diagnosis model of the SF6 electrical equipment can be constructed on the Hadoop platform, and the parallel algorithm of the diagnostic algorithm can be realized through the Hadoop platform.
In the process of establishing a random forest, each decision tree is created in a serialized manner. Only when the current decision tree is generated will the next tree be created. Decision trees are independent of each other, and there is no need to rely on other trees when creating a decision tree, so parallelization can be achieved.

Using the MapReduce framework to achieve the parallelization of the random forest algorithm mainly has two stages: Map and Reduce. Each Map task establishes a decision tree, and finally uploads them to HDFS in the Reduce task to form a forest. The parallel forest construction process of random forest is shown in Figure 5 and the specific tree construction process of Map task is shown in Figure 6.

**Figure 5.** The parallel forest construction process of random forest.

**Figure 6.** The specific tree construction process in Map task.

Specific steps are as follows:

**Step 1:** Create a sample subset. Using Bagging method to extract subsets from the original sample as the sample subset of each decision tree, where treelD is the number of the decision tree and
dataset is its corresponding sample subset. In this paper, the number of trees is set to 7. So the range of the treeID is 1 to 7.

**Step 2:** Use the sample subset to create a decision tree and initialize the number of Map tasks based on the number of decision trees.

The input of the Map function is \(<treeID, dataset>\). This function mainly completes the decision tree construction, and the output is \(<treeID, list<feature>>\). MapReduce parallelism is also used to select the splitting attribute of the node. Every time the non-leaf node selects the split attribute, it needs to calculate the Gini value of the remaining feature attributes, and return the best split attribute and its value by comparison. The pseudo code is shown in Figures 7 and 8.

**Step 3:** All the Map tasks are completed, which means that the construction of the decision tree has been completed. At this time, the Reduce task is executed, and the split rule of each decision tree is written into HDFS to obtain a random forest classifier.

```
/// Map Task ///
input: <treeID,dataset> output: <treeID,list<feature>>
treeID: the Id of the decision tree
Dataset={X,Y}, X: the sample subset of n*m matrices,
   Xij: the jth attribute value of the i-th row in X
Y: the decision attribute vector, Yi: the output class of the i-th row data.

If all sample in dataset belong to the same class then
   return a leaf node;
End if
For j=1:m
   %call the function that calculate the Gini value
   calculateGini();
   j* = minGini;
   %Divide the dataset into two subsets
   XL = {X|Xij<a*}, YL=decision attribute of XL, datasetL=(XL,YL);
   XR = {X|Xij>=a*}, YR=decision attribute of XR, datasetR=(XR,YR);
End for
Return list<feature>; % record split attributes and split values
```

*Figure 7.* The pseudo code to construct a decision tree in Map task.

```
/// Calculate the Gini value ///
function: Calculate()
% j:split attribute, tmp: split value
For j=1:n-1
   tmp = (Xij+Xij+1)/2;
   Gini(j,tmp);
End for
Return min(Gini);
```

*Figure 8.* The pseudo code to calculate Gini value.
4.2. Implementation of Neural Network Algorithm Based on MapReduce

Parallelization of the back propagation (BP) neural network model achieves parallelization of data processing. Multiple compute nodes are deployed on the Hadoop platform, and each node has a complete BP neural network model to process part of the sample data, so the calculation of each node is parallelized. Training the neural network model on the Hadoop platform includes the Map, Reduce, and Combine stages, as shown in Figure 9.

![Figure 9. The process of training back propagation (BP) neural network model.](image)

① Map stage

In the Map stage, the setup() function reads the initial network weights from the file system HDFS and initializes the neural network. The Map() function reads the sample data and trains on that node. After the set condition is reached (if the set number of iterations is reached, or the output error reaches the set value, the number of iterations reached in this section is used), and the model training ends. The pseudo code of Map is shown in Figure 10.

```
///bp Mapper ///
Input: <key, value>
Output: <key, weightWritable>
Setup()
    read the network weights file from HDFS;
}
Map(key, value)
    while(!meet the max iterations)
        % Calculate the weight change for each sample
        calculateWeight();
    }
    % Instantiate weightWritable to record weight changes
    new weightWritable;
    context.write(key, weightWritable);
}
Cleanup()
    Release resources;
```

![Figure 10. The pseudo code of Map stage.](image)

② Combine stage

Combine is used to merge the results of the Map. The input is the output of the Map, and its output is the input of Reduce. The type is the same as the output type of Map.
Reduce stage

In the Reduce stage, the <key, weightWritable> output of the Combine stage is used as the input of Reduce stage. The Reduce() function is to calculate the average value of the value for each key, and compare it with the network weight stored on the HDFS to determine whether to perform the next loop operation, and finally re-write the updated weights to HDFS. The pseudo code of Reduce stage is shown in Figure 11.

```
/////////    bp Reducer   /////////
Input: <key, weightWritable>
Output: <key, 1 or 0>
Setup()
    read the network weights file from HDFS; }
Map(key,value)
    %Calculate the average value of weightWritable for each sample
    while(weightWritable has not been processed)
        total = sum(weightWritable);
    }
    ave = average (total);
    %Compare the initial weights and the updated weights
    combine(weightFile,ave );
    upload the updated weights to HDFS;
}
Cleanup()
    Release resources;
```

Figure 11. The pseudo code of Reduce stage.

4.3. Fault Diagnosis Experiment of SF6 Electrical Equipment

4.3.1. Experimental Programming

Two MapReduce tasks are set up on the Hadoop platform to diagnose SF6 electrical equipment, as shown in Figure 12. In the first task MR1, the random forest model is used for initial diagnosis. In the second task MR2, the neural network model is used for diagnosis of fault categories.

Figure 12. The diagnosis process of SF6 electrical equipment.
MR1 is mainly divided into Map and Reduce stages, namely:

(1) Map stage

1. Setup () loads the random forest file in HDFS.
2. Enter the sample to be diagnosed for the first level of diagnosis and statistically determine the diagnosis result of the decision tree.
3. Output <key, value>, where the key is the flag of whether the sample may be fault data. If the key is 1, it means the sample is normal data. If the key is 2, it means that the majority of the decision tree is abnormal for the diagnosis of the sample; the value is the sample data and the diagnosis result.

(2) Reduce stage

1. Enter the output of the Map stage.
2. Use different files for different key values to save, where the key is 1 and the data is saved in a file named normal, no further processing is required. The data with key 2 is stored in a file named input2 as the input file for the second job.

The MapReduce process of MR2 is as follows:

(1) Map stage

1. Use the setup () to initialize the met matrix object, import the weight matrix of each layer of the trained neural network, and build a neural network model.
2. The input data in map () is the input2 file data of MR1, and the sample is judged to obtain the diagnosis result. The reliability of the diagnosis result is calculated.
3. Output <key, value>, where the key is the flag for successful diagnosis. If the key is 1, it indicates that the reliability of the diagnosis result of the sample reaches the set credibility threshold requirement, otherwise the credibility threshold is not reached; the value is the sample data and the diagnosis result.
4. Close the met object in cleanup ()

(2) Reduce stage

1. Enter the output of the Map stage.
2. Save different files for different key values, where a key of 1 means no further processing is required and the result is saved directly. Data with a key of 2 requires expert re-diagnosis and is stored in a file called to Expert.

4.3.2. Experimental Results and Analysis

(1) Verify the validity of the model.

There are three types of faults in SF6 electrical equipment: arc, spark and corona discharge. In this experiment, normal data and data including arc and corona discharge are sent to the model for training, and then the monitoring data including spark discharge is diagnosed.

1. Random forest model results.

First, 70% of the 2000 samples (10% of which are fault data and the rest are normal data) are used to train the random forest model, and 30% (including 60 sets of data is fault data) are used to test the model. After testing, the results are shown in Table 8, where category 1 is fault data and category 2 is normal data.
Table 8. The test results of random forest model.

<table>
<thead>
<tr>
<th>Target Category 1</th>
<th>Target Category 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of diagnostic category 1</td>
<td>60</td>
</tr>
<tr>
<td>Number of diagnostic category 2</td>
<td>0</td>
</tr>
</tbody>
</table>

As can be seen from Table 8, one piece of data is normal, but it is classified as possible fault data. All fault data can be identified. From the results of the diagnosis, the random forest model can distinguish between normal and fault data.

2 Neural network model results.

70% of the samples were used to train the neural network model, and 30% of the samples were used to test the model. Some diagnostic results are shown in Table 9.

Table 9. Some diagnostic results of neural network model.

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>Normal</th>
<th>Corona Discharge</th>
<th>Arc Discharge</th>
<th>Credibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>0.0031</td>
<td>−0.0019</td>
<td>0.9914</td>
<td>0.9950</td>
</tr>
<tr>
<td>61</td>
<td>0.0012</td>
<td>0.0051</td>
<td>0.9950</td>
<td>0.9937</td>
</tr>
<tr>
<td>64</td>
<td>0.0063</td>
<td>1.0019</td>
<td>0.0013</td>
<td>0.9925</td>
</tr>
<tr>
<td>99</td>
<td>0.0016</td>
<td>0.9933</td>
<td>0.0074</td>
<td>0.9910</td>
</tr>
</tbody>
</table>

The reliability of the sample diagnosis shown in Table 9 is greater than the set confidence threshold (threshold is 0.8), so the neural network model is considered to be able to determine the diagnosis. In practice, the samples numbered 58 and 61 are the monitoring data when corona discharge occurs, and the samples numbered 64 and 99 are the monitoring data when the arc discharge occurs, so the diagnosis result is correct.

3 Monitoring data diagnosis results and analysis.

After training the model, 10,010 groups (10,000 normal data and 10 sets of fault data) to be diagnosed into the diagnostic model, after the random forest diagnosis model, all 10 groups of fault data are identified, and after the neural network model is substituted, the diagnosis is made. The results are shown in Table 10.

Table 10. The results of neural network model.

<table>
<thead>
<tr>
<th>Number</th>
<th>Normal</th>
<th>Arc Discharge</th>
<th>Corona Discharge</th>
<th>Credibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−0.0015</td>
<td>0.0019</td>
<td>0.9988</td>
<td>0.9966</td>
</tr>
<tr>
<td>2</td>
<td>0.0024</td>
<td>−0.0073</td>
<td>0.9982</td>
<td>0.9904</td>
</tr>
<tr>
<td>3</td>
<td>0.0056</td>
<td>−0.0026</td>
<td>0.9988</td>
<td>0.9919</td>
</tr>
<tr>
<td>4</td>
<td>−0.0021</td>
<td>−0.0074</td>
<td>0.9872</td>
<td>0.9905</td>
</tr>
<tr>
<td>5</td>
<td>0.0179</td>
<td>0.6740</td>
<td>0.3243</td>
<td>0.6633</td>
</tr>
<tr>
<td>6</td>
<td>0.0261</td>
<td>0.5271</td>
<td>0.4721</td>
<td>0.5141</td>
</tr>
<tr>
<td>7</td>
<td>−0.0041</td>
<td>1.0012</td>
<td>0.0019</td>
<td>0.9940</td>
</tr>
<tr>
<td>8</td>
<td>−0.0011</td>
<td>1.0008</td>
<td>0.0085</td>
<td>0.9905</td>
</tr>
<tr>
<td>9</td>
<td>0.0063</td>
<td>1.0010</td>
<td>0.0087</td>
<td>0.9852</td>
</tr>
<tr>
<td>10</td>
<td>0.0019</td>
<td>1.0009</td>
<td>0.0047</td>
<td>0.9966</td>
</tr>
</tbody>
</table>
It can be seen from Table 10, the credibility of samples 5 and 6 is less than the threshold (threshold is 0.8), so they are sent to experts for evaluation. They have been identified as a new fault type—spark discharge. Then the neural network model is retrained and the network weights are updated so that the model can identify the type of spark discharge. The diagnosis results of the updated neural network model for samples Nos. 5 and 6 are shown in Table 11.

Table 11. The diagnosis results of the updated neural network model.

<table>
<thead>
<tr>
<th>Number</th>
<th>Normal</th>
<th>Spark Discharge</th>
<th>Arc Discharge</th>
<th>Corona Discharge</th>
<th>Credibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.0019</td>
<td>0.9972</td>
<td>0.0089</td>
<td>0.0136</td>
<td>0.9761</td>
</tr>
<tr>
<td>6</td>
<td>0.0008</td>
<td>0.9988</td>
<td>−0.0088</td>
<td>0.0019</td>
<td>0.9886</td>
</tr>
</tbody>
</table>

It can be seen from Table 11 that the reliability of the sample diagnosis results of the numbers 5 and 6 is greater than threshold, that is, the updated neural network model can accurately diagnose the new fault type.

(2) Experiment 2: Comparison of fault diagnosis performance between Stand-alone and cluster mode.

The speedup $S$ is usually used to measure the performance of parallel algorithms, which are defined as follows:

$$S = \frac{T_s}{T_m}$$

(7)

where $T_s$ is the time taken for diagnosis by one node, $T_m$ is the time taken for nodes to perform parallel operations.

The datasets of different sizes are run in the cluster mode of the Hadoop platform. The diagnostic times under different slave nodes are shown in Table 12.

Table 12. The diagnostic times under different slave nodes.

<table>
<thead>
<tr>
<th>Number</th>
<th>10 MB</th>
<th>100 MB</th>
<th>500 MB</th>
<th>1 GB</th>
<th>2 GB</th>
<th>3 GB</th>
<th>4 GB</th>
<th>5 GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Slave</td>
<td>10.2</td>
<td>15.2</td>
<td>35.4</td>
<td>60.2</td>
<td>110</td>
<td>165.3</td>
<td>215</td>
<td>265.5</td>
</tr>
<tr>
<td>2 Slave</td>
<td>19.6</td>
<td>21</td>
<td>35</td>
<td>54</td>
<td>80</td>
<td>116.3</td>
<td>148</td>
<td>178.6</td>
</tr>
<tr>
<td>3 Slave</td>
<td>21</td>
<td>21</td>
<td>25</td>
<td>39.6</td>
<td>47</td>
<td>69.2</td>
<td>88.5</td>
<td>104.2</td>
</tr>
</tbody>
</table>

In order to more intuitively compare the speed of the fault diagnosis algorithm, the Table 12 is converted into a histogram as shown in Figure 13. It can be seen from Figure 13 that when the size of data to be diagnosed is less than 500 MB, the time consumed by single node and multiple nodes is similar. But as the size increases, the time of the fault diagnosis system in a stand-alone model is longer than cluster model. Therefore, when the size of the data that needs to be diagnosed reaches a certain level, using Hadoop’s cluster mode can greatly improve the speed of fault diagnosis.

Through calculation, the speedup of fault diagnosis under different node numbers is shown in Table 13. It can be seen that when the amount of data to be diagnosed is small (less than 500 MB in the experiment), the acceleration ratio of the diagnosis is less than 1, indicating that the operating efficiency of cluster mode is lower than the stand-alone at this time. However, when the amount of data increases, with the increase in the number of nodes, the speedup also increases.

In order to further compare the influence of the number of nodes on the diagnosis speed, the experiment was carried out for the case where the data amount was 5 GB, and the result is shown in Figure 14.
problematic data in a large amount of data. The second-level diagnostic model provides an in-depth analysis of the fault types of SF6 electrical equipment and can be learned in real time to improve the presentation.

Engineers can check the operating status of SF6 electrical equipment in real time on the client side, which saves time in field inspections. Engineers can quickly and accurately identify problems to avoid security incidents.

In the future, smart substations can be equipped with SF6-derived gas composition detection equipment (such as infrared spectrum analyzers) and upload data to cloud databases in real time.

The first-level diagnostic model can quickly diagnose monitoring data and filter out the problematic data in a large amount of data. The second-level diagnostic model provides an in-depth analysis of the fault types of SF6 electrical equipment and can be learned in real time to improve the presentation.

As can be seen from Figure 14, when the amount of data is 5 GB, the speedup does not increase linearly with the increase in the number of nodes. This is because as the cluster size increases, the time spent communicating between nodes increases. Therefore, in practice, the size of the cluster can be selected according to the size of the amount of data.

5. Conclusions

In this paper, a two-level fault diagnosis model for SF6 electrical equipment is designed, and fault diagnosis for SF6 electrical equipment is implemented on the Hadoop platform in this paper. Before training the fault diagnosis model, the monitoring data is preprocessed first, and different data filling methods are adopted for different missing values. Secondly, the fault diagnosis algorithms are parallelized on the Hadoop platform. Finally, the time consumption of fault diagnosis of SF6 electrical equipment in stand-alone mode and cluster mode is compared by simulation, and the advantages of cluster mode in processing massive data are verified.

The first-level diagnostic model can quickly diagnose monitoring data and filter out the problematic data in a large amount of data. The second-level diagnostic model provides an in-depth

Table 13. The speedup of fault diagnosis under different node numbers.

<table>
<thead>
<tr>
<th>Speedup</th>
<th>10 MB</th>
<th>100 MB</th>
<th>500 MB</th>
<th>1 GB</th>
<th>2 GB</th>
<th>3 GB</th>
<th>4 GB</th>
<th>5 GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Slave</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 Slave</td>
<td>0.52</td>
<td>0.72</td>
<td>1.01</td>
<td>1.11</td>
<td>1.38</td>
<td>1.42</td>
<td>1.45</td>
<td>1.49</td>
</tr>
<tr>
<td>3 Slave</td>
<td>0.48</td>
<td>0.72</td>
<td>1.41</td>
<td>1.52</td>
<td>2.34</td>
<td>2.39</td>
<td>2.43</td>
<td>2.55</td>
</tr>
</tbody>
</table>

Figure 13. The diagnostic times under different slave nodes.

Figure 14. The speedup of fault diagnosis under different slave nodes.

The authors declare no conflict of interest.
analysis of the fault types of SF6 electrical equipment and can be learned in real time to update the fault type library.

In the future, smart substations can be equipped with SF6-derived gas composition detection equipment (such as infrared spectrum analyzers) and upload data to cloud databases in real time. Engineers can check the operating status of SF6 electrical equipment in real time on the client side, which saves time in field inspections. Engineers can quickly and accurately identify problems to avoid security incidents.

**Author Contributions:** All authors have equally contributed substantially to the work reported. Conceptualization, H.M.; Data curation, H.Z.; Formal analysis, H.Z.; Funding acquisition, B.Q.; Investigation, M.C.; Methodology, H.M.; Project administration, H.M.; Resources, M.C. and J.L.; Software, H.Z. and J.L.; Supervision, H.M.; Visualization, H.Z. and B.Q.; Writing—original draft, H.Z.; Writing—review and editing, M.C.

**Funding:** This paper is supported by the national Natural Science Foundation of China (51607057). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

2. Liu, M. *Electrical Device Failure Diagnosis Research Based on the Analysis of the SF6 Gas Decomposition*; Hunan University: Changsha, China, 2013.


