Research on Fault Prevention and Maintenance System of Automatic Substation Primary Equipment Based on Decision Tree Algorithm

Xinyue Wang†

1. Henan University of Technology, Jiaozuo, Henan, 454003, China

Abstract

Electric power enterprises are developing rapidly in the era of big data information digitization. At this stage, the total number of substations is gradually increasing, the structure of the power engineering system is slowly becoming complicated, and the video monitoring system instantly collects a lot and contains a lot of noisy data information, which affects the power supply system’s access to effective data information and fault detection. To prevent the above phenomenon, this paper selects a decision tree algorithm to obtain and analyze meaningful operation-confirming information from a large amount of data information, and then can quickly and confirm the diagnosis of common fault machines and equipment in substations, reduce the running time of common fault machines, and improve the safety and reliability of primary equipment in substations with automation technology. The paper describes the basic concept of big data mining common algorithm and its data mining algorithm in the automation technology substation primary equipment fault detection, selected the typical alarm signal to start the analysis, and categorization and collocation solution. A decision tree algorithm entity model is built, several classical decision tree algorithms are described, and their data analysis is carried out for each attribute, and then the decision tree algorithm is improved. According to build the decision tree algorithm according to improve the decision tree algorithm under the fuzzy set base theory, mainly by expertise in the four on cannot identify the association, rough set and up close and down close, similar and membership relationship, expertise concise to optimize the calculation method. And the common ID3, C4.5, and CRAT algorithm of each property is compared and analyzed, and the results show that: compared with C4.5 and ID3, the boosted optimization algorithm has higher classification accuracy and can model rate more quickly. The research in this paper can quickly diagnose the automation technology substation primary equipment and fault phenomena, and its establishment of the whole process is easy, the scope of application is relatively high, and it has wide applicability.

Keywords: Big data; Decision tree algorithm; Fuzzy set theory; Automated substation; Maintenance system

AMS 2020 codes: 78-02
1 Introduction

In recent years, data mining algorithms have long been widely used in the automation technology substation primary equipment condition assessment industry because they have the advantage of solving large amounts of data and rapid response time in building models [1]-[2]. Automation technology substation primary equipment needs to provide many kinds of information in the operation and operation and maintenance management and repair process, such as the operating status of machines and equipment, experimental data, safety inspection data information, and status statistics [3]-[4]. With the application and promotion of power information systems such as power production management systems, the data related to the primary equipment of automated substations have been effectively collected and stored [5]-[6]. Knowledge reflecting the state deterioration process of the primary equipment of the automated substation is implied in these data, but it is difficult to extract relevant knowledge by manual statistical analysis of these data because more state quantities can reflect the state change of the primary equipment of the automated substation, and the state change mechanism is more complicated [7]-[8]. The influence of these objective factors and the disadvantage of the manual approach’s low efficiency leads to poor knowledge acquisition work quality and makes it difficult to extract knowledge more quickly and comprehensively [9]-[10]. Therefore, a method capable of efficient condition assessment knowledge acquisition for fault prevention data of primary equipment in automated substations is needed [11]-[12]. To ensure that the maintenance work is carried out as expected, it is necessary to explore in depth the equipment maintenance control work and to master the maintenance principles and the specific equipment maintenance control methods [13]. To truly prevent problems before they occur and effectively deal with various hidden problems in the equipment to achieve good equipment operation [14]-[15].

Reference [16] built the software for MapReduce distributed processing system on the Hadoop platform according to Hadoop. The cloud computing technology was completed based on selecting the C4.5 decision tree algorithm, and the algorithm’s improvement was implemented based on MapReduce. The parallel processing algorithm is the key to enhancing the parallel processing method on the Hadoop platform to complete the parallel computing of the database. The experimental results show that the method proposed in this paper can significantly improve the data processing efficiency, enhance the accuracy of fault diagnosis and identification of power engineering equipment, and reasonably complete the real-time detection of electrical equipment in intelligent substations. The literature [17] proposes the knowledge indication and acquisition method according to the decision tree algorithm for the defects of the traditional knowledge indication and acquisition method. This method flexibly uses the advantage of the decision tree algorithm to integrate knowledge indication and acquisition into one, so that knowledge representation and knowledge acquisition are carried out simultaneously and solve the defects of knowledge indication and knowledge acquisition separated in the traditional type artificial intelligence system. It is used in the acquisition and representation of substation fault detection knowledge. The experimental results show that the proposed approach can not only accomplish automatic knowledge acquisition and representation but also that the obtained knowledge expressed in the decision tree algorithm has high efficiency of logical reasoning. In reference [18], an accurate identification method of substation protective switches based on an improved decision tree algorithm is given to address the challenges of poor anti-interference and accuracy of current protective switch positioning and identification methods in complicated scenarios, which are prone to severe dust adhesion and lubricant dryness after long-term operation of substation protective switches. Under the premise of fusing the decision tree algorithm and theoretical Hough transform to complete the accurate localization of protective switches, the in-work state of protective switches is screened according to the close combination of the decision tree algorithm and particle swarm algorithm, and the substation protective switches are fully prepared for power cleaning. Reference [19] took the monthly safety operation condition of Jinggu County substation for three
years from 2012 to 2017 as the subject, analyzed the practical significance of power engineering
weather monitoring, and carried out relatively deep research and completion on the issue of tree
building time efficiency with the algorithm in the decision tree algorithm, and made corresponding
improvements, and by improving the ID3 algorithm according to The decision tree model for power
engineering weather monitoring algorithm was set up, and an experimental report was made to show
that the computational characteristics of the power engineering weather monitoring algorithm are
extremely strong. Reference [20] proposed the knowledge acquisition method of distribution
transformer condition assessment based on the generation of very few oversampling technicalities
(SMOTE) algorithms and decision tree algorithm and the first for the condition that the sample size
of the abnormal condition of the transformer is small, the SMOTE algorithm is chosen to fill the
sample size of the abnormal condition, which overcomes the problem of the unbalanced sample set
category of the transformer.

Then the whole process of transformer state assessment is regarded as the whole process of
categorization, and the decision tree model is used as a white box model to transform the problem of
obtaining knowledge of transformer state assessment into a problem of building a decision tree
algorithm. Finally, the c4.5 decision tree algorithm is chosen to build the decision tree algorithm,
from which the substation condition assessment knowledge is obtained, and important substation
condition quantities and assessment criteria are obtained. The literature [21] gives a rough set-based
fault diagnosis method for substation maintenance machinery and bionic equipment technology. The
algorithm consists of four important components: substation sub-region planning method, rough set
characteristic reduction algorithm, binary logic inference monopulse neural system membrane system
(BRSNPS), and parallel processing logic inference algorithm. In general, the substation sub-region
planning method and rough set simplification algorithm are chosen to explore the problem of
simplification for each sub-region resulting in a rule set that optimizes the problem multiplicity and,
simultaneously, can solve the uncertainty of fault warning information content.

In this paper, we analyze the software of the substation’s primary mechanical fault prevention and
maintenance system for automation technology based on a decision tree algorithm, build a decision
tree algorithm in data mining algorithm, discuss several types of classical decision tree algorithms
and carry out data analysis for each characteristic, and implement improvement of the decision tree
algorithm entity model, mainly from the knowledge and indiscernible association, rough set and up
and down approximation, similarity and membership, and simplicity of knowledge under fuzzy set
base theory. The final sample data is selected, and the improved decision tree algorithm entity model
is certified based on four aspects of fuzzy set grounded theory mainly from the knowledge and
indistinguishable association, rough set and up and down approximation, similarity and membership
relationship, and knowledge conciseness, and the results show that the optimized algorithm is better
than the traditional algorithm in many aspects such as classification accuracy and model rate. The
improved decision tree algorithm was then used in the automation technology substation primary
mechanical fault and maintenance system software, significantly improving the speed and accuracy
of the automation technology substation primary mechanical fault and maintenance across the board.

2 Decision tree algorithm framework

Knowledge acquisition is removing redundant information and acquiring effective data from a large
amount of data. The decision tree optimization algorithm is used in the paper to sort out the primary
equipment condition assessment knowledge from the automation technology distribution station’s
primary equipment statistics to obtain knowledge accurately and quickly. The conclusion of
transformer condition assessment applies state level to carry out the indication, divided into four
types: normal state, attention state, abnormal state, and more serious state. Thus the whole process of
primary equipment condition assessment is classified as a classification problem. The decision tree algorithm is a kind of random forest algorithm with a relatively small operation volume and high categorization precision, which is suitable for solving the whole process of primary equipment condition assessment. Because the decision tree entity model is a white box model, it is possible to obtain the management process from this.

The research perspective of this paper is divided into 3 main parts: formation of a sample set, construction of a decision tree of primary equipment condition assessment knowledge, and knowledge acquisition based on the decision tree. The section on sample set formation describes the construction of the sample set and its processing; the section on primary equipment condition assessment knowledge decision tree construction describes the selection of the decision tree algorithm and the basic theoretical process of constructing the primary equipment condition assessment knowledge decision tree model; the section of knowledge acquisition based on decision tree describes the whole process of knowledge acquisition from within the constructed decision tree.

2.1 Generation of the sample set

The data information mainly comes from manual entry and automatic classification of monitoring devices, and the quality of data information varies greatly, with many problems such as missing data and cache overflow. To improve the data quality, the collected information is subjected to data cleaning, data transformation, data aggregation, and other processes to produce the initial sample set of primary equipment of the automation technology distribution station. Therefore, the samples of the normal state in the initial sample set of the primary equipment of the automation technology distribution station far exceed the samples of the attention state, abnormal state, and more serious state. The imbalance in the number of samples of each state will make the initial sample set of automation technology distribution station primary equipment occur type imbalance problem. That is, the number of samples of the normal state far exceeds the number of samples of the abnormal state, which may lead to the random forest algorithm in the screening discrimination, focusing on ignoring the abnormal state samples, resulting in categorization characteristics reduced.

2.2 Decision tree ID3 algorithm

The ID3 algorithm is based on information theory and uses information entropy and information gain as evaluation indexes to sort and categorize the database, and the ID3 optimization algorithm selects attributes with maximum information gain and maximum entropy drop as detection attributes. The specific approach is shown below:

Each of the N labeled patterns belongs to the set of patterns in category $c_i$, $i = 1, 2, 3, \cdots c$. The number of patterns in the category is $N_i$. Each pattern has K attributes, and each attribute has J values (each feature is set to J values, just for simplicity).

Calculate the initial entropy:

$$H(I) = \sum_{i=1}^{c} - \left( \frac{N_i}{N} \right) \ln \left( \frac{N_i}{N} \right)$$

(1)

For the sample set, the classes of all patterns are known, so there are N labeled patterns that constitute the initial draft of the system.

Select an attribute as the root node of the judgment tree:
For each attribute $A_k$, $k = 1, 2, 3, \cdots, k$, the original patterns are divided into first level pattern groups according to the $J$ with $a_{kj}$ values of feature $A_k$.

For each classification of $n_{kj}$ patterns, the number of patterns belonging to category $c_i$ is $n_{kj}(i)$. Use the following relation to find the entropy of this branch.

$$H(I, A_k, j) = \sum_{i=1}^{c} -\frac{n_{kj}(i)}{n_{kj}} \ln \left( \frac{n_{kj}(i)}{n_{kj}} \right)$$  \hspace{1cm} (2)

Further find the arithmetic mean entropy $H(I, A_k)$ under this attribute.

The entropy drop caused by the test attribute $A_k$ is:

$$\Delta H(k) = H(I) - H(I, A_k)$$  \hspace{1cm} (3)

Choose the property $A_{k0}$ that produces the maximum entropy drop, i.e., $A_{k0}$ satisfies:

$$\Delta H(k_0) > \Delta H(k)$$  \hspace{1cm} (4)

Attribute $A_{k0}$ is the root of the judgment tree.

To build the next level of the judgment tree, choose an attribute $A_k$ as the node of this level so that the maximum entropy drop of information is obtained after testing $A_k$ on all branches, thus forming the second level of the judgment tree.

### 2.3 Defects of decision tree ID algorithm

The ID3 optimization algorithm makes it difficult to remove noisy and irrelevant attributes reasonably and cannot reduce the number of attributes in the detection, thus reducing the process comprehensibility of the decision. Therefore, when the decision tree is built, because the sample data is centralized with noise and isolated points, many branches are densely practiced to appear anomalous, so the most unsafe branches need to be cut out by pruning. This procedure is usually complex and not easy to understand. Secondly, most decision trees are restricted to detecting only individual attributes for each connection point, a restriction that drives increasingly difficult or indescribable representations of much of the complicated theory. The primary expressions are: ignoring the correlation between attributes, the same subtree in a decision tree repeatedly, and having some attributes detected several times directly on a certain path in a decision tree. To get rid of the problems of the ID3 optimization algorithm, the paper tries to introduce the rough set theory into the structure decision tree, selects the surface roughness in the rough set theory as the check attribute of the decision tree hair branch, and finally structures the decision tree of the primary equipment fault diagnosis of the distribution station of automation technology according to the approximate simplification, kernel and extensive of the decision table, and then provides a rapid and simple way of fault classification.

### 2.4 Decision tree algorithm based on rough set theory

Rough set theory can measure inconsistency and uncertainty in mathematical software and effectively analyze and deal with imprecision, inconsistency, incompleteness, and other imperfect information, thus discovering the implied underlying knowledge and revealing the hidden regularity. In treating the issue of variability, rough set theory can obtain natural numbers by formula measurement without all the preparation and additional information of the relevant data, and therefore it is proved to be
particularly suitable for data simplification, data correlation, similarity search, and discovery of data ways. Through continuous scientific research and development, rough set theory has long been making great progress in theoretical and practical applications, especially in machine learning algorithms, discovering expertise from database files, decision analysis, and analytics. At this stage, it has been used more successfully in many aspects, such as artificial intelligence technology, knowledge and data discovery, system identification and categorization, and fault diagnosis. As a mathematical class theory, rough sets apply equivalence relations to represent categorization in a flowing formal way. This expertise can then be seen as applying the set of equivalence relations \( R \) to discrete variables indicates the space \( U \) to carry out the zoning, and the knowledge is the conclusion of the \( R \) to \( U \) distinction. So under the value of \( U \) and \( R \), the knowledge base system can be defined as the association of all possibilities in the attribute \( R \) to the differentiation of \( U \). The expression is:

\[
K = (U, R)
\]

Further, to describe the degree of certainty of knowledge, rough set theory introduces the concepts of up-approximation and down-approximation and uses these concepts to define the degree of coincidence of a subset \( B \) in \( U \) with \( U \) after being divided by the relation \( R \), called roughness.

### 2.4.1 Knowledge and Indistinguishability Relationships

Assume that \( U \neq \emptyset \) as a relatively limited combination of interested targets composed of, called the theoretical domain \( U \). Subset \( X \subseteq U \) as one of the definitions, the basic knowledge in \( U \) that is embodied as a family set of theories, a classification family in \( U \) defined as the abstracted expertise of \( U \), commonly referred to as knowledge. This knowledge base expresses the various types of basic taxonomies of a functional management department or a group of them, composing the prerequisite prefabricated building blocks of that required definition and environment or its own association.

To facilitate the derivation of mathematical formulas, the theory of rough set foundations replaces the classification with an equivalence relation, let \( R \) be an equivalence relation on \( U \) and \( U^R \) as the union of all equivalence classes (or classifications in \( U \)) composed of. \([x]_R\) as the equivalence class of \( R \) including the element \( x \in U \). A knowledge base is an associative system software \( K \in (U, R) \), in which \( U \) is a non-empty finite set called a theoretical domain and \( R \) is a family of equivalence relations in \( U \).

If \( P \subseteq R \) and \( P \neq \emptyset \) then \( \cap P \) then (the association of all equivalence relations in \( P \)) is also an equivalence relation, called the indistinguishable association on \( P \), denoted as \( IND(P) \).

\[
[x]_{IND(P)} = \cap[x]_R
\]

The indistinguishability relation is the equivalence relation in the domain \( U \) of the argument when a set of attributes expresses the object \( P \). It suggests the granular structure of knowledge, which is responsible for the inability to represent certain concepts precisely using existing knowledge.

### 2.4.2 Rough set and upper approximation, lower approximation

Let \( X \) be a subset of \( U \), i.e., \( X \subseteq U \). When the set \( X \) can be represented as a concatenation of \( R \) basic equivalent classes, \( X \) is said to be \( R \)-definable, otherwise \( X \) is said to be \( R \)-undefinable.

An \( R \)-definable set is a subset of a theoretical domain \( U \) that can be defined exactly in a knowledge base \( K \), while an \( R \)-indefinable set cannot be defined in this knowledge base. An \( R \)-definable set is
also called an R-exact set, while an R-indefinable set is also called an R-imprecise set or an R-rough set, and a set \( X \subseteq U \) is called an exact set in \( K \) when there exists an equivalence relation \( R \in IND(K) \) and \( X \) is a \( R \in IND(K) \)-exact set; when for any of \( R \in IND(K) \), \( X \) is a \( R \)-rough set, then \( X \) is called a rough set of \( K \).

Uncertainty and fuzziness in rough set theory is a boundary-based concept, i.e., a fuzzy concept has a fuzzy boundary. Each uncertain concept is represented by a pair of exact concepts called upper approximation and lower approximation. Given a knowledge base \( K \in (U, R) \), for each subset \( X \subseteq U \) and an equivalence relation \( R \in IND(K) \), define two subsets:

\[
\mathcal{R}X = \bigcup \{ x \in U \mid [x]_R \subseteq X \} \tag{7}
\]

\[
\mathcal{R}X = \bigcup \{ x \in U \mid [x]_R X \neq \emptyset \} \tag{8}
\]

where \([x]_R\) denotes the class containing elements \( x \in U \) and \( R \) equivalents.

They are called the \( R \) lower and \( R \) upper approximation sets of \( X \), respectively.

The set \( \text{pos}_R(X) = \mathcal{R}X \) is called the positive domain of \( R \) of \( X \). It is the set consisting of those elements in \( U \) that definitely belong to \( X \) according to the knowledge \( R \);

The set \( \bar{R}X \) is the set consisting of those elements of \( U \) that, according to the knowledge \( R \), may belong to \( X \);

The set \( b_{nR}(X) = \bar{R}X - \mathcal{R}X \) is called the \( R \)-bounded domain of \( X \). It is the set consisting of those elements of \( U \) which, according to the knowledge of \( R \), can be judged to belong neither to \( X \) nor to \( (U - R) \) with certainty;

The set \( n_{egR}(X) = U - \bar{R}X \) is called the negative domain of \( R \) of \( X \). It is the set consisting of those elements in \( U \) that are definitely not part of \( X \) according to the knowledge \( R \).

It is easy to see that the rough set basis theory is very much in line with human cognitive characteristics, and the imprecision of knowledge may be caused by its large granularity distribution, and the way to deal with it is to generate a partial order construction of the knowledge base and to calculate, by computing, the least approximate and thus flowing minimum dependencies in the middle of formal standard attributes and management decision attributes, without having to be for each standard set of attributes.

### 2.4.3 Proximity and Membership

Another new concept, membership relation, can be derived from the concept of set approximation;

\( A \) is called a membership relation under \( R \) of \( X \) when and only when \( x \in \bar{R}x \).

Making \( \text{card} \) as a functional formula for finding the number of bonded group members, called the number of bonds or potential, then the corresponding level of a subset \( X \) within \( U \) with \( U \) after being compartmentalized by the equivalent circuit association \( R \) can be defined as a similar precision with the expression

\[
\alpha_R(X) = \frac{\text{card} \ (\mathcal{R}X)}{\text{card} \ (\bar{R}X)} \tag{9}
\]
Roughness of \( R \) of \( X \) \( \rho_R \)

\[
\rho_R(X) = 1 - \alpha_R(X) = 1 - \frac{\text{card}(RX)}{\text{card}(\bar{R}X)}
\]  

(10)

\( \alpha_R(X) \) is used to reflect the degree of completeness of our knowledge about understanding the set \( X \). Obviously \( 0 \leq \alpha_R(X) \leq 1 \).

\( \alpha_R(X) = 1 \) is called the set \( X \) is clear with respect to \( R \). \( \alpha_R(X) < 1 \), then, is called the set \( X \) is coarse with respect to \( R \). \( \alpha_R(X) \) can be thought of as approximating the accuracy of the set \( X \) under the equivalence relation \( R \).

As can be seen, this membership relationship is the fundamental point that distinguishes rough sets from traditional sets. While traditional sets consider that a set is completely determined by its elements and the elements can only be in two states of belonging or not belonging to the set, in rough sets, the uncertainty is related to the membership relation, while the ambiguity is expressed in the set itself.

### 2.4.4 Knowledge parsimony

Knowledge simplification is one of the main elements of modern logic for rough sets. The knowledge base document knowledge (attributes) is not equally important, and even some knowledge in it is redundant. By knowledge reduction, we mean deleting irrelevant or indifferent knowledge while maintaining the knowledge base categorization level unchanged. There are two basic elements in knowledge simplification: simplification and kernels. Before discussing their simplification and nucleation, the following definitions are given.

Let \( \pm \) be a family of equivalence relations, \( r \in R \), if \( \text{IND}(R) = \text{IND}(R - \{r\}) \) then Chen \( r \) is unnecessary in \( R \); otherwise it is said that \( r \) is necessary in \( R \). For every \( r \in R \) that is necessary in \( R \), then \( R \) is said to be independent; otherwise \( R \) is said to be dependent.

If \( R \) is independent, \( P \subseteq R \), then \( P \) is also independent. Let \( Q \subseteq P \), if \( Q \) is independent and \( \text{IND}(Q) = \text{IND}(P) \), then \( Q \) is said to be a simplification of \( P \). Clearly, \( P \) can have multiple simplifications. Clearly, there can be multiple simplifications of \( G \). The set of all necessary relations in \( P \) is called the kernel of \( P \) and is denoted as \( \text{core}(P) \).

The kernel is related to the simplification as follows:

\[
\text{core}(P) = \cap \text{red}(P)
\]

(11)

Where \( \text{red}(P) \) denotes all simplices of \( P \).

The use of the concept of the nucleus is divided into two aspects: firstly, it can be used as the basis for all simplification calculations since the nucleus is included in all simplifications and the measurement can be carried out directly; secondly, it can be interpreted as a combination of knowledge features that cannot be eliminated when knowledge is simplified.

### 2.4.5 Knowledge representation system

Within the rough set grounded theory, the knowledge of an object is represented by the essential features (attributes) of the particular object with their characteristic roots.
A knowledge representation system is defined as:

\[ S = (U, R, V, F) \]  \hspace{1cm} (12)

Where \( U \) is the thesis domain; \( R \) is the set of all attributes, which is divided into conditional attribute \( A \) and decision attribute \( D \) and \( A \cap D = \emptyset \); \( V = \bigcup_{P \in R} V_P \), \( V_P \) is the value domain of attribute \( P \), and \( f: U \times Q \to V \) is an information function that specifies the attribute value of each object \( x \) in \( U \).

\[ \text{SplitInfo}_A (X) = - \sum_{j=1}^{i} \frac{|x_j|}{|X|} \log_2 \frac{|x_j|}{|X|} \]  \hspace{1cm} (13)

The information gained of the sample set after splitting according to attribute \( A \):

\[ \text{InfoGain} (X, A) = H(X) - H_A(X) \]  \hspace{1cm} (14)

Information gain of the sample set after splitting of attribute \( A \):

\[ \text{GainRatio} (X, A) = \frac{\text{InfoGain} (X, A)}{\text{SplitInfo}_A (X)} \]  \hspace{1cm} (15)

These definition methods enable the knowledge of objects to be easily and quickly described in data diagrams, and such data analysis tables are called knowledge representation systems, also known as information content systems. The information of a knowledge representation system is shown in the form of a relational diagram, where the rows of an association table match the object to be explored, the columns match the attributes of the object, and the values of each attribute of the particular object represent the data of the object. It can be learned that an attribute matches an equivalence relation, and a table can be understood as a family of equivalence relations defined, i.e., a knowledge base. The knowledge parsimony explored in the previous section can be translated into attribute parsimony.

3 Results and Analysis

Along with the gradual development of decision tree algorithms in data mining algorithms, many derived optimization algorithms have been explored. For example, it will be widely used at the power engineering level. Because the automation technology distribution station’s primary equipment prevention and maintenance system is a linear system of large interconnected systems, in operation also constantly causes and accumulates a large amount of data and information. In the automation technology distribution station primary equipment prevention and maintenance system, the use of basic theoretical research methods has long overcome many problems, if you can use data mining algorithms, then you can make more comprehensive use of this equipment operation state, reveal the automation technology distribution station primary equipment prevention and maintenance system in previous years behind the accumulation of information contains the basic principles, standards, looking for more effective ways to deal with the problem It is also possible to provide stronger scientific arguments for management decisions. To make the data mining algorithm based on the decision tree algorithm more suitable for fault detection in the automation technology distribution station primary equipment prevention and maintenance system, the following we make reasonable improvements to the most primitive optimization algorithm and the selection of typical alarm signals to carry out an in-depth analysis, and it is carried out to categorize the combined solution, each attribute of the common ID3, C4.5, CRAT algorithm The comparison and analysis are carried out.
3.1 Research process of automated substation primary equipment fault prevention and maintenance system based on decision tree algorithm

The improved decision tree algorithm can perform operations on different types of data sets. This study is based on the improvement of the nearest neighbor method, collecting suitable samples to construct the decision tree with the best effect, screening the remaining samples as candidates, and obtaining a decision tree with higher spatiotemporal complexity. In practical applications, more continuous types are used, so the discretization of continuous types is an important link in the construction of decision trees, and the discretization effect of continuous types determines the classification accuracy after the construction of decision trees to a certain extent and also has a greater impact on the construction time of decision trees. In this paper, after improvement, a simple and clear discretization method is proposed. Through optimization, a better tree is obtained, due to the elimination of the pruning link is eliminated, so that the spatiotemporal complexity of the tree decreases. The final obtained decision tree algorithm is more accurate in classification and easier to use rules. The process of constructing a quality decision tree algorithm, which is improved in many ways, consists of the following steps:

Step 1: The samples are collected using the optimal screening method, and the suitable samples are added to the training set;

Step 2: Improving the dichotomous lookup-based method for discretizing continuous attributes;

Step 3: Obtain the attribute values corresponding to each sample data set by the improved sample selection attribute;

Step 4: Improve the selection criteria of the test attributes and start the construction of the decision tree based on this;

Step 5: Finish the construction of the decision tree until all the basic attributes are used up, or the training sample set under a branch corresponds to one type, and the accuracy meets the requirements, then end the algorithm.

3.1.1 Functional design of data mining fault diagnosis

Data mining technology is used in the substation monitoring system to diagnose the system status, which makes the system reasoning much more efficient. In this paper, based on the demand analysis of the major modules of the system, the functional model and interface of the system are designed so that the system has the corresponding functions, and the user and the expert are the main participants of the system, and the user can interact with different modules, including the interaction with the system expert to input the problem, and after reasoning, get the approved conclusion; the expert interacts with the system module, entering the knowledge and organizing it, as well as setting the parameters. In this study, the previously optimized decision tree algorithm is applied to the fault diagnosis of the power system to achieve rapid fault diagnosis, timely fault detection, and fault resolution with the help of the superior performance of the algorithm, thus reducing the losses of the power system due to faults. The design of this system class mainly includes forward reasoning, backward reasoning, backward comprehensive reasoning, inference class, decision tree class, interpreter class, system setting and knowledge editing class, and human-computer interaction interface. The decision tree class mainly includes the data set, number of rules, generation rules, number of nodes (intermediate nodes), execution function, and generation
rule function. The inference class keywords include name and number, and the inference process is realized by keyword description, keyword logic, and suggestion result.

3.1.2 Database and database tables

The database of this system consists of three main aspects, namely, the fact database, the knowledge database, and the system settings database. The knowledge database consists of several tables, including the keywords table, the rules table, the inference rules table, the hypothesis expression explanation table, the keywords logic table, and the factors table. Among them, the main contents of the keyword table are names, logos, and related explanations. The rule table is used to access rule information. The inference rule table is the same as the rule table; the hypothesis expression explanation table is used to access hypothesis expression names and logos, the keyword logic table is responsible for names and logos, and the factor table is responsible for accessing factor names and values. Each table stores different contents, and the tables do not affect each other. This split table storage is a means to classify data information and store it in the correct table according to the type of data information to improve the efficiency of data query. A rule table is shown in Table 1.

<table>
<thead>
<tr>
<th>Columns</th>
<th>Name</th>
<th>Type</th>
<th>Is it empty</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Rule ID</td>
<td>int</td>
<td>No</td>
<td>Primary Key</td>
</tr>
<tr>
<td>ifstr</td>
<td>Facts</td>
<td>Varchar50</td>
<td>Yes</td>
<td>Conditions</td>
</tr>
<tr>
<td>thenstr</td>
<td>Assumptions</td>
<td>Varchar100</td>
<td>Yes</td>
<td>Conclusion</td>
</tr>
<tr>
<td>ifm</td>
<td>Fact Factor</td>
<td>float</td>
<td>Yes</td>
<td>Default is 1</td>
</tr>
<tr>
<td>thenm</td>
<td>Determining factors</td>
<td>float</td>
<td>Yes</td>
<td>Default is 1</td>
</tr>
<tr>
<td>prior</td>
<td>Rule priority</td>
<td>int</td>
<td>Yes</td>
<td>Incremental</td>
</tr>
<tr>
<td>intime</td>
<td>Time</td>
<td>datetime</td>
<td>Yes</td>
<td>Input time</td>
</tr>
</tbody>
</table>

The system setting and fact database mainly has the system setting table, rule learning table, and fact table. The fact table is mainly used to store various relevant factors, firstly, the amount of data in the learning or rule table is searched, if the number of collected information is less than the set value, it enters the system for inference without using the fact table, if the amount of data is greater than the set value, inference is made according to the data recorded before the set value, the new rule is stored in the rule learning table, in the rule query, according to the rule table to determine whether the rule searched in the rule query, according to the rule table to determine whether the rule searched for exists, if the search is not available, this new rule is generated. The main task of the system setting table, detailed setting table and setting association table is to access the parameters related to the system and roles, including the number, inference rules and policies for handling conflicts. The rule learning table and the system setting table are shown in Table 2 and Table 3, respectively:
### Table 2 Rule Learning table

<table>
<thead>
<tr>
<th>Name of school section</th>
<th>Name</th>
<th>Type</th>
<th>Is it empty</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Rule ID</td>
<td>int</td>
<td>No</td>
<td>Primary Key</td>
</tr>
<tr>
<td>ifstr</td>
<td>Facts</td>
<td>Varchar50</td>
<td>Yes</td>
<td>Conditions</td>
</tr>
<tr>
<td>thenstr</td>
<td>Assumptions</td>
<td>Varchar100</td>
<td>Yes</td>
<td>Conclusion</td>
</tr>
<tr>
<td>ifm</td>
<td>Fact Factor</td>
<td>float</td>
<td>Yes</td>
<td>Default is 1</td>
</tr>
<tr>
<td>themm</td>
<td>Determining factors</td>
<td>float</td>
<td>Yes</td>
<td>Default is 1</td>
</tr>
<tr>
<td>prior</td>
<td>Rule priority</td>
<td>int</td>
<td>Yes</td>
<td>Each increment</td>
</tr>
<tr>
<td>intime</td>
<td>Time</td>
<td>datetime</td>
<td>Yes</td>
<td>Input time</td>
</tr>
</tbody>
</table>

### Table 3 System settings

<table>
<thead>
<tr>
<th>Columns</th>
<th>Name</th>
<th>Type</th>
<th>Is it empty</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Rule ID</td>
<td>int</td>
<td>No</td>
<td>Primary Key</td>
</tr>
<tr>
<td>name</td>
<td>Set name</td>
<td>Varchar50</td>
<td>Yes</td>
<td>Name</td>
</tr>
<tr>
<td>remark</td>
<td>Remarks</td>
<td>Varchar500</td>
<td>Yes</td>
<td>Marker</td>
</tr>
</tbody>
</table>

### 3.1.3 System Function Reference

Decision tree algorithms are applied in the system software, and a series of exercises are used to obtain models with relatively high accuracy. The models are created from a part of the data information and the rest of the data information is used for testing and certification of the model. The way obtained in the data mining session is fully considered may not match the customer’s needs when it is necessary to back up to the previous session and select the data information again. In order to advance the data mining method of data visualization, the decision tree algorithm used in the power inspection system applies the “If-Then” criterion. It can be concluded that the conditions leading to SF6 gas leakage from the circuit breaker are gas leakage from the body, poor contact of the auxiliary contact of the gas pressure gauge, short circuit between the signal terminals in the terminal block of the circuit breaker terminal box, loose signal terminals in the external terminal block of the measurement and control device, and disconnection of the internal logic contact of the measurement and control device. The conclusion reached through reasoning is that the signal terminals in the circuit breaker terminal box terminal row short circuit between, need to disconnect the terminal box DC power. The function of this system still needs to be extended, and it can also use artificial neural networks, fuzzy identification, and other algorithms to carry out the extension of the function. Therefore the staff can refer to the relevant conclusions of reasoning and diagnosis to provide guidance direction and reference suggestions for the final power fault diagnosis.

### 3.2 Data Analysis

The cloud computing platform set proposed by a university is used to test the performance of the improved optimization calculation method, and the actual effect of the optimization algorithm is verified according to the respective comparison of C4.5 and ID3. This university does not explicitly specify the total number of training and test sample sets in the database system, and this time the selection of training sample sets is carried out in an arbitrary form, 80% of which are selected, and the remaining 20% of the sample sets are used for testing, and the optimization algorithm is implemented ten times, and the average value is sought based on the ten conclusions. The data of this scientific study are shown in Table 4, and the before and after comparison of the accuracy rate is shown in Figure 1 below.
Table 4 Research data sets

<table>
<thead>
<tr>
<th>Data set name</th>
<th>Number of samples in the data set</th>
<th>Optimized sample size</th>
<th>Number of attributes</th>
<th>Number of categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPECT heart</td>
<td>91</td>
<td>62</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Hayes-Roth</td>
<td>140</td>
<td>130</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Monk</td>
<td>175</td>
<td>150</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Balance-scale</td>
<td>624</td>
<td>610</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Nursery</td>
<td>12965</td>
<td>12740</td>
<td>18</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 1 Comparison chart of correct rates

As can be seen from Figure 1: The new algorithm optimizes the training data sample set for screening, and the final decision tree is constructed with higher classification accuracy and simpler and more understandable generation rules. The larger the training sample data set for constructing the decision tree, the higher the computational cost, where discretization and pruning incur a significant computational cost, and the same is true for the computation of attribute entropy values, as shown in Figure 2 for the comparison of model building speed, in Figure 3 for the comparison of the number of leaf nodes, and in Figure 4 for the comparison of the total number of nodes.

Figure 2 Speed Comparison Chart
Combining the graphical patterns given in Figure 2, Figure 3, and Figure 4 can learn that the optimized decision tree possesses a lower number of leaf nodes and total nodes, so the formation criteria are simplified to some extent, and the new algorithm has a higher model efficiency compared to the traditional algorithm. In this paper, according to the traditional decision tree, the optimized sample set is used to carry out the achievement, and in addition, the characteristic selection specification is optimized, and the discretization method is also improved, and through a series of optimization, a smaller decision tree is constructed, and if there is no explicit provision, and no pruning measurement is done, so the optimized algorithm in the processing method of smaller-scale data, the measurement cost is significantly reduced, and the performance is significantly improved.

Based on the above data analysis, the following points are summarized:

Compared with C4.5 and ID3, the optimized algorithm has higher classification accuracy; faster modeling speed, and simpler production rules. In the production of the new algorithm, its sample dataset is filtered by the optimized algorithm without considering the noise and confounding regions.
in the sample dataset, the classification of the new algorithm becomes more accurate and the production rules become simpler. Not only that, the improved algorithm simplifies the selection criteria of test attributes and uses balancing coefficients in constructing the tree, which overcomes the multi-value bias problem and makes the decision tree easier to understand. Comprehensive analysis shows that the optimized algorithm has better application efficacy compared with C4.5 and ID3 algorithms.

4 Conclusion

This paper studies the application of the decision tree algorithm to automatic substation primary equipment fault diagnosis. Firstly, we analyze the research background and research significance of automatic substation primary equipment fault diagnosis, collect the related literature and data at home and abroad, and study and research with this thesis. After that, the concept of data mining algorithm and mining steps are introduced, and several classification methods of commonly used data mining are introduced, which are decision tree algorithm, optimization decision tree algorithm based on rough set theory method, and, at the same time, some commonly used methods of automatic substation primary equipment diagnosis are introduced., it is seen that decision tree algorithm has been widely in automatic substation primary equipment diagnosis system. The decision tree algorithm has been widely used in automatic substation primary equipment diagnosis systems. The selection of typical alarm signals is also analyzed, and they are classified and merged. The various attributes of the commonly used ID3, C4.5, and CRAT algorithms are compared and analyzed to provide references for the subsequent optimization of the algorithms, followed by the improvement in the three aspects of optimizing the number of samples, optimizing the test attributes, and discretization, respectively, and it is verified that the optimized decision tree algorithm has a better performance compared with the traditional C4.5 and ID3. Finally, data mining technology is used in the automatic substation primary equipment fault prevention and maintenance system. The knowledge expression of the system utilizes if-then. The optimized decision tree algorithm realizes the combination of knowledge expression and acquisition, based on which a simpler and clearer reasoning process is realized, which can effectively improve the performance of automatic substation primary equipment fault diagnosis, and because the rules of the algorithm can be automatically generated, so it can greatly reduce the time of work and improve the efficiency of automatic substation primary equipment fault diagnosis.

References


