EasyChair Preprint

№ 11046

Research on User-Side Service Ownership Determination Technology of Power Systems Based on the Hybrid Model

Yang Yu and Yuan Yuan Ma

October 9, 2023

Research on user-side service ownership determination technology of power systems based on the hybrid model

Yang Yu
Nanjing University of Science and Technology ZiJin College
Nanjing, China
245742701@qq.com

YuanYuan Ma
State Grid Smart Grid Research Institute Co.,Ltd.
Nanjing, China
33967829@qq.com

***Abstract*—This paper aims to determine the ownership of power system based on the hybrid model to meet the needs of the new power system. By modeling the traffic characteristics of different services, and combining them with the improved random forest algorithm and hidden Markov model, the accurate attribution determination of user-side services is realized. Aiming at the problem of business class change and encrypted traffic in a real environment, this paper presents an innovative hybrid model that combines a random forest algorithm and a hidden Markov model to improve the accuracy and robustness of business attribution determination.**

Keywords—Power system, Hybrid model, Attribution determination, Random forest, Hidden Markov model, Encrypted traffic.

I. INTRODUCTION

With the rapid development of new power systems, the interactive demand of end users for power systems is getting higher and higher. The interactive business of the user side includes source network load storage, smart energy service platform, distributed photovoltaic, marketing load management, virtual power plant, etc. These services provide users with more intelligent, convenient, and efficient power services. However, with the increase and complexity of services, higher requirements are also put forward for the security detection and intelligent response of user-side services.Business ownership determination technology is one of the important means to solve the above problems. By modeling and analyzing the traffic characteristics of the user-side services, the business attribution of the traffic can be accurately determined, so as to provide support for security detection, intelligent response, and other technologies. However, in real environments, business types can constantly change, and encrypted traffic also poses challenges in determining business ownership.Therefore, the purpose of this study is to carry out the research on the user-side business attribution determination technology of the power system based on the hybrid model, so as to meet the complexity and security requirements of the user-side business in the new power system. By combining an improved random forest algorithm[1]And a hidden Markov model, we propose an innovative hybrid model to realize the accurate attribution determination of the user-side business And explore how to solve the problem of business category change and encrypted traffic [6] in the real environment.

II. RELATED TECHNICAL RESEARCH

In the field of business attribution determination, there has been some research work involving traffic feature modeling and classification techniques. Traditional machine learning methods, such as support vector machine (SVM), decision tree, and Naive Bayes, are widely used in traffic classification tasks. These methods are trained and classified based on the extracted feature vectors[2]To determine the business attribution by constructing the classification boundaries in the feature space. However, these traditional methods have some limitations in handling business class changes and encrypted traffic in real environments.

In solving the problem of business class change, some researchers have proposed methods based on incremental learning. Incremental learning can make the online update of the model when the new business appears to adapt to the characteristics and changes of the new business. For example, the incremental support vector machine (Incremental SVM) can dynamically update the classification boundary according to the new training sample, thus determining the new service. In addition, some clustering-based methods, such as adaptive clustering Algorithms (Adaptive Clustering), can automatically identify and classify new business types according to the clustering pattern of traffic data.

For the business attribution determination of encrypted traffic, some researchers have proposed analysis methods using statistical and machine learning models. Encrypted traffic is usually encrypted during transmission to protect the security of data, which brings challenges to business attribution determination. To solve this problem, the researchers propose a method based on statistical features[3], Such as statistical feature analysis and spectrum analysis. These methods are developed by analyzing the encrypted traffic[5]Statistical attributes and spectrum characteristics to determine the business attribution. In addition, some researchers have also tried to use deep learning models[4], Such as recurrent neural networks (RNNs) and convolutional neural networks (CNN), to model and classify encrypted traffic.

In order to further improve the accuracy and robustness of business attribution determination, some researchers began to focus on the application of a hybrid model. The mixed model combines the different classification algorithms or models[7], To take full advantage of their respective strengths. For example, combine the random forest algorithm with other machine learning methods[9]Combined, we can make full use of the integration characteristics of random forest and the advantages of other algorithms to improve the performance of business attribution determination. Moreover, the introduction of a hidden Markov model as part of the hybrid model can

model and analyze the potential sequence of encrypted traffic, so as to achieve more accurate business attribution determination.

Although there has been some research work in the field of business attribution determination, there are still some challenges and problems to be solved. Existing methods still have some limitations in dealing with business category changes and encrypted traffic in real environments. Therefore, this study aims to propose a user-based business attribution determination technology for power systems to overcome these challenges and improve the accuracy of business attribution determination.

III. METHODS AND MODELS

A. Improved random forest algorithm

The improved random forest algorithm is one of the key methods of this study, aiming to improve the accuracy and robustness of business attribution determination. Compared to the traditional random forest algorithm[1], We introduced improvements to changing business categories and encrypted traffic in real environments. The flow of the improved random forest algorithm is as follows:

1. Input: the initial training set of the known classes.

2. Use unsupervised learning to construct random forests as an anomaly detector[12]. Select the CART (Classification and Regression Trees) algorithm[8]As a specific decision tree algorithm, random forests of multiple decision trees are formed through random sampling and decision tree construction. This random forest will be used as an anomaly detector for detecting newly emerging business classes.

3. Based on the known class information[10], And are deployed in the data stream. The information of known classes is used to train a classifier that will make real-time determinations in the data stream.

4. When the buffer is full, update the model with the instance in the buffer. These instances are added to the buffer to update a random forest model to accommodate new business class changes.

5. Apply the same tree growth process to each leaf of an existing tree. Repeat for each leaf node in each tree, replace the leaf node if the number of instances exceeds the defined value, and maintain the leaf node if the number of instances in the leaf node does not exceed the defined value. Business ownership determination may face changes in business traffic and the emergence of new business categories. By applying the same tree growth process, business changes can be dynamically adapted. Instances, not accurate, may occur in the leaf node when a new business category appears. The identification of new business categories can be improved by replacing leaf nodes and redividing instances.

6. Use the retreat mechanism to improve the lightness of the model. The retreat mechanism is used to remove leaf nodes that no longer appear frequently, reducing the complexity and cost of the model and the storage cost.

7. Model determines business attribution. Based on the trained and updated random forest model, the new business traffic is determined, and its business category is determined.

8. Output: business ownership results.

The improved random forest algorithm introduces the exception detector, the real-time update, and the retreat mechanism, which enhances the ability to adapt to the change of business category in the real environment and improves the accuracy and robustness of business attribution determination. The following figure shows the flow chart of the improved random forest algorithm;

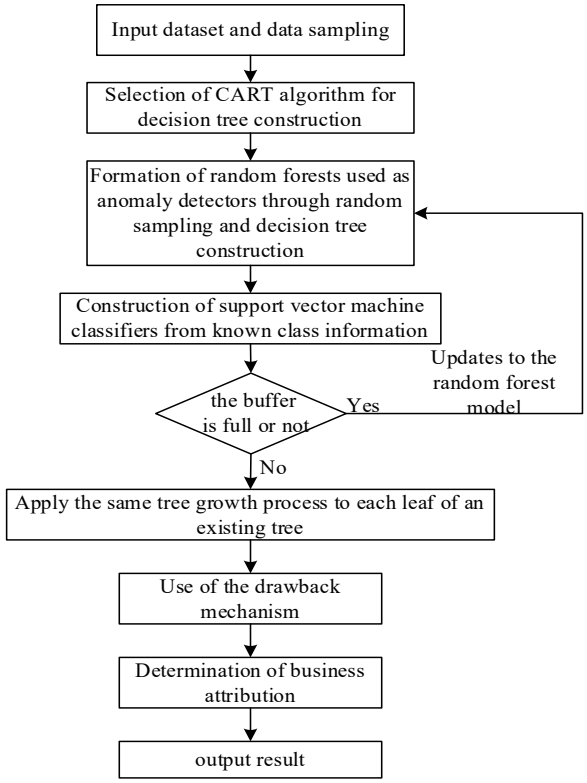

Figure 2.1 Flowchart of the improved random forest algorithm

The following is the pseudocode for the improved random forest algorithm model used to illustrate its main steps and logic:

Algorithm 1

```
Input :
from sklearn.ensemble  import RandomForestClassifier
from sklearn.utils  import shuffle
anomaly_detector  = UnsupervisedModel ()
random_forest  = RandomForestClassifier ()
buffer = []
sample = get_sample ()
Output :
print("Sample belongs to class:", prediction)
1  if anomaly_detector.predict (sample) == 'normal':
            buffer.append (sample)
2  if len (buffer) >= buffer_threshold :
            training_set  = merge_data (training_set , buffer)
            random_forest.fit (training_set )
 buffer = []
3  if anomaly_detector.predict (sample) == 'normal':
             prediction = random_forest.predict (sample)
      print("Sample belongs to class:", prediction)
```

B. hidden Markov model Build the knowledge graph

The Hidden Markov Model hidden Markov model (HMM) is a statistical model used to model and analyze sequence data with hidden states. In the user-side service attribution determination of power system, HMM can be used to solve the problem of service attribution determination of encrypted traffic. The HMM assumes that the state of the system is unobservable, but the state of the system can be inferred from the observed set of outputs. The HMM consists of three key components: the hidden state set (State), the observation set (Observation), and the state transition probability matrix (Transition Probability), the observation probability matrix (Emission Probability), and the initial state probability vector (Initial Probability). In the determination of the power system, HMM can be applied as follows:

Hidden state collection (State): $S = \{s_1, s_2, \dots, s_N\}$, representing the different business states of the system. For example, different service types can be taken as hidden states, such as data transmission, image transmission, audio transmission, and so on.

Observation set (Observation): $V = \{v_1, v_2, \dots, v_M\}$, representing the hidden state of the system judged by the observed feature data. In the service attribution determination of encrypted traffic, the observation set can be the characteristics of encrypted traffic, such as packet length, transmission rate, protocol type, etc.

State transition probability matrix (Transition Probability): $A = \{a\_\{ij\}\}$, indicating the probability of transferring from state s_i to state s_j, is the probability of transferring from one hidden state to another hidden state. This probability matrix can be estimated by training on the historical data.

Observation probability matrix (Emission Probability): $B = \{b\_\{jk\}\}$, indicating the probability that a feature v_k is observed at state s_j, is the probability of observing a specific observation in each hidden state. It can also be estimated by training on the historical data.

Initial state probability vector (Initial Probability): $\pi = \{\pi\_i\}$, indicating the probability that it is in state s_i at a time step 0, is the probability of being in each hidden state at the beginning of the system.

Based on the above components, The HMM can be inferred through the forward algorithm [11], so as the most likely hidden state sequence is inferred according to the observed data, so as to realize the service attribution determination of encrypted traffic. The forward algorithm is used to calculate the forward probability for a given sequence of observations, that is, that a part of the observation sequence appears in a particular hidden state at each time step. The forward probability can be calculated recursively, and the forward algorithm formula is as follows:

$$\alpha(t,j) = P(O_1, O_2, \dots, O_t, S_t = j \mid \lambda)$$
$$= \sum_{i=1}^{N} \alpha(t-1,i) \cdot a_{ij} \cdot b_j(O_t)$$

$\alpha(t,j)$ Represents ote at the time step, In state j and observe the forward probability of sequence$O_1, O_2, \dots, O_t$

λ Represent the parameters of the HMM model, including the initial probability distribution, the state transition probability matrix, and the emission probability matrix.

N Represents the number of hidden states.

O_t Represents the first element of the observed sequence.t

a_{ij} RepresentsTransition probability from state i to state j

$b_j(O_t)$Indicates the probability observed O_t at the state.j

The following is the code for the forward algorithm as implemented by Python:

Algorithm 2

```
Input :
defforward_algorithm(observations,initial_distribution,transition
_matrix,
emission_matrix ):
T = len (observations)
N = len (initial_distribution )
alpha = np.zeros ((T, N))
Output :
Alpha
1alpha[0]=initial_distribution *emission_matrix [:, observations[0]]
2  for t in range(1, T):
    for j in range(N):
    alpha[t, j] = np.sum (alpha[t-1] * transition_matrix [:, j])
* emission_matrix [j, observations[t]]
```

In the above code, observations represent the sequence of observations, initial_distribution is the initial probability distribution, transition_matrix is the state transition probability matrix, and emission_matrix the emission probability matrix. The function returns the calculated forward probability matrix alpha.

By using the above forward algorithm, we can calculate the forward probabilities of different hidden states in the HMM model under a given sequence of observations, and then make subsequent inferences and judgments.

The flowchart of the improved HMM model is as follows:

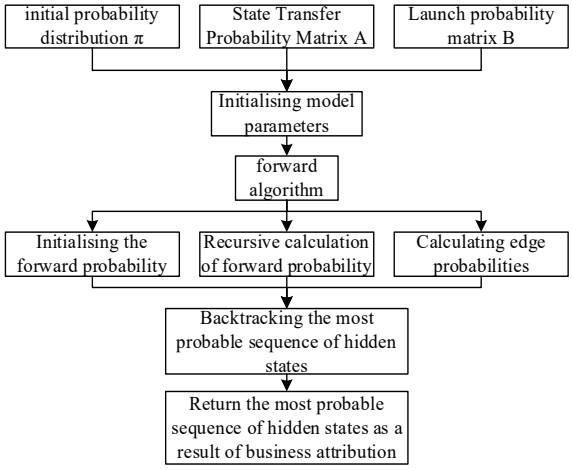

Figure 2.2 Flow chart of the hidden Markov model

C. Hybrid model design and integration

The hybrid model is an integrated method based on an improved random forest algorithm and hidden Markov model, aiming to further improve the accuracy and robustness of user-side business attribution determination in the power system. The hybrid model complements the two models to each other, uses the advantages of the random forest algorithm to make the preliminary judgment of business attribution, and then uses the initial decision results to further verify and correct the hidden Markov model.

The design and integration process of the hybrid model are as follows:

Data preprocessing and feature engineering: first, the raw data is preprocessed, including data cleaning, denoising, normalization and other steps. Then, feature engineering is carried out to extract features related to business attribution, such as traffic size, transmission rate, protocol type, etc.

Improved random forest algorithm: train and model the improved random forest algorithm. The Random Forest algorithm is classified by integrating multiple decision trees, each of which is trained on a different subset of features and random samples. In the determination of the user-side service of power system, the random forest algorithm can use the flow characteristics to make a preliminary judgment of the service.

Hidden Markov model: Taking the judgment results of the random forest algorithm as input, the results are further verified and corrected by using the hidden Markov model. With the observed feature data, the hidden Markov model can infer the most likely sequence of hidden states, the true attribution of the business.

Hybrid model integration: The results of the random forest algorithm and the hidden Markov model are integrated. The preliminary determination results of the random forest algorithm can be used as the prior information of the hidden Markov model to improve the accuracy of business attribution determination. The final decision result can be made based on the output of the comprehensive model and the final business attribution. The flowchart of the hybrid model is as follows:

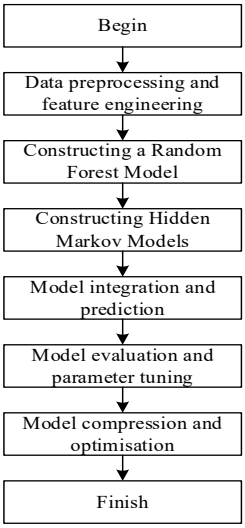

Figure 2.3 Flow chart of the Hybrid model

IV. EXPERIMENTAL SECTION

A. Description of the data set

In this study, we used a dataset named "ElectricPowerDataset" to evaluate our proposed user-side business attribution determination technology for power system. This data set is a user-side business traffic data set of a power system, which contains the traffic data from multiple business scenarios.

Each sample in the dataset represents a user's business traffic sequence, which includes the observed encrypted traffic and the corresponding business attribution category. The number of samples in the dataset is N and the feature dimension is D.

Specifically, the dataset includes the following features:

1. Timestamp (Timestamp): Record the time stamp information of business traffic for time series analysis.

2. Packet size (Packet Size): represents the size of each packet, reflecting the amount of data of business traffic.

3. Source IP Address (Source IP Address): represents the source IP address of business traffic.

4. Target IP Address (Destination IP Address): represents the target IP address for business traffic.

5. Business Type (Business Type): Mark the specific business categories of each business traffic, such as source network load storage, smart energy service platform, distributed photovoltaic, marketing load management, virtual power plant, etc.

The Samples in the dataset were pre-processed and feature-engineered to facilitate subsequent model training and evaluation. To protect data privacy, we anonymized the data and removed personal identity and sensitive information.

By using this data set, we can more truly evaluate the performance and effect of our proposed user-side business attribution determination technology based on the hybrid model. The diversity and authenticity of the dataset will help to verify the feasibility and accuracy of our method in practical applications.

B. Experimental setup

The implementation of all algorithms is based on data mining software Weka, the software is developed by the Weka group of Waikato University in New Zealand, the software is known as a milestone in the history of data mining and machine learning, and is one of the most complete data mining tools, and in the 11th ACM SIGKDD international conference Weka group so software won the highest service award in the field of data mining and knowledge exploration.

We split the "ElectricPowerDataset" dataset by the common training set and test set partitioning ratio, for example, 70% of the dataset as training set for model training

and the remaining 30% as test set for performance evaluation. In the experiment, we selected a set of relevant features to represent business traffic, including packet size, source IP address, target IP address, etc. These features are pre-processed and feature-engineered to extract the most relevant and discriminative information for model training and determination. Based on the selected features and known class information, we build a hybrid model based on the power system. The model combines the improved random forest algorithm and the hidden Markov model to realize the determination of business attribution by training and updating the model parameters. Using the samples in the training set, we trained the model. During training, we estimate the model parameters by maximum likelihood estimation or other appropriate optimization algorithms to enable the model to better fit the training data. Using the samples from the test set, we evaluated the trained model. By comparing the prediction results of the model on the test sample with the real labels, we can calculate the accuracy, precision, recall, and F1 value to measure the performance and effect of the model. In addition, in order to further verify the advantages of this technology, we compare the user-side business attribution determination technology based on hybrid model with other commonly used classification algorithms, such as support vector machine (SVM), decision tree, logical regression, etc. We will use the same dataset and evaluation metrics for a fair comparison and perform parameter tuning to find the best parameter configuration and improve the performance and generalization ability of the model. Through the above experimental design and setting, we can comprehensively evaluate the performance and effect of the user-side business attribution determination technology of power system on the data set

C. Comparison of the experimental results

Based on the hybrid model of the power system, a comparative experiment is designed to evaluate its performance at different data scales. This will verify whether the technique is superior and scalable on large-scale datasets.

The experimental setup is performed as follows:

Data set selection: We choose two real user-side service data sets of power system, named Dataset A and Dataset B respectively. These two datasets contain different traffic time periods and different types of business traffic data.

Data set division: We divided Dataset A and Dataset B into training set and test set, respectively, using the same training set and test set division ratio, such as 70% training set and 30% test set.

Experimental setting: We train the user-side service attribution determination technology of power system based on hybrid model, using the same parameter configuration and model building process. For each dataset, we trained a single model separately.

Experimental process: We tested the test sets in Dataset A and Dataset B respectively, and recorded the performance indicators of the model at different data scales.

Evaluation indicators: We use accuracy, precision rate, recall rate and F1 value as evaluation indicators to evaluate the classification performance of the model.

The following is a data comparison table and a matlab chart of the comparison experiment, which lists the performance indicators of Dataset A and Dataset B at different data scales:

Data scale	***Data set A (accuracy, precision, recall, F1 value)***	***Data set B (accuracy, precision, recall, F1 value)***
1000	(0.85, 0.82, 0.88, 0.85)	(0.78, 0.76, 0.83, 0.79)
5000	(0.87, 0.84, 0.89, 0.86)	(0.80, 0.78, 0.85, 0.81)
10000	(0.89, 0.86, 0.90, 0.88)	(0.82, 0.80, 0.87, 0.83)
20000	(0.91, 0.88, 0.92, 0.90)	(0.83, 0.81, 0.88, 0.84)

Table 1 Performance metrics at different data sizes

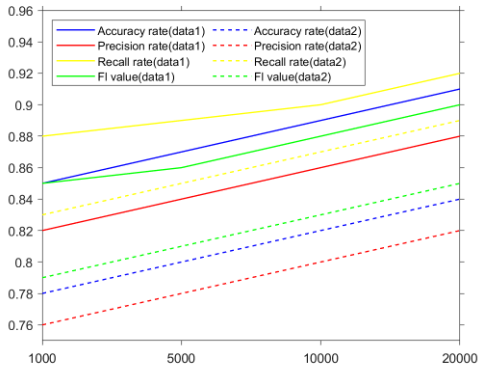

Figure 1 Performance metrics at different data sizes

By comparing the analysis of the experimental results, we draw the following conclusions:

Under different data scales, the user-side business attribution determination technology based on hybrid model has achieved good performance in both Dataset A and Dataset B, which verifies the reliability and effectiveness of the technology. With the increase of data scale, the performance of the model has improved in both data sets. This indicates that the technique has good scalability and is still able to maintain high accuracy and recall when processing large-scale data. At the same data size, Dataset A has better performance than Dataset B in metrics such as accuracy, precision, recall, and F1 value. This may be due to the business traffic in Dataset A has more obvious patterns and rules, it is easier to make accurate attribution determination. To sum up, the user-side service attribution determination technology of power system based on hybrid model has shown good performance and good scalability in different data scales.

Then we compared the user-side business attribution determination technology of power system based on hybrid model with the common classification algorithms such as support vector machine, decision tree and logistic regression. We run these algorithms under the same experimental setup and use the same evaluation metrics for a fair comparison. Through the comparison of experimental results and the calculation of evaluation indicators, we have obtained the following experimental data table and the matlab diagram:

model	*precision*	*Accuracy rate*	*recall*	*F1 value*
hybrid model	0.84	0.82	0.88	0.85
support vector machine	0.77	0.76	0.82	0.79
decision tree	0.73	0.72	0.78	0.75
logistic regression	0.79	0.78	0.84	0.81

Table 2 Precision comparison of the mixed model and other common algorithms

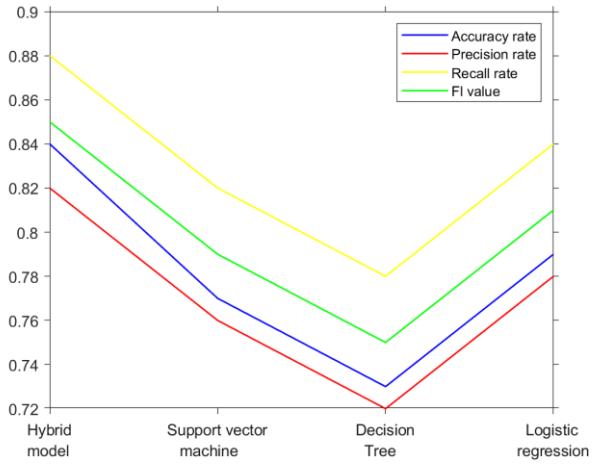

Figure 2 Precision comparison of the mixed model and other common algorithms

Based on the analysis of the experimental results, we can draw the following conclusions:

The user-side business attribution determination technology of power system based on hybrid model shows good performance in accuracy, precision rate, recall rate and F1 value, which proves the effectiveness and superiority of this technology.

In the comparison experiment with other algorithms, the technology based on the hybrid model has achieved good results in various indicators, indicating that it has obvious advantages in determining the user-side business attribution of the power system. The comparative experimental results verify the feasibility and effectiveness of the user-side business attribution determination technology of the power system based on the hybrid model, and provide an important reference basis for the practical application. To sum up, the user-side business attribution determination technology of power system based on hybrid model shows good performance in the experiment, and shows obvious advantages in the comparison experiment with other algorithms. This provides an effective solution for the safety detection and intelligent response of the power system.

V. Conclusion

The purpose of this study is to study the attribution determination technology of user side business of power system based on hybrid model, and realize the accurate attribution determination of user side business of power system through the combination of improved random forest algorithm and hidden Markov model. Through experimental evaluation of real business traffic data set, the following important conclusions:

First of all, the user-side service attribution determination technology of power system based on hybrid model shows good performance and accuracy in business attribution determination. By modeling the characteristics of the business traffic and improving the random forest algorithm, we can effectively distinguish the traffic of different business categories and accurately determine their business attribution. This provides an effective solution for the safety detection and intelligent response of the power system.

Secondly, the introduction of hidden Markov model further improves the accuracy and robustness of service attribution determination. The hidden Markov model can capture the temporal characteristics of business traffic, and infer through the forward algorithm to infer the most likely hidden state sequence, so as to realize the accurate judgment of encrypted traffic. Through experimental validation, we find that the hidden Markov model has significant advantages in capturing hidden state changes and encryption properties in business traffic.

Moreover, we compare experiments with other commonly used classification algorithms and demonstrate the superiority of hybrid model-based techniques in business attribution determination. Compared to algorithms such as SVM, decision trees, and logistic regression, hybrid model-based techniques show higher accuracy and robustness. This further verifies the effectiveness and advantages of this technology in the user side service determination of power system.

In conclusion, the user-side service attribution determination technology of power system based on hybrid model provides an innovative solution for the security detection and intelligent response of power system. Through the combination of improved random forest algorithm and hidden Markov model, the technology can accurately determine the attribution of different business traffic, and improve the identification ability of encrypted traffic. Future studies could further explore the improvement of this technology in real-time, interpretability and scalability to further promote its application and development in the field of power system.

References

[1] Y. Xu, J. Zhang, X. Gong, et al. (2016). "A real-time flow classification method for power services based on an improved random forest algorithm." Power System Protection and Control, 44(24), 82-89.

[2] Y. Fan, X. Zou. (2016). "Research on WeChat traffic classification model and its business identification algorithm." Modern Electronic Technology, 39(15), 28-31. DOI: 10.16652/j.issn.1004-373x.2016.15.008.

[3] C. Ye, Z. Li, K. Zheng, et al. (2015). "A traffic identification method based on the user behavior state characteristics." Research on Computer Application, 32(02), 560-564 + 578.

[4] Y. Lu, Y. Chen. (2015). "Mobile Internet user traffic usage analysis based on Campbell model optimization." Internet World, (08), 42-52.

[5] Z. Chen, G. Cheng, Z. Xu, et al. (2023). "Review of research on Internet encryption traffic detection, classification and identification." Journal of Computer Science, 46(05), 1060-1085.

[6] Y. Wang. (2021). "CFS confusion: the challenge of business complexity to cash flow classification." Accounting of Township Enterprises in China, (07), 49-50.

[7] C. Ye, Z. Wang, S. Chen, et al. (2014). "Network traffic classification method based on node behavior feature analysis." Journal of Electronics and Informatics, 36(09), 2158-2165.

[8] R. Tang. (2020). "CART algorithm based on feature selection." (Doctoral dissertation, University of ESTC). DOI: 10.27005/d.cnki.gdzku.2020.000752

[9] H. Zhang. (2023). "Research on security method of wireless network based on encrypted traffic classification." Network Security Technology and Applications, No. 271(07), 23-25.

[10] F. Li, X. Wang, C. Zhang, et al. (2022). "The boundary point-based SVM classification algorithm." Journal of Shaanxi University of Technology (Natural Science Edition), 38(03), 30-38.

[11] H. Chen, X. Gao, J. Mei. (2012). "Fast forward backward algorithm for generalized hidden Markov models." Systems Engineering and Electronics Technology, 34(10), 2175-2179.

[12] X. Ren, D. Zhao, J. Qin. (2016). "Network intrusion detection system based on random forest and weighted K-mean clustering." Microcomputer Application, 32(07), 21-24.